Drivers and Mechanisms of Historical Sahel Precipitation Variability

Rebecca Jean Herman

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

COLUMBIA UNIVERSITY

2023

Abstract

Drivers and Mechanisms of Historical Sahel Precipitation Variability

Rebecca Jean Herman

The semiarid region between the North African Savanna and Sahara Desert, known as the Sahel, experienced dramatic multidecadal precipitation variability in the 20th century that was unparalleled in the rest of the world, including devastating droughts and famine in the early 1970s and 80s. Accurate predictions of this region's hydroclimate future are essential to avoid future disasters of this kind, yet simulations from state of the art general circulation models (GCMs) do a poor job of simulating past Sahel rainfall variability, and don't even agree on whether future precipitation will increase or decrease under global warming. Furthermore, climate scientists are still not in agreement about whether anthropogenic emissions played an important role relative to natural variability in dictating past Sahel rainfall change. Because the climate system is complex and coupled, it is difficult to determine which processes should be considered causal drivers of circulation changes and which should be considered part of the climate response, and therefore many theories for monsoon rainfall variability coexist in the literature. It is difficult to evaluate these competing theories because observational studies generally cannot be interpreted causally, but simulated experiments may not represent the dynamics of the real world.

The Coupled Model Intercomparison Project (CMIP) provides a wealth of data in which GCMs maintained at research institutions worldwide perform similar experiments, allowing the researcher to reach conclusions that are robust to differences in parameterization between GCMs. The scientific community has been using a wide range of statistical techniques to analyze this data, and each has notable limitations. This dissertation explores two statistical techniques for

leveraging CMIP to explore the drivers and mechanisms of historical Sahel rainfall variability: analysis of ensemble-mean responses to prescribed variables, and causal inference.

In Chapter 1, we give an overview of the climatology and variability of Sahel rainfall and present relevant physical theory.

In Chapter 2, we examine the roles of various types of anthropogenic forcing in observations and coupled simulations, using a 3-tiered multi-model mean (MMM) to extract robust climate signals from CMIP phase 5 (CMIP5). We examine "20th century" historical and single-forcing simulations—which separate the influence of anthropogenic aerosols, greenhouse gases (GHG), and natural radiative forcing on global coupled ocean-atmosphere system, and were specifically designed for attribution studies—as well as pre-Industrial control simulations, which only contain unforced internal climate variability, to investigate the drivers of simulated Sahel precipitation variability. The comparison of single-forcing and historical simulations highlights the importance of anthropogenic and volcanic aerosols over GHG in generating forced Sahel rainfall variability that reinforces the observed pattern, with anthropogenic aerosols alone responsible for the low-frequency component of simulated variability. However, the forced MMM only accounts for a small fraction of observed variance. A residual consistency test shows that simulated internal variability cannot explain the residual observed multidecadal variability, and points to model deficiency in simulating multidecadal variability in the forced response, internal variability, or both.

In Chapter 3, we investigate the causes for discrepancies in low-frequency Sahel precipitation variability between these ensembles and for model deficiency in reproducing observations. In the most recent version of CMIP – phase 6 of the Coupled Model Intercomparison Project (CMIP6) – the differences between observed and simulated variability

are amplified rather than reduced: CMIP6 still grossly underestimates the magnitude of low-frequency variability in Sahel rainfall, but unlike CMIP5, historical mean precipitation in CMIP6 does not even correlate with observed multi-decadal variability. We continue to use a MMM to extract robust climate signals from simulations, but now additionally include sea surface temperature (SST) as a mediating variable in order to test the proposed physical processes. This partitions all influences on Sahel precipitation variability into five components: (1) teleconnections to SST; (2) atmospheric and (3) oceanic variability internal to the climate system; (4) the SST response to external radiative forcing; and (5) the "fast" (not mediated by SST) precipitation response to forcing.

Though the coupled simulations perform quite poorly, in a vast improvement from previous ensembles, the CMIP6 atmosphere-only ensemble is able to reproduce the full magnitude of observed low-frequency Sahel precipitation variance when observed SST is prescribed. The high performance is due entirely to the atmospheric response to observed global SST – the fast response to forcing has a relatively small impact on Sahel rainfall, and only lowers the performance of the ensemble when it is included. Using the previously-established North Atlantic Relative Index (NARI) to approximate the role of global SST, we estimate that the strength of simulated teleconnections is consistent with observations. Applying the lessons of the atmosphere-only ensemble to coupled settings, we infer that both coupled CMIP ensembles fail to explain low-frequency historical Sahel rainfall variability mostly because they cannot explain the observed combination of forced and internal variability in SST. Though the fast response is small relative to the simulated response to observed SST variability, it is influential relative to simulated SST variability, and differences between CMIP5 and CMIP6 in the simulation of

Sahel precipitation and its correlation with observations can be traced to differences in the simulated fast response to forcing or the role of other unexamined SST patterns.

In this chapter, we use NARI to approximate the role of global SST because it is considered by some to be the best single index for estimating teleconnections to the Sahel. However, we show that NARI is only able to explain half of the high-performing simulated low-frequency Sahel precipitation variability in the atmospheric simulations with prescribed global SST. Statistical techniques commonly applied in the literature cannot distinguish between correlation and causality, so we cannot analyze the response of Sahel rainfall to global SST in more depth without atmospheric CMIP simulations targeted at every ocean basin of interest or a new method.

In Chapter 4, we turn to a novel technique called *causal inference* to qualify the notion that NARI can adequately represent the role of global SST in determining Sahel rainfall. We apply a causal discovery algorithm to CMIP6 pre-Industrial control simulations to determine which ocean basins influence Sahel rainfall in individual GCMs. Though we find that state of the art causal discovery algorithms for time series still struggle with data that isn't generated specifically for algorithm evaluation, we robustly find that NARI does not mediate the full effect of global SST variability on Sahel rainfall in any of the climate simulations. This chapter lays the foundation for future work to fully-characterize the dependence of Sahel precipitation on individual ocean basins using the non-targeted simulations already available in CMIP – an approach which can be validated by comparing the composite results to the interventional historical simulations that are available. Furthermore, we hope this chapter will guide algorithm improvement efforts that are needed to increase the performance and usefulness of time series causal discovery algorithms on climate data.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank Yochanan Kushnir, Michela Biasutti, and Alessandra Giannini for bringing me to Lamont to apply my diverse background to this important subject. Special thanks to Alessandra for training me how to use the Lamont data libraries, to Michela for introducing me to conferences, meetings, seminars, reading groups, and books that ignited my excitement about monsoons and causal inference, and to Yochanan Kushnir for being a constant source of support and guidance throughout my PhD. I also can't express enough gratitude to Adam Sobel and Ronald Miller, who inspired me early in my PhD with their excellent classes, and who later became my academic advisors.

I acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and thank the climate modeling groups for producing and making available their model output, which has provided the vast majority of the data I've analyzed during my PhD. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. All CMIP simulations are freely available through the Earth System Grid (see https://esgf-node.llnl.gov/projects/esgf-llnl/). I thank Haibo Liu for preparing CMIP5 (CMIP5, Taylor et al. 2012) data for use on the IRI/LDEO Climate Data Library, and Naomi Henderson, Julius Busecke, and Charles Stern for preparing CMIP6 (Eyring et al. 2016) data for use on the Pangeo project (https://pangeo.io/) Google cloud storage (https://console.cloud.google.com/storage/browser/cmip6), and I thank all of them again for their support when I needed to access the data.

I am indebted to Elias Barenboim for catering to my research interests while teaching me causal inference in a wonderful year-long course, to Amirkasra Jalaldoust for exploring the concepts with me and connecting me to the causal inference community, and to Adele Ribeiro for her excellent support and guidance throughout the class and also afterwards as I worked with her on the last chapter of this dissertation. I also acknowledge Andreas Gerhardus and Jacob Runge for designing LPCMCI—the causal discovery algorithm I use in my last chapter—as well as the entire Causal Inference and Climate Informatics Group at DLR's Institute of Data Science for developing causal discovery algorithms applicable to climate data in tigramite, and for welcoming me to join the team in Fall 2022.

Throughout my PhD, I received invaluable technical support from many people. Thank you to Naomi Henderson for maintaining the Lamont machines and setting up my account, and to Kyle T. Mandli for helping me learn how to effectively use Columbia's High Performance Computers. Further thanks to Naomi and Tom Nicholas for their support with python, to Yash Amonkar and my sister Arielle Herman for their patience and support with R, and to Sagar Simha for implementing LPCMCI in python and for the continuous technical support in using it.

I'd also like to thank Michela Biasutti, Oded Stein, and my family for invaluable editing at various stages of my PhD, and Meir Brooks for support during the substantial final effort to finish the writing of this dissertation.

I would like to thank Nathan Lenssen, whose friendship and scientific discourse has been invaluable to me throughout my PhD and will continue to be treasured going forward. Finally, I extend my deepest gratitude to Nathan Steiger of Hebrew University, who encouraged me to pursue my interest in applying causal inference to the field of climate science despite the many challenges of adopting a novel approach.

Much of this research was supported by the U.S. National Science Foundation grant AGS-1612904. Alessandra Giannini also benefited from French state support managed by the Agence Nationale de la Recherche under Investissements d'Avenir Program contract ANR-17-MPGA-0015, and Michela Biasutti and Yochanan Kushnir received funding from DOE Grant DESC0014423.

Chapter 1. Sahel Climatology and Variability

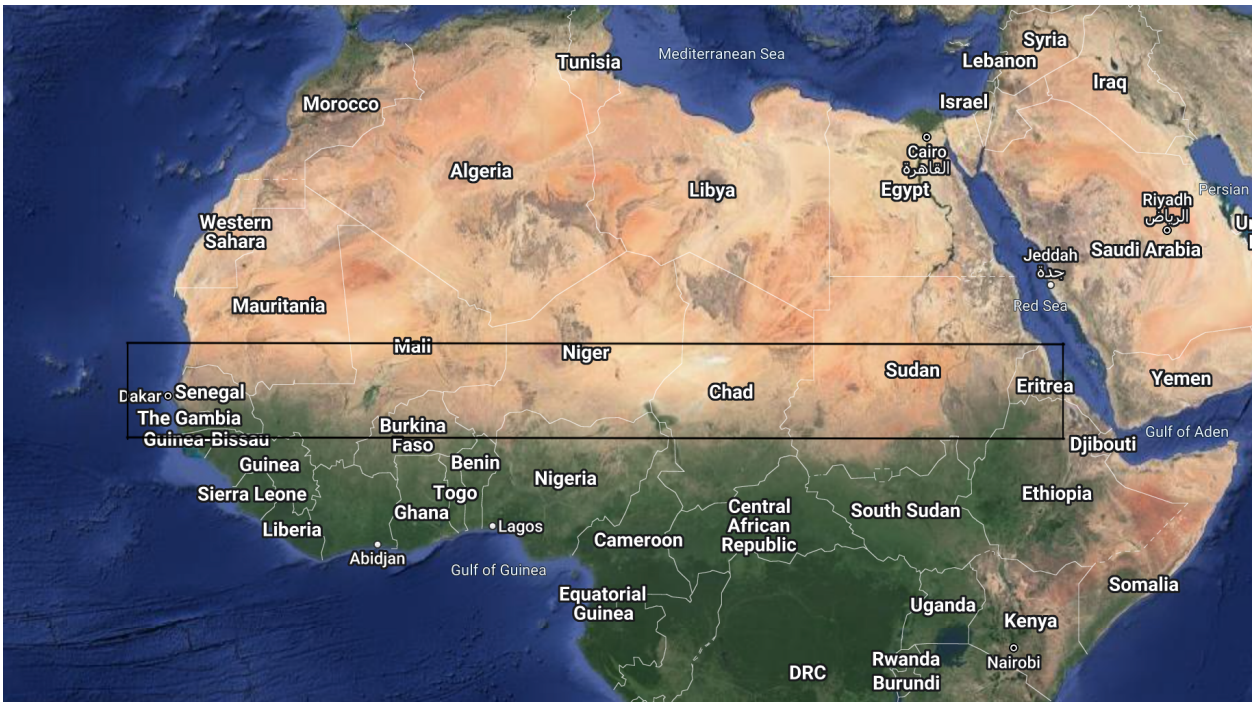

Figure 1.1: Satellite view of the Sahel (outlined in black) from Google Earth.

1.1. Introduction and Climatology

The Sahel is the semi-arid region bordering the North African Savanna and the Sahara Desert and spanning the width of Africa. The satellite imagery in Figure 1.1 highlights the zonal (East-West) symmetry of vegetation and albedo (surface reflectivity) in the region, and outlines the area that defines the Sahel for the purposes of this dissertation: 12°-18°N and 20°W-40°E. Rainfall in the Sahel is also quite zonally homogenous in its climatology (mean over the observed record, Figure 1.2a in the black box) and its historical variability (Figure 1.3d shows the spatial pattern of the leading mode of Sahel precipitation variability), though this uniformity may not hold under future global warming (Chou et al. 2001; Marvel et al. 2020). The Sahel receives its rainfall almost exclusively between July and September (JAS, see Figure 1.2b, turquoise contour) during the West African Monsoon (WAM), when zonal-mean rainfall

migrates north following seasonal heating from the sun (yellow-red shading shows the mean temperature at different latitudes over the course of a year). The WAM is additionally characterized by changes in the horizontal wind field (streamlines, Figure 1.2a): the annual-mean northeasterly winds (originating in the northeast) between the equator and the Sahel reverse and become southwesterly, while the annual-mean northeasterly winds over the Sahara Desert, known as the harmattan winds (Nicholson 2013), remain unchanged. (See Table 1.1 for a list of acronyms used in this dissertation.)

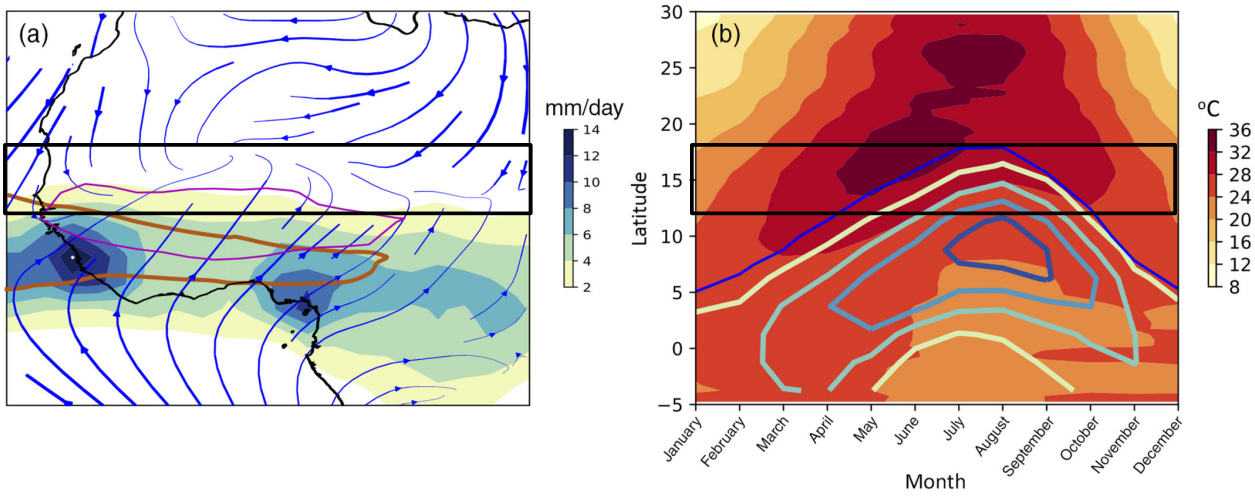

Figure 1.2: (Figure 1 in Biasutti 2019) The main features of the rainfall climatology in West Africa, with the latitude of the Sahel denoted in black. (a) Mean May–October (MJJASO) rainfall (color shading) and near surface (925 hPa) wind (streamlines; ERA40, Uppala et al. 2005). All fields are for the 1979–1998 period, for consistency with Thorncroft and Hodges (2001). (b) The seasonal cycle of sector mean (20°W-30°E) surface air temperature (warm shaded colors; CPC Monthly Global Surface Air Temperature, Fan and Van den Dool 2008), precipitation (contours, same colors as in (a); GPCP Huffman et al. 1997), and the intertropical discontinuity (ITD), where the WAM southwesterlies meet the harmattan northeasterlies (indicated in blue by the zero contour of the 925 hPa meridional wind from ERA40 reanalysis). Note how the advance of the ITD is connected with the warming of the Sahara and how the rain band stays to the south of the ITD. Panel (a) additionally shows the maximum African Easterly Waves (AEW, orange color; contour retraced from Figure 5a in Thorncroft and Hodges 2001) as indicated by the a track density scaled to number density per unit area (106 km^2) per season (MJJASO) greater than 6, and the location of the African Easterly Jet (AEJ, magenta contour) as indicated by the 9 m/s contour of the easterly wind at 600 hPa; these are features are important for the regional circulation paradigm of Sahel rainfall change.

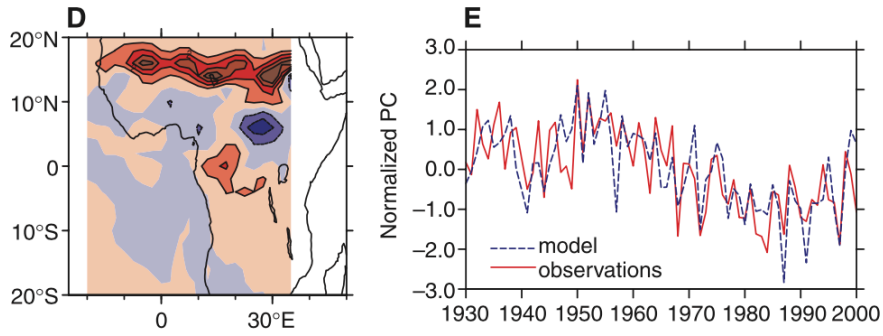

Figure 1.3: (From Figure 2 in Giannini et al. 2003) Principal Component Analysis (PCA) of simulated northern summer rainfall over tropical Africa during 1930-2000. (d) Second leading spatial pattern (Empirical Orthogonal Function, EOF) where red are positive precipitation anomalies and blue are negative anomalies. (e) Principal Component (PC) time series (blue). Panel (e) additionally displays the PC from the corresponding analysis using observations (red); the two PCs correlate at 0.73. The first EOF and PC don't project onto the Sahel; the second (pictured) explain 15% of observed and 21% of ensemble-mean precipitation variance across tropical Africa, and capture the well-known low-frequency variability in Sahel rainfall.

The tropical oceans also have a zonal band of rainfall that migrates north during the summer and is generally co-located with low surface pressure, ascent, and surface wind convergence, and is therefore called the Intertropical Convergence Zone (ITCZ). Because these features co-locate, all of them have been used to identify the location of the marine ITCZ (Nicholson 2013); however, over land, the features separate. The WAM is characterized by two areas of ascent, which can be seen in the dark shading in the top panel of Figure 1.4: a deep ascent associated centered near 10°N, and a shallow ascent centered at 20°N (which is closer to the latitude of maximum temperature at 25°N, Figure 1.2b). Each area of ascent is associated with a track of African Easterly Waves (AEW), separated both by latitude and by vertical extent: the shallow circulation with the African Easterly Jet (AEJ) in the mid-troposphere, and the deep ascent with the Tropical Easterly Jet (TEJ) in the upper troposphere (Nicholson 2009). The shallow ascent is part of a Shallow Meridional Circulation (SMC) in which air rises and then turns south to enter the deep ascent at around 700 mb (Nicholson 2009, 2013). It is associated

with surface wind convergence and a low-pressure system known as the Saharan Heat Low (SHL), which also induces geostrophic counter-clockwise near-surface flow. However, it is the deep ascent that is associated with rainfall (see bottom panel). To avoid confusion between the location of surface convergence and the location of land rainfall, we follow Nicholson's lead and refer to the zonal band of rain over Africa as the tropical rainbelt. The Sahel lies between the areas of strongest ascent (Figure 1.4, bottom); it experiences the northern edge of the rainbelt, and can be classified as a "convective margin." Figure 1.5 shows a simplified schematic of the zonal mean overturning circulation.

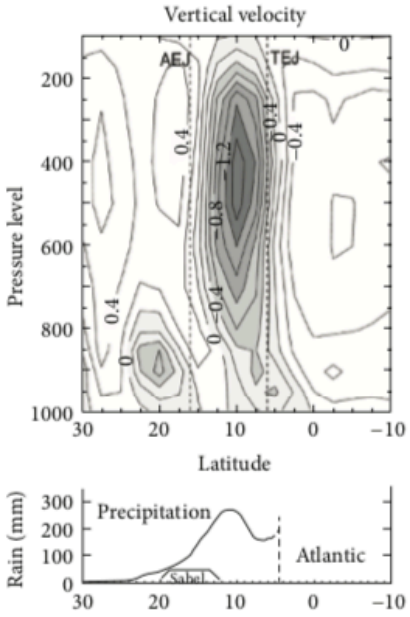

Figure 1.4: (Figure 11 in Nicholson 2009) Schematic of the rainbelt over West Africa. Top diagram is a vertical cross-section of mean vertical motion (10^{-2} Pa s^{-1}) in August. The main region of ascent lies between the axes of the African easterly jet (AEJ) and the tropical easterly jet (TEJ). A shallow region of ascent corresponds to the center of the Saharan Heat Low (SHL) and that maximum in surface convergence, which is sometimes associated with the land Intertropical Convergence Zone (ITCZ). The bottom diagram gives mean August rainfall (mm mo^{-1}, averaged for 10^{o} W to 10^{o} E) as a function of latitude, with the location of Sahel indicated on the latitudinal axis.

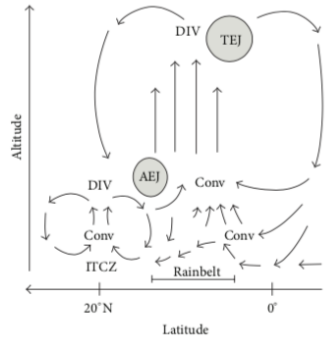

Figure 1.5: (Figure 18 in Nicholson 2009) Schematic illustration of Nicholson's view of the West African monsoon. Surface convergence is labelled ITCZ; other sources associated the rainbelt with the ITCZ even though surface convergence is lower there.

State of the art coupled climate simulations do a poor job of simulating past Sahel rainfall variability, and have wildly different projections of how Sahel rainfall will change in a warming world. Simplified conceptual depictions of the WAM system are necessary to diagnose the reasons for these differences, evaluate the simulations relative to observations and theory, and improve projections of the future. Because the climate system is complex and coupled, it is difficult to determine which processes should be considered causal drivers of circulation changes and which should be considered part of the climate response, and therefore many theories for monsoon rainfall variability coexist in the literature. The different paradigms used to explain and understand variability in Sahel monsoon rainfall can be broadly categorized into three different spatial scales (see Biasutti 2019 for a discussion of physical processes governing the WAM at different spatial scales). (1) At one extreme, some scientists portray rainfall at a given location as uniquely determined by moist atmospheric thermodynamics within an arbitrarily narrow column of air above it. In this view, horizontal processes are only relevant insofar as they affect column moisture and heat content. (2) On the other extreme, some scientists view rainfall across the tropics as an expression of the shifting global circulation, which is in turn determined by differences in heating and energetics on a planetary scale. (3) Lastly, some scientists depict

rainfall as determined by local regional circulation features on the order of hundreds to thousands of kilometers, such as land-sea contrast or variability in African Easterly Waves. Of course, climate phenomena at all spatial scales must coexist in the coupled climate system, so the truth may encompass all of these arguments. All of these paradigms have merit and deserve analysis. But in the interest of time, this introduction will focus on the column view and discuss the relevant theory for understanding the arguments that follow in the dissertation.

1.2. Basic Theory

Precipitation occurs when a parcel of air rises in the atmosphere until it reaches a pressure and temperature too cold to sustain its water vapor or suspended water droplets. In the absence of a mechanical disturbance, such vertical motion is called convection and is driven by differences in density: when an air parcel's density differs from its surroundings, it experiences a vertical buoyant force proportional to the fractional difference between its density and the density of the surrounding air. The density of a parcel of air can be expressed to a good approximation in terms of pressure (p) and temperature (T) according to the ideal gas law: $\rho = p/RT$, where R depends on the composition of gases in the parcel. The biggest differences in the composition of air over time and space are associated with the presence of water vapor, so if an air parcel's pressure matches the surrounding air, the difference in temperature and moisture content between the air parcel and the environment determines its buoyancy. Virtual temperature (T_v) corrects the ideal gas law for differences in density due to the presence of water vapor, allowing us to rephrase the ideal gas law using the ideal gas constant for dry air (R_d) and virtual temperature: $\rho = p/R_d T_v$. If we can predict the virtual temperature of the parcel as it rises (T_{vp}), and if we know the virtual temperature profile of the column (the "environment" surrounding the parcel, T_{ve}), we can quantify column (in)stability as the integral of the difference in buoyancy

due to virtual temperature from the level of free convection (p_f, where a parcel might begin to rise) to the level of "neutral" buoyancy (p_n, where the virtual temperature of the air parcel matches that of the column), termed convective available potential energy: CAPE = $\int_{p_f}^{p_n} R_d\,(T_{vp} - T_{ve})/p\;dp$ (Glossary 2009). Here, positive CAPE is associated with instability and convection, while zero CAPE is associated with stability and subsidence (sinking air).

It is not automatically obvious how to predict the virtual temperature of a parcel as it rises: processes such as incoming or outgoing radiation, mixing with surrounding air from the column or from neighboring columns, condensation of water vapor, freezing of liquid water, and precipitation can all affect the temperature and water vapor content of the air. Moist static energy (MSE $= h = gz + c_pT + L_vq$)—defined as the sum of gravitational potential energy (g, the acceleration of gravity, times z, height), sensible heat (the heat capacity of air at constant pressure, c_p[1], times temperature, T), and latent heat (latent heat of water, L_v, times the partial pressure of water vapor, q)—is often used to predict temperature along the parcel's path of motion. It is a common choice because—under the assumption of hydrostatic balance, which defines the pressure profile such that changes in height and temperature roughly balance each other in the MSE budget for a rising parcel of air (Bohren and Albrecht 2000; Yano and Ambaum 2017)—it is relatively simple to calculate, and it is almost[2] conserved under vertical

[1] Most appropriately defined as $(1 - q_t)c_{pd} + q_tc_l$, where q_t is the total water content (equal to the sum of the specific vapor and specific liquid water), and c_{pd} and c_l are the heat capacities of dry air at constant pressure and of liquid water, respectively (Emanuel et al. 1994), but often (poorly) approximated with c_{pd}, the heat capacity of dry air (Yano and Ambaum 2017).

[2] Moist entropy, which is truly conserved under reversible adiabatic processes, is a more rigorous choice. The conservation of MSE is derived from conservation of entropy along with a number of approximations (Bohren and Albrecht 2000).

motions and condensation if there is no exchange of heat with the surroundings (adiabatic), no precipitation (reversible), and no freezing of water droplets or vapor[3].

If the virtual temperature of the environment just above the parcel is warmer than the parcel would be at that pressure, this may inhibit convection even when CAPE is positive. This state is called "conditional instability," and allows us to sometimes measure positive values of CAPE in the atmosphere before a convection occurs (e.g. Figure 1.6). But convection is a fast process, and CAPE is rapidly consumed when a column is actively convecting. This means that CAPE remains near a critical neutral value (Sobel 2007), and the column reaches a state of convective quasi-equilibrium (CQE; Arakawa and Schubert 1974; Emanuel et al. 1994) where the temperature profile of the column approximates a moist adiabat: the virtual temperature profile of the rising air. Thus, the thermodynamic properties of a precipitating atmospheric column are determined roughly by the near-surface MSE from which the adiabat arises, and precipitation rate is more readily associated with CAPE *production* and column energetics.

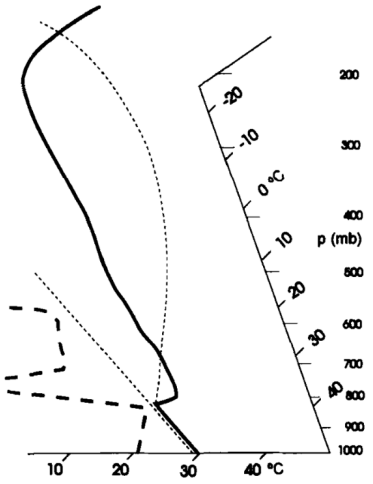

Figure 1.6: (Figure 1 in Emanuel et al. 1994) Conditionally unstable thermodynamic sounding, 00 GMT 7 May 1986, at Oklahoma City, Oklahoma. If raised through the column, a

[3] This can be simply added to the definition (Hill 2016; Yano and Ambaum 2017)

surface parcel of air would react to the change in height and pressure by changing temperature according to the dry adiabat (straight light dashed line) until the lifting condensation level (intersection of the straight and curved light dashed lines), at which point the temperature has decreased enough to saturate the parcel so that any further decrease will cause water vapor to condense. After that, the parcel follows the moist adiabat (curved light dashed line). The dark curve shows the temperature of the column. Between 800 and 700 mb, it is warmer than the moist adiabat, inhibiting convection. But above 700 mb, it is cooler than the moist adiabat. Convective available potential energy is proportional to the area between the curves.

1.3. Column Energetics

The column energetics framework assumes CQE in precipitating regions, and associates precipitation with latent heating that balances CAPE *production* by top-of-atmosphere radiation (R_t), surface fluxes including latent enthalpy (L_v) due to evaporation (E) and sensible heat (H), and column-integrated ($\{\cdot\}$) quantities including mean-flow ($\overline{\boldsymbol{u}}$) advection $\left\{\overline{\boldsymbol{u}} \cdot \nabla_p \bar{h}\right\}$, transient eddy flux divergence ($\nabla \cdot \left\{\overline{h'\boldsymbol{u}'}\right\}$), and MSE storage ($\partial\left\{\bar{h}\right\}/\partial t$).

The MSE budget ($\frac{\partial}{\partial t}\left\{\bar{h}\right\} + \left\{\overline{\boldsymbol{u}} \cdot \nabla_{\mathrm{p}} \bar{h}\right\} + \left\{\overline{\omega}\frac{\partial \bar{h}}{\partial p}\right\} + \nabla \cdot \left\{\overline{h'\boldsymbol{u}'}\right\} \approx L_v E + H + R_t + R_s$),

approximately relates changes in vertical motion ($\overline{\omega}$) that could produce precipitation to the MSE profile of the column ($\partial\bar{h}/\partial p$), MSE storage ($\partial\left\{\bar{h}\right\}/\partial t$), and the terms mentioned above. Because it contains many of the terms important for CAPE production, many studies thus set aside CAPE and focus instead on the MSE budget, which must hold in non-precipitating as well as precipitating regions (see Hill 2019 for a helpful overview of the MSE framework for monsoons and impactful studies). Gross moist stability is a method for quantitatively predicting precipitation using the MSE budget; it relates vertically integrated horizontal convergence of MSE (or some other conserved quantity) to the strength of moist convection. Using assumptions about the profile, one can predict precipitation increases and decreases via direct atmospheric responses to anthropogenic emissions or other factors.

1.4. Column Instability in the Tropics

Another approach approximates CAPE with a simple index directly representing the MSE profile of the column. In non-precipitating regions, the atmospheric column is not constrained to the moist adiabat, and upper-tropospheric MSE can and does vary independently from near-surface MSE. In the tropics, the free troposphere cannot sustain severe horizontal temperature gradients, since the small Coriolis parameter near the equator allows Inertial Gravity waves including eastward propagating Kelvin and westward propagating Rossby waves to rapidly spread any temperature perturbation over the whole tropics (Sobel et al. 2001). This affinity for a horizontally-uniform free tropospheric temperature is known as the weak temperature gradient (WTG) constraint. Thus, when deep convection in the tropics heats the upper troposphere, the resulting local temperature anomaly is quickly spread throughout the tropics (Chiang and Sobel 2002; Parhi et al. 2016), and the upper-tropospheric temperature in non-convecting regions is set by the near-surface MSE in precipitating regions (Zhang and Fueglistaler 2020).

The closeness of the tropical free tropospheric temperature profile to a single moist adiabat and the separation of near-surface temperatures from this adiabat in non-precipitating regions leads to a natural simplification of the seasonal-mean thermodynamic MSE profile: near-surface MSE in the boundary layer, and uniform MSE at upper levels that approaches tropical-maximum near-surface MSE (Zhang and Fueglistaler 2020). Thus, one may roughly associate perturbations to CAPE with perturbations in the difference between local near-surface and upper tropospheric MSE (i.e. Giannini 2010). As a near-surface parcel approaches the upper troposphere, ambient temperatures are so low that humidity is near zero both in the parcel and in the environment. This means that MSE at upper levels is essentially a measure of temperature,

and this allows us to approximate CAPE in terms of near-surface humidity (q) and temperature (T) as well as upper tropospheric temperature (TT). (This simplification ignores intrusion of dry air from the SMC, but could be modified to take this into account by examining lower-tropospheric mean MSE instead of near-surface MSE, in a fashion similar to that suggested by Shekhar and Boos (2016)). While many analyses use column-integrated values, the results from such studies are often still discussed using this conceptual simplification (e.g. Chou and Neelin 2004).

In the planetary boundary layer over ocean regions, where moisture supply is unlimited and relative humidity is relatively constant at 80% (Sobel 2010), sea surface temperature (SST) directly determines humidity; thus local near-surface MSE can be estimated solely using local SST (Sobel 2007), and global tropical TT can be estimated using a precipitation-weighted mean of global tropical SST (Sobel et al. 2002). Over land such as the Sahel, evaporation and moisture supply are limited, so the surface MSE maximum is strongly associated with moisture convergence (Chiang and Sobel 2002; Giannini et al. 2008) and often occurs at a different latitude than the temperature maximum. High surface temperatures alone are never able to create enough buoyancy to sustain deep convection, so latent heat release is crucial for convection (Giannini et al. 2008), and precipitation over land is associated not with the maximum of temperature, but of MSE.

1.5. Moist Static Energy Budget Analyses

Unfortunately, moisture convergence cannot be viewed strictly as a cause of precipitation – due to conservation of mass, an increase in vertical motion associated with precipitation must be balanced by an increase in mass (and moisture) convergence, so causality can flow both ways and the two are tightly coupled. However, large-scale external factors, such

as orography or changes in remote evaporation, may affect moisture convergence in the context of quasi-steady circulations, and so diagnostic analysis of moisture budgets is still useful (Sobel 2007).

For example, Chou and Neelin (2004) use column-integrated MSE budget analysis and employ both the column energetics (Section 1.3) and column instability (Section 1.4) frameworks to explain why tropical convective margins (such as the Sahel) often dry under global warming while neighboring convection intensifies.

The temperature of the upper troposphere sets the surface MSE threshold for convection, and when it warms for any reason—whether due to increased absorption of infrared radiation by greenhouse gases (GHG) or because near-surface MSE in convecting oceanic regions elsewhere has increased—tropical regions must match the increased MSE threshold in order to convect. (Indeed, the combination of WTG in the tropics and CQE in convecting regions means that, even at daily time scales, convection throughout the tropics only occurs where near-surface MSE is close to the tropical maximum; Zhang and Fueglistaler 2020.)

Global warming also increases temperature on and near Earth's surface, which in turn increases absolute humidity by increasing evaporation according to the Clausius-Clapeyron relation and relative humidity (Held and Soden 2006). If the climatological wind field did not change, this increase in absolute humidity would be expected to increase moisture convergence and MSE in convecting regions in a thermodynamic mechanism termed the "direct moisture effect". This helps precipitating regions meet the higher MSE threshold for convection, and overall would be expected to increase precipitation in precipitating regions.

But—because the Clausius-Clapeyron relation is non-linear and because non-convecting regions are associated with surface divergence—absolute humidity and MSE in non-convecting

regions increase less than in convecting regions, causing horizontal MSE gradients to increase under uniform global warming. Chou and Neelin (2004) argue that advection of air with relatively lower MSE from non-convecting regions prevents neighboring convective margins from meeting the convective threshold for MSE set by the upper troposphere, and suppresses precipitation. (This MSE budget analysis is diagnostic by nature, but Hill et al. (2017) later confirmed that uniform oceanic warming causes an increase in MSE gradients, advection of relatively lower-MSE air from the Sahara to the Sahel, and subsidence in the Sahel, though this doesn't imply reduced Sahelian precipitation in all models (see also Hill et al. 2018).) This mechanism, termed the "upped ante" mechanism, is represented in Figure 1.7.

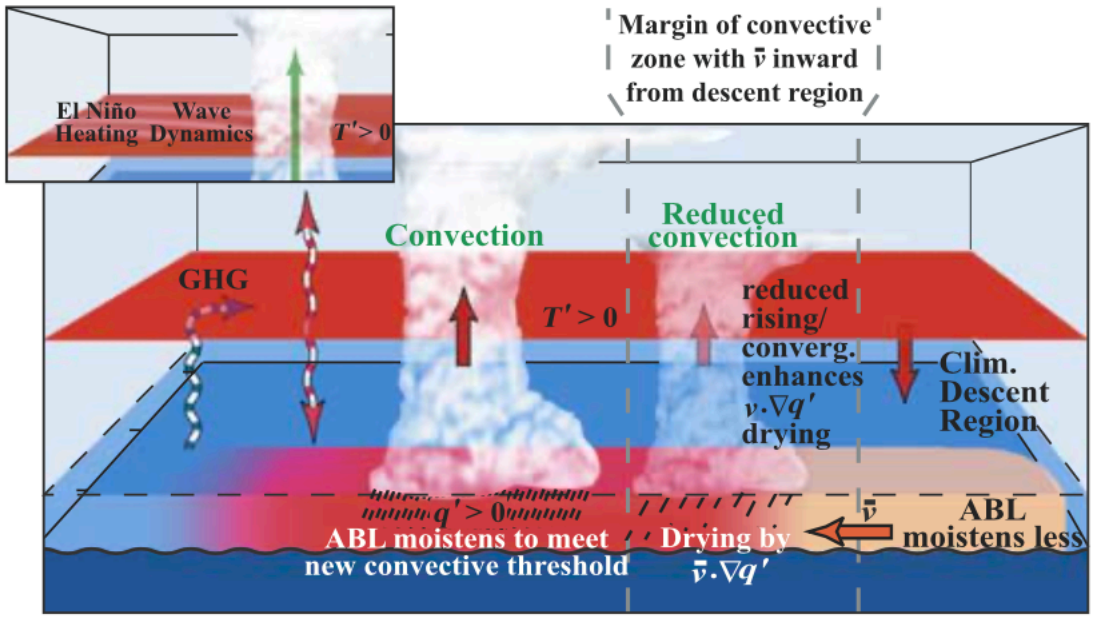

Figure 1.7: (Figure 3 from Neelin et al. 2003) Schematic of the "upped ante" mechanism for negative precipitation anomalies. For the global warming case, the tropospheric temperature warms due to increased absorption of infrared radiation (dashed curves) by greenhouse gases (GHG). For the El Niño case (inset) warming is spread from the Pacific by wave dynamics. The rest of the pathway via convective interactions is common to both. Adjustment of atmospheric boundary layer (ABL) moisture in convective regions, to meet the new convective "ante", establishes a gradient of ABL moisture anomalies q' relative to nonconvecting regions. This creates a tendency where low-level flow v moves into the

margin of a convective zone. Feedbacks reducing upward motion and low-level convergence enhance this drying tendency.

The reduction of convection in precipitating regions further triggers a dynamic feedback whereby the mass (and moisture) convergence that used to occur at the convective margins is relocated to the center of convecting regions, causing further decreases in precipitation in convective margins, but moistening the atmospheric boundary layer and causing further increases in precipitation in convective areas. This dynamic mechanism is termed the "anomalous gross moist stability mechanism" because it involves changes in horizontal convergence and convection that affect gross moist stability without directly changing moisture content in the atmosphere. This effect and the direct moisture effect together enhance the pattern of precipitation minus evaporation – an effect of global warming which is robust to model parameterization (Held and Soden 2006) – and both have been called "rich-get-richer" mechanisms (Chou et al. 2009). (For the rest of the dissertation, in this context, "thermodynamic" is used to refer to the "direct moisture effect" while "dynamic" refers to differences that can be traced to changes in the wind fields.)

1.6. El Niño Southern Oscillation

Neelin et al. (2003) outline the upped ante/anomalous gross moist stability framework in the context of the El Niño Southern Oscillation (ENSO), which is more complicated than greenhouse gas warming because SST warming is far from uniform over the tropics and its pattern changes over time. Warm upper tropospheric temperatures resulting from deep convection in the Pacific during the development phase of El Niño in spring are rapidly spread horizontally throughout the tropics before remote SST can respond. This increases temperature aloft without a matching increase in moisture supply, meaning all remote tropical precipitation might initially respond to El Niño like a convective margin responds to global warming. Later,

the tropospheric temperature anomalies are communicated vertically to SST in remote oceans via turbulent surface energy fluxes with a lag of a 3-6 months (Chiang and Sobel 2002), resulting in local thermodynamic increases in moisture supply during the mature phase of ENSO that counteract the "upped ante" mechanism. Parhi et al. (2016) explore this possibility, represented in Figure 1.8, with regards to African rainfall. They show that, in June, upper tropospheric temperature over the Sahel has already increased but neighboring SST has not yet changed, allowing the "upped ante" mechanism to have its full effect in the Sahel during the development of El Niño. Though the El Niño event may last into the next year, Global Tropical SST in May is strongly affected by El Niño the previous winter (Chiang and Sobel 2002), so an El Niño event may thermodynamically increase moisture supply to the Sahel during the spring of its demise, preventing a second dry year. It is worth note that in many GCMs, the atmospheric response to El Niño and the corresponding response of the WAM is delayed by a year (Joly and Voldoire 2009).

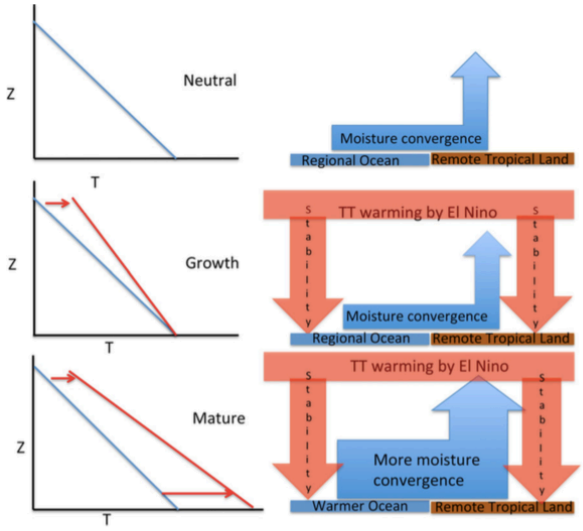

Figure 1.8: (Figure 1 from Parhi et al. 2016) Schematic showing our hypothesized mechanism acting upon remote tropical land and adjacent regional ocean (moisture source). (left) The environmental lapse rates are shown for the remote tropical land regions with temperature T on the x axis and height Z on the y axis. The lapse rate is shown in blue for the neutral case and in red for the El Niño case. Note how in the mature phase of El Niño the surface warming by moist static energy convergence increases the lapse rate (more negative slope).

(right) The moisture and heat fluxes are shown for the remote tropical ocean and land region for (top) neutral, (middle) growth, and (bottom) mature cases. The thick red horizontal lines in the cases of growth and mature phase indicate the TT warming advection in the free troposphere. The blue arrows indicate the ocean-to-land moisture advection and convergence in the rainy season. For mature phase, the blue arrow is thicker suggesting the bottom-up instability dominates over the stability caused by the TT warming acting top down.

1.7. The Sahel

Giannini (2010) first used the upped ante/anomalous gross moist stability framework to develop storylines for future Sahel precipitation change. She argued that future climate projections for the Sahel from different general circulation models (GCMs) diverge because differing sensitivities of local evaporation to global warming give some, but not all, GCMs enough moisture supply to meet the upped ante for convection, triggering the gross moist stability mechanism rather than the upped ante mechanism. Giannini et al. (2013) (G13 from here on out) then focus on historical Sahel rainfall variations. Arguing that moisture supply from the North Atlantic—defined from 10 to 40°N—has had a leading role in determining Sahel rainfall in observations and most simulations of the 20th century, they claim that the cooling and subsequent warming of North Atlantic SST in the 20th century thermodynamically (via changes in humidity over the North Atlantic) controlled anomalous moisture convergence in the Sahel and determined whether near-surface MSE there was able to meet the threshold for convection. They estimate the convective threshold (TT) using average SST in the global tropics[4], resulting in a single SST index for predicting Sahel precipitation: the North Atlantic Relative Index (NARI), defined as the difference in SST between the North Atlantic (NA) and Global Tropics

[4] Average of SST over the global tropics appears to have a computational correspondence to TT that is about as strong (or stronger) than that of precipitation-weighted mean tropical SST in the past (Sobel et al. 2002) and in many simulations of the future (Johnson and Xie 2010), but it has a weaker theoretical relationship (Sobel et al. 2002), and uncertainties in observational estimates of observed SSTs draw out the difference between these two indices in atmospheric simulations with prescribed SST (Flannaghan et al. 2014).

(GT, 20°S-20°N). Giannini and Kaplan (2019, hereafter GK19) tied Sahelian moisture supply to competing responses of North Atlantic SST to greenhouse gases and aerosols emissions via the thermodynamic mechanism discussed above (G13), or potentially via other dynamic mechanisms reviewed by Giannini et al. (2008) from a column viewpoint.

Because it is the best known linear indicator of Sahel precipitation, in Chapter 3 we will use NARI to represent the role of the global oceans in mediating the effect of anthropogenic emissions on 20th century Sahel precipitation variability, and we will employ physical arguments such as these to qualify our results. In Chapter 4, we will evaluate how well NARI represents global oceans in climate simulations. In both chapters, we will use the language of "causal diagrams," where variables are represented by nodes, and the causal dependence of node A on node B is represented by an arrow from node B to node A (this will be explained in detail in Chapter 4). This graphical picture is a way of representing pictorially which variables would be included in a functional definition defining another, without having to specify the form of the relationship.

Figure 1.9 summarizes the argument in GK19 for 20th century rainfall change—that precipitation (P) is determined by moisture supply (q) and upper tropospheric temperature (TT)—and additionally includes surface temperature (T), which is necessary for defining MSE. Upper tropospheric temperature is determined by SST in the Global Tropics (GT), which is in turn driven solely by greenhouse gases (GHG). As discussed in the previous section, greenhouse gases are also believed to affect upper tropospheric temperatures directly. This effect is neglected in GK19 because Giannini et al. (2003) show that Sahel precipitation is primarily SST-driven. This view may still be overly simplistic: Hill (2019) showed that the uniform cooling component of the global SST response to Anthropogenic Aerosols (AA) triggers a reverse upped ante

mechanism (presumably by cooling TT) that dominates the precipitation response over the Sahel in some GCMs. In GK19, the role of Anthropogenic Aerosols (AA) is to determine North Atlantic (NA) SST and moisture supply. Because G13 argue that moisture is directly supplied to the Sahel via atmospheric transport from the North Atlantic, we represent surface temperature as a function of North Atlantic SST and greenhouse gases.

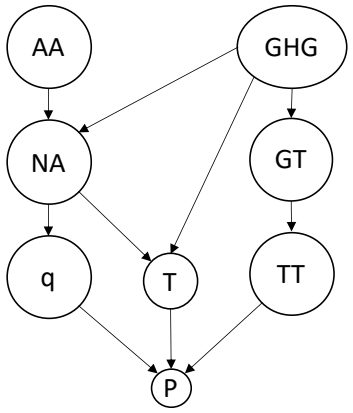

Figure 1.9: Causal diagram summarizing GK19's hypothesis for Sahel precipitation variability. Anthropogenic Aerosols (AA) and GHG affect North Atlantic SST (NA), which determines the humidity (q) and the temperature in the Sahel (T). GHG also affect SST of the global tropics (GT), which determines upper tropospheric temperature (TT). Q, T, and TT together determine Sahelian precipitation (P).

1.8. Moisture Supply and its Variability

G13's claim that the (subtropical) North Atlantic region supplies critical moisture to the Sahel is not well established in the literature. In the climatological (mean over tens of years) seasonal mean, the main sources of moisture to the Sahel are thought to include the Atlantic cold tongue in the Gulf of Guinea (Keys et al. 2014; Nicholson 2013), transported to the Sahel via an atmospheric "moisture river" (Lélé et al. 2015) in May and June (reaching a maximum buildup in June) and then via the WAM flow in JAS (Lélé et al. 2015; Thorncroft et al. 2011); and the Tropical Atlantic (mostly 8-11°N; Druyan and Koster 1989; Lélé et al. 2015; Pu and Cook 2012), transported to the Sahel via westerlies known as the West African Westerly Jet (WAWJ;

Pu and Cook 2010). Seth et al. (2013) find that springtime moisture supply is crucial for the accumulation of moisture in the boundary layer necessary to counteract the upped ante effect due to greenhouse gases (called the "remote" effect in this paper) in monsoon regions across the globe during early-summer, and that a lack of springtime moisture supply causes early-summer Sahel rainfall to be redistributed later in the season in future climate projections. Sahelian moisture may additionally originate in the Mediterranean (Keys et al. 2014; Peyrillé et al. 2007; Rowell 2003), or may be traced to farther reaches of the Gulf of Guinea (Keys et al. 2014). Figure 1.10 shows the moisture river in the Gulf of Guinea in May (a) and June (b, vectors and contours) as well as the WAM flow from the Gulf of Guinea and WAWJ inflow from the Tropical Atlantic in August (c, July and September are similar). Figure 1.11 shows the Sahel "precipitationshed," or region where at least 5mm of evaporated moisture later precipitates in the Sahel during the WAM, and the magnitude of moisture supply from each location, estimated from reanalysis data with an Eulerian moisture tracking method. It also shows the first two Empirical Orthogonal Functions (EOFs) of variability in the moisture origins field, which notably do not include moisture supply from the subtropical North Atlantic.

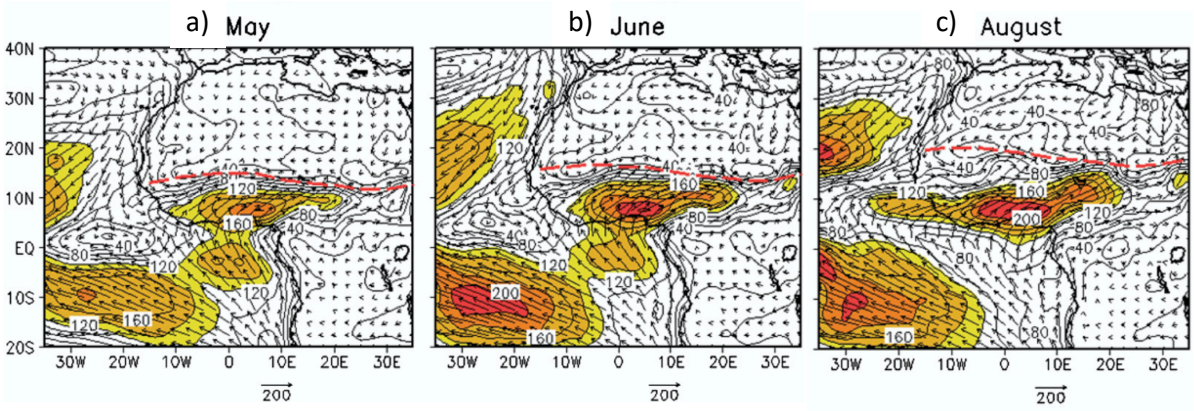

Figure 1.10: (From Figures 3 and 4 of Lélé et al. 2015) Mean surface-850-hPa seasonal distribution of vertically integrated mean moisture flux (vectors; kg m^{-1} s^{-1}) over the WAM region, overlaid by moisture flux magnitude (contours) in (a) May, (b) June, and (c) August averaged from 1979 to 2008. The Unit vector is displayed at the bottom of each panel and the

contour interval for flux magnitude is 20 kg m^{-1} s^{-1}. Values of moisture flux magnitude ≥ 100 kg m^{-1} s^{-1} are shaded. Thick dashed red line indicates the intertropical front latitude, defined as the 15°C dewpoint temperature.

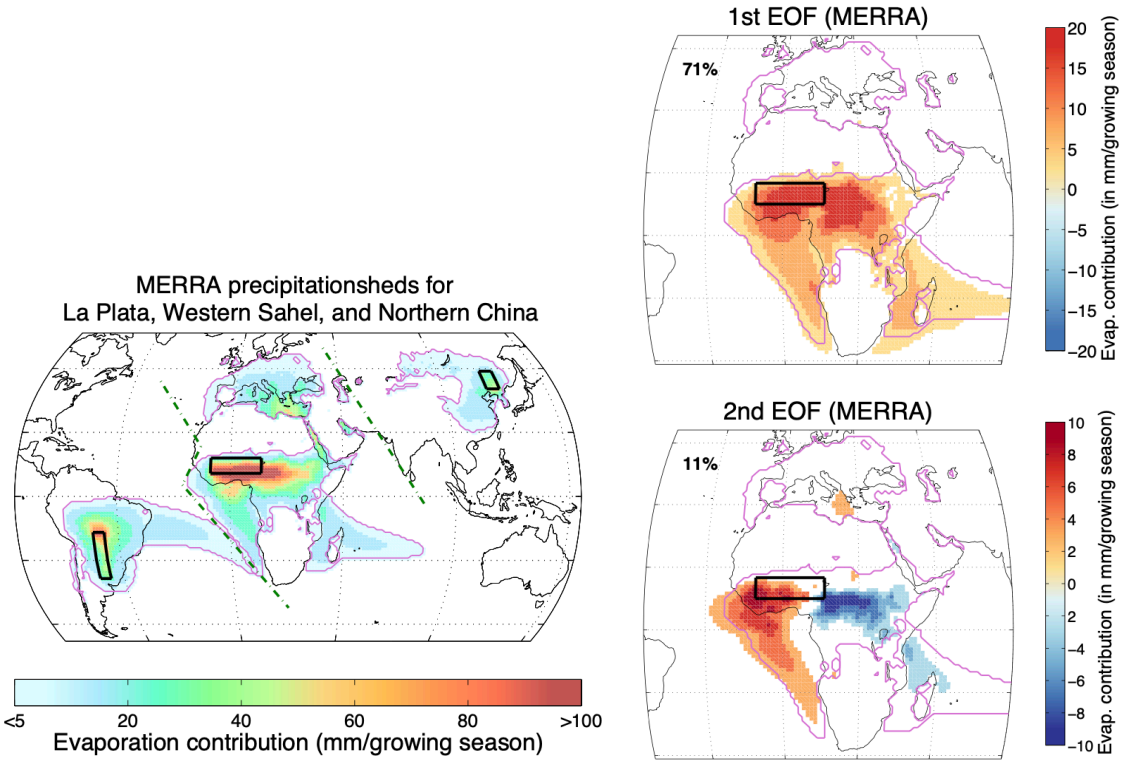

Figure 1.11: (Taken from Figures 2b and 6 of Keys et al. 2014) (Left) Mean precipitationshed (the upwind ocean and land surface that contributes evaporation to a specific location's precipitation) extents for MERRA for the period 1980-2011. Values less than 5 mm are excluded from the precipitationsheds. The green dotted-dashed lines separate the Sahelian precipitationshed (middle) from the La Plata and Northern China precipitationsheds. Note that the Mediterranean sources belong to the western Sahel. (Both) The magenta line indicates the 5 mm growing season^{-1} precipitationshed boundary, and the black box indicates the sink region. (Right) Comparison of first and second Empirical Orthogonal Functions (EOFs) for the western Sahel for MERRA during the period 1980-2011. The bold number in the upper left corner indicates the amount of variance explained by the associated pattern. Values < 2 mm growing season^{-1} are excluded for the sake of clarity of the figure.

One study that focused on a GCM with a questionable representation of the observed climate system[5] reached conclusions that are not inconsistent with climatological moisture

[5] It focuses on an old version of the NASA Goddard Institute for Space Studies model (GISS) that does a poor job of reproducing Sahel rainfall change and has an unrealistic climatological wind field

transport from the North Atlantic (though it fails to separate the North Atlantic from the Mediterranean Sea; Druyan and Koster 1989), and a couple of moisture advection studies show anomalous wind fields during wet years consistent with a small amount of moisture advection from part or all of this region (Park et al. 2015; Pu and Cook 2012). But the main mechanism for variability in moisture supply during the monsoon season (even in studies potentially consistent with moisture supply from the North Atlantic) is determined to be "dynamic" changes in atmospheric circulation that reduce transport of moisture from the Tropical Atlantic (and perhaps, to a lesser extent, the Gulf of Guinea) to the Sahel (Lélé et al. 2015; Liu et al. 2014; Park et al. 2015; Pu and Cook 2012) or reduce convergence of transported moisture in the Sahel (Druyan and Koster 1989). The North Atlantic may instead play a role in instigating dynamic changes in moisture supply.

The apparent leading-order role for dynamic changes in moisture supply does not preclude an important causal role for "thermodynamic" mechanisms mediated by changes in evaporation. For example, Chou and Neelin (2004) find that 70% of the greenhouse gas-induced increase in precipitation in regions of convection in climate projections is due to the (dynamic) "anomalous gross moist stability mechanism", which involves changes in MSE transport via circulation changes. But they maintain that this mechanism is a feedback triggered by the (thermodynamic) "direct moisture effect". Thermodynamic moisture supply mechanisms have been proposed for the Mediterranean Sea—which is thought moisten the north of the Sahel when it is warm via the Harmattan winds, allowing precipitation to penetrate further into the Sahel (Peyrillé et al. 2007; Rowell 2003)—and for the Tropical Atlantic (Liu et al. 2014).

Table 1.1: Definitions of acronyms used in this dissertation.

Climate Variables and Phenomena	
JAS	July-September
WAM	West African Monsoon
ICTZ	Intertropical Convergence Zone
AEW	African Easterly Waves
AEJ	African Easterly Jet
TEJ	Tropical Easterly Jet
SMC	Shallow Meridional Circulation
SHL	Saharan Heat Low
CAPE	Convective Available Potential Energy
MSE	Moist Static Energy
CQE	Convective Quasi-Equilibrium
q	(near-surface) absolute humidity
T	(near-surface) temperature
TT	Upper-tropospheric temperature (in Chapter 4: over the Sahel)
SST	Sea Surface Temperature
NA	North Atlantic
GT	Global Tropics
NARI	North Atlantic Relative Index = NA – GT
WAWJ	West African Westerly Jet
ENSO	El Niño Southern Oscillation
AMV	Atlantic Multidecadal Variability
AMOC	Atlantic Meridional Overturning Circulation
IOBM	Indian Ocean Basin Mode
PMM	Pacific Meridional Mode
NAO	North Atlantic Oscillation
Simulations, Climate Forcings, and Observational Products	
GCM	General Circulation Model
CMIP5	Coupled Model Intercomparison Project, phase 5 (Taylor et al. 2012)
CMIP6	Coupled Model Intercomparison Project, phase 6 (Eyring et al. 2016)
AA	Anthropogenic Aerosols
GHG	Greenhouse Gases
NAT	natural radiative forcing
ALL	The combination of AA, NAT, and GHG
piC	Pre-Industrial control simulations
AMIP	Atmospheric Model Intercomparison Project
amip-piF	amip-piForcing simulations (prescribed SST and pre-Industrial radiative forcing)
amip-hist	AMIP historical simulations (prescribed SST and ALL radiative forcing)
$P_{nonNARI}$	MMM precipitation – NARI, scaled by the amip-piF teleconnection strength (0.87)
GPCC	Global Precipitation Climatology Center dataset (Becker et al. 2013)
CRU	Climate Research Unit precipitation dataset (CRU; Harris et al. 2014)
Papers	
GK19	Giannini and Kaplan (2019)
G13	Giannini et al. (2013)
Statistical Jargon	
MMM	Multi-Model Mean: 3-tiered weighted mean over runs, models, institutions
IM	Institution Mean: 2nd tier of the MMM.
$R_{(LF)}$	Correlation (at low frequency)
$sRMSE_{(LF)}$	Standardized Root Mean Squared Error (at low frequency)
PDF	Probability Distribution Function
PS	Power Spectrum / Spectra

Causal Inference Jargon	
iid	Independent and identically distributed
SCM	Structural Causal Model
$A \perp\!\!\!\perp B \mid C$	A is "conditionally independent" of B given C
PAG	Partial Ancestral Graph
MAG	Maximal Ancestral Graph
ADMG	Acyclic Directed Mixed Graph
DC	Distance Correlation
CMI	Conditional Mutual Information
CMIknn	K nearest-neighbor estimator of CMI with a permutation-based significance test (Runge 2018b)
MCI	Momentary Conditional Independence
knn	K nearest neighbors for CMIknn
SN	Shuffle neighbors for CMIknn
PCMCI	Time-series causal discovery algorithm based on the Peter and Clark algorithm (PC) and MCI (Runge et al. 2019a)
PCMCI+	PCMCI with contemporaneous relationships (Runge 2020)
LPCMCI	"Latent PCMCI" with latent confounding (Gerhardus and Runge 2021)
α	Significance parameter for the conditional independence test for (L)PMCIC(+)
p	Number of preliminary iterations for LPCMCI
τ_{max}	Maximum considered time lag for (L)PCMCI(+)

Chapter 2. The Effects of Anthropogenic and Volcanic Aerosols and Greenhouse Gases on Twentieth Century Sahel Precipitation in CMIP5

Note: This chapter has been published in very near its present form as "The effects of anthropogenic and volcanic aerosols and greenhouse gases on twentieth century Sahel precipitation" in *Sci. Rep.* (2020), Vol. 10.1, pp. 1-11, doi: 10.1038/s41598-020-68356-w.[6] Minor edits have been made for clarity.

2.1. Introduction

The Sahel experienced dramatic, multi-decadal rainfall variability in the twentieth century which was unparalleled in the rest of the world. This variability was marked by a striking decline in rainfall between about 1960 and the early 1980s, including devastating droughts and famine in the early 1970s and 80s, which left 100,000 people dead and 750,000 dependent on food aid (Bird and Medina 2002). Scientists immediately began to explore potential relationships between Sahel rainfall and a wide variety of local (Charney 1975; Taylor et al. 2002a) and global (Folland et al. 1986; Rowell et al. 1995) climatic factors. Giannini et al. (2003) confirmed the importance of global over local processes by showing that an atmospheric model forced with observed global sea surface temperature (SST) alone could reproduce the profile of the first

---

[6] AUTHORS: Rebecca Jean Herman[a]*, Alessandra Giannini[a,c,d], Michela Biasutti[a,b], and Yochanan Kushnir[a,b]

[a] Columbia University,
[b] Lamont-Doherty Earth Observatory of Columbia University
[c] International Research Institute for Climate and Society, The Earth Institute at Columbia University
[d] Laboratoire de Météorologie Dynamique/IPSL, École Normale Supérieure, PSL Research University, Sorbonne Université, École Polytechnique, IP Paris, CNRS, Paris, France

* Corresponding author: Rebecca Jean Herman, rebecca.herman@columbia.edu

principal component of Sahel 20th century rainfall variability, if not the amplitude, at a correlation of approximately 0.7. Studies since then have continued to focus on various global processes, reinforcing the connections between the Sahel and the temperature of ocean basins across the globe, and establishing links to internal variability—such as the El Niño Southern Oscillation (Okonkwo et al. 2015; Parhi et al. 2016; Pomposi et al. 2016) and the Atlantic Multidecadal Oscillation (Okonkwo et al. 2015; Pomposi et al. 2015)—and external forcing—such as greenhouse gases (GHG; Ackerley et al. 2011; Dong and Sutton 2015; Giannini and Kaplan 2019; Giannini et al. 2013; Haarsma et al. 2005; Held et al. 2005) and volcanic and anthropogenic aerosols (Ackerley et al. 2011; Haywood et al. 2013; Polson et al. 2014). However, the relative importance of internal variability and different sources of external forcing remain unclear.

There is a developing consensus in the literature that anthropogenic aerosols have contributed to the Sahel drought, though there is disagreement over the prominence of this contribution and the physical mechanism that governs it (Ackerley et al. 2011; Biasutti and Giannini 2006; Chang et al. 2011; Giannini and Kaplan 2019; Haywood et al. 2013; Held et al. 2005; Hwang et al. 2013a; Iles and Hegerl 2014; Kawase et al. 2010; Polson et al. 2014; Pomposi et al. 2015; Robock and Liu 1994; Rotstayn and Lohmann 2002; Undorf et al. 2018). The magnitude of the contribution is somewhat contentious because of disagreement about the strength of the indirect aerosol effects (McCoy et al. 2017; Penner et al. 2006; Stevens and Feingold 2009), which may influence SSTs and global precipitation much more than the direct radiative effect (Booth et al. 2012; Lin et al. 2018; Wang et al. 2015), and which may cause non-linear interactions affecting both the spatial pattern (i.e. Polson et al. (2014) on the Asian monsoon) and even the mean (Lohmann and Feichter 2005) of the precipitation and temperature

responses to other sources of forcing. The role of greenhouse gases (GHG) is even more widely debated—not just in the 20th century (Ackerley et al. 2011; Booth et al. 2012; Chang et al. 2011; Dong and Sutton 2015; Giannini and Kaplan 2019; Giannini et al. 2013; Haarsma et al. 2005; Haywood et al. 2013; Kawase et al. 2010), but even in the future when GHG forcing dominates (Biasutti and Giannini 2006; Dong and Sutton 2015; Haarsma et al. 2005; Held et al. 2005). Some argue that there are also non-linear interactions between different effects of increasing GHG (Biasutti 2013; Biasutti et al. 2008) or between GHG and other external forcings (Giannini and Kaplan 2019) and internal processes (Neupane and Cook 2013). Finally, many studies claim that SST and Sahel rainfall variation are primarily of internal origin (Hoerling et al. 2006; Sutton and Hodson 2005; Ting et al. 2009).

Many of the above studies on the Sahel focus on one or two types of forcing or on one model, or are limited to CMIP3 (Meehl et al. 2007), in which most models did not include indirect aerosols effects. Some, such as Giannini and Kaplan (2019), use a storyline approach—focusing on proposing physically-consistent pathways by which anthropogenic emissions could affect societal welfare rather than proving the existence or significance of a climate response—in order to avoid underestimating regional impacts that result from "type II errors" such as scientific hesitation to make a claim and take action (Shepherd 2019). Others (Polson et al. 2014; Undorf et al. 2018) use fingerprinting, extracting distinct spatial and/or temporal patterns associated with different forcings and scaling the model response to match observations in order to correct sensitivity biases and avoid compensating errors in the models (Hegerl and Zwiers 2011). We attempt to enrich the debate about the influence of external forcing and internal variability on Sahel rainfall over the 20th century by performing an attribution study using multi-model means (MMM) of simulations from the Coupled Model Intercomparison Project phase 5

(CMIP5; Taylor et al. 2012). This is the first large ensemble of coupled models to include aerosol indirect effects (albeit of varying quality). Additionally, while previous ensembles have run simulations where all sources of external radiative forcing are set to historical values (ALL) and also control simulations where all sources of external radiative forcing are held constant at pre-Industrial levels (piC), CMIP5 is the first to run "single-forcing" model simulations, in which one source of external radiative forcing—such as GHG, anthropogenic aerosols (AA), or natural forcing (which includes volcanic aerosols and solar and orbital variations, NAT)—varies historically while the other external forcings are held at constant pre-Industrial values. This model ensemble will allow us to accurately characterize simulated responses to different forcing agents that are robust to model parameterization.

2.2. Methods

2.2.1. Data

Our index of Sahel rainfall variability is land-averaged precipitation anomalies for the monsoon season (July – September; JAS) over the region 12°-18°N, 20°W-40°E. For precipitation observations we use the Global Precipitation Climatology Center (GPCC) dataset (Becker et al. 2013), which is quite similar to the Climate Research Unit (CRU; Harris et al. 2014) dataset in average precipitation over the Sahel (see Figure 2.1a and Figure 2.1b). The two are compared in Figure 2.1, and GPCC is used for the rest of the paper. Model simulations come from the Coupled Model Intercomparison Project phase 5 (CMIP5; Taylor et al. 2012), which includes simulations by over 50 models from 20 different research institutions. Not all models contribute simulations to all four historical experiments; we use all available runs (between 1 and 10 for a given model) from all models (with a distinct name and physics number) and research institutions that have complete data from 1901 (where the observed rainfall record begins) to

2003 (where some models stop their historical simulations). There are 14 models from 8 institutions that contributed model simulations to the AA experiment, 21 models from 15 institutions that contributed to the GHG experiment, 22 models from 15 institutions that contributed to the NAT experiment, and 51 models from 20 institutions that contributed to the ALL experiment (Tbl. A.1). Here, if the physics number is changed, it is treated as a different model under the same institution.

2.2.2. The Multi-Model Mean

The MMM is defined as a 3-tiered, weighted average: (1) across individual simulations (runs) to get an ensemble mean (EM) for each model, (2) across EMs to get an institution mean (IM) for each research institution, and (3) across IMs to get the MMM for that experiment. While any averaging helps to filter internal variability from the MMM, the first tier focuses on reducing internal variability present in the individual runs, the second tier focuses on reducing variability between models from uncertainty in parameter values, and the third tier focuses on reducing variability between institutions from uncertainty in parameterization. A simple mean across all model simulations is very similar to the tiered mean (not shown), but tiers are used to prevent over-representation of some parameterizations and parameter choices (associated with models that provide more simulations) in the MMM and in the uncertainty and significance calculations described in Section 2.2.4. (These differences are easier to see in the significance calculations; not shown.)

If a random variable (such as the internal variability component of yearly JAS Sahel precipitation) has a variance of σ^2, then the mean over n realizations of that variable will have a variance of σ^2/n. The forced variability component may experience some attenuation as well due to differences in the simulated response to forcing between models. Given that the forced

signal ought to be similar across simulations from a given model, we expect attenuation of internal variability to overwhelm attenuation of forced variability. Thus, means over models with more runs or over institutions with more models will have a higher signal (forced variability) to noise (internal variability) ratio than their counterparts. However, they will also have less total variability, causing them to (counterproductively) contribute *less* to the MMM than their noisier model-mean counterparts. We counteract this by using weights which are inversely proportional to the expected attenuation of noise in the MMM tiers due to the number of ensemble members.

For a weighted mean $\sum_i w_i X_i$ between independent random variables X_i with mean μ_i, variance $a_i\sigma^2$, and weight w_i, where $\sum_i w_i = 1$, we find that:

$$\sigma^2_{\sum_i w_i X_i} = \mathrm{E}[(\sum_i w_i X_i)^2] - \mathrm{E}[\sum_i w_i X_i]^2 = \sum_i w_i^2 \, (\mathrm{E}[X_i^2] - \mu_i^2) = \sigma^2 \sum_i w_i^2 a_i$$

Thus, to counteract the attenuation from a previous tier, captured in a_i, we define the weights as $w_i = a_i^{-\frac{1}{2}} / \sum_i a_i^{-\frac{1}{2}} \propto a_i^{-\frac{1}{2}}$. Specifically, let f, i, m, r, N_f, N_{fi}, and N_{fim} be such that each forcing experiment f is simulated by N_f institutions, with N_{fi} models from each institution i, and N_{fim} runs from each model m, and assume that the JAS Sahel precipitation in a given year for each run r has a variance of σ^2. In the first tier, where $a_{fimr} = 1$ and $w_{fimr} = \frac{1}{N_{fim}}$ (an unweighted mean), we find that the variances of the EMs are $\sigma^2_{EM_{fim}} = \sigma^2 \sum_r \frac{1}{N_{fim}^2} = \frac{\sigma^2}{N_{fim}}$, giving $a_{fim} = \frac{1}{N_{fim}}$ for the second tier. To combat this attenuation, in the second tier we define weights $w_{fim} = \frac{\sqrt{N_{fim}}}{\sum_m \sqrt{N_{fim}}} = \frac{\sqrt{N_{fim}}}{M_{fi}} \propto \sqrt{N_{fim}}$, where $M_{fi} = \sum_m \sqrt{N_{fim}}$ is the normalization constant for those weights. Using these weights, the variances of the IMs are $\sigma^2_{IM_{fi}} = \sigma^2 \sum_m \frac{N_{fim}}{M_{fi}^2} \frac{1}{N_{fim}} = \frac{N_{fi}}{M_{fi}^2}\sigma^2$, giving $a_{fi} = \frac{N_{fi}}{M_{fi}^2}$ for the third tier. Then in the third tier, $w_{fi} \propto \frac{M_{fi}}{\sqrt{N_{fi}}}$.

2.2.3. Approach

MMMs are compared to observations using correlations, which capture similarity in frequency and phase, and standardized root mean squared errors (sRMSE), which capture differences in magnitude and are expressed as a fraction of observed variance. When comparing the observations to themselves, the correlation would be 1 and the sRMSE would be 0; when comparing the observations to a constant prediction, the correlation would be 0 and the sRMSE would be 1 (or 100% of observed variance).

2.2.4. Uncertainty and Significance: Bootstrapping and Randomized Bootstrapping

Estimates of sampling uncertainty over all possible model parameterizations are obtained by bootstrapping (resampling with replacement) available forced IMs before calculating the MMM and corresponding correlations and RMSE, yielding probability density functions (PDF) around the MMM correlation and RMSE. This PDF can also be interpreted as a measure of agreement between CMIP5 models.

In addition to uncertainty derived from model parameterization, the MMM still contains noise from lingering coincident internal variability, and because bootstrapping underestimates variance when sample size is small, this procedure does not capture the full magnitude of that uncertainty. To measure this noise, we perform a similar analysis on the piC simulations, which contain only internal variability. We improve the procedure by artificially increasing the sample size by choosing random, continuous, 103-year subsets from the long, constant-forcing piC runs before each bootstrap. This "randomized bootstrapping" procedure captures the noise uncertainty in the MMM, as evident from the nearly-uniform confidence intervals of the piC MMMs (yellow, Figure 2.2), which contain no time-varying signal—due to forcing or to internal variability—by construction. (When randomizing is not used while bootstrapping the piC

MMMs, for comparison, the piC confidence interval expresses high-frequency variability similar to that seen around the forced MMMs in Figure 2.2: not pictured). However, it is worth note that it may be a slight overestimate since the MMM is less effective at filtering noise when there are fewer runs per model, and there is almost always only one piC run per model.

We test the null hypothesis—that all results from the forced experiments are consistent with noise in the MMM derived from modelled internal variability alone—by repeating the randomized bootstrapping procedure once for each of the four forced experiments, using piC runs from the set of models contributing historical simulations to that experiment.

2.2.5. Residual Consistency

We evaluate consistency between modelled and observed internal and externally-forced variability by examining and comparing the power spectra (PS) of individual ALL and piC simulations. For increased resolution of sampled frequencies, we zero-pad the time series before taking the PS, and for clarity and decreased uncertainty, we average across PS from the same model before presenting the PS. This averaging is precluded for the piC simulations, which usually contain only one (long) simulation per model. To help reduce uncertainty in the piC PS, we divide them into consecutive, non-overlapping segments of 103 years, calculate the PS of the segments separately, and average them together. To account for the effect of climatological rainfall biases on spectral power, we calculate "rescaled PS" by scaling the individual ALL and piC runs from a given model by (mean observed 20th century precipitation)/(mean precipitation from the ALL runs for that model) before taking the PS. We then average the PS by model and present the 66% and 95% range of the PS. We also present 3-tiered, unweighted means over all simulations of the rescaled ALL, AA, NAT, and GHG PS. We use an unweighted mean because

differences in the timing of different simulated realizations of internal variability do not affect the magnitude of the spectral peaks characterizing that variability.

2.3. Results

2.3.1. Multi-model mean performance

In Figure 2.1a and Figure 2.1b, we compare the MMM of Sahel 20th century precipitation anomalies for the ALL MMM (blue line) to individual ALL runs (blue-grey lines, background) and IMs (cyan lines), and to observations from the Global Precipitation Climatology Center (GPCC, black line; Becker et al. 2013) and the Climatic Research Unit (CRU, red dotted line; Harris et al. 2014). Despite disagreement in the first three years, the spatial averages of the two observational records look similar enough that the choice of observational product should not affect the results, and only GPCC is used throughout the rest of the paper. Despite the spread of the IMs, the standardized anomalies (Figure 2.1a) reveal a striking similarity between observations and the MMM, which captures much of the time series' multi-decadal variation by reproducing the drought of the 70s and 80s and its recovery, and even many episodes of dramatic interannual rainfall changes, most notably near 1984, the driest year in observations. Assuming the averaging was successful in preferentially filtering out internal variability present in individual model simulations, the MMM represents a consensus, forced Sahelian rainfall profile which is recognizable in the observations (Figure 2.1a). However, the actual rainfall anomalies (Figure 2.1b) reveal substantial attenuation of variance in the ALL MMM compared to individual simulations and to the observations.

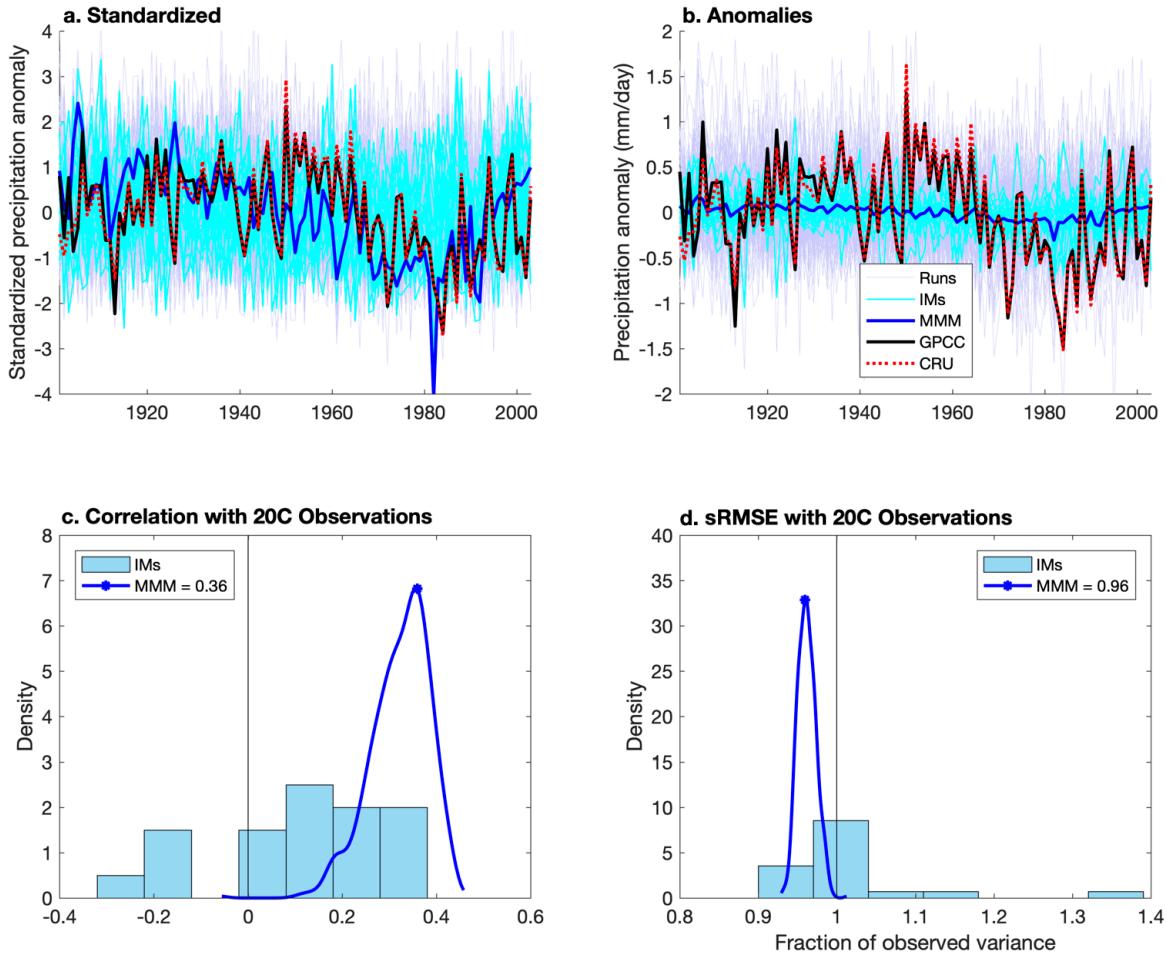

Figure 2.1: MMM Performance: Standardized (a) and actual (b) departures from climatology of 20th century Sahel precipitation in Individual ALL runs (blue-grey solid lines), ALL institution means (IMs, cyan), the ALL MMM (blue), and observations from GPCC (black) and CRU (red dotted line). Histogram (cyan) of correlations (c) and sRMSE (d) between GPCC observations and the IMs, actual correlation (c) and sRMSE (d) of the MMM with observations (blue dot), and the bootstrapping PDFs (blue curve) of the correlation (c) and sRMSE (d) between the ALL MMM and observations.

The remaining panels of Figure 2.1 display the correlations (Figure 2.1c) and the RMSE (Figure 2.1d) of individual IMs (cyan histogram) and of the MMM (blue dot) with observations. The blue curves show probability density functions (PDF) from bootstrapping over the IMs, and represent how those statistics might change with a slightly different set of models. The correlation measures the similarity in the shape of one time series with respect to the other but is

independent of relative amplitude, whereas the sRMSE estimates the difference in amplitude of the simulated and observed yearly rainfall time series.

The MMM performs as well as or better than most individual IMs in both metrics, consistent with previous studies which compared other versions of multi-model means to individual models (Gillett et al. 2002). Though some research institutions may appear to outperform the MMM in correlation and RMSE with 20th century observations (notably, GISS outperforms the MMM in both), as we are comparing only one variable (precipitation) to one realization of observations in which forced and internal variability are indistinguishable, it is unclear whether these models truly capture the underlying mechanisms better than the ensemble. The RMSE values for the MMM and the IMs are near 100% of observed variance, partially reflecting the severe attenuation seen in Figure 2.1b.

2.3.2. Model response to different forcing experiments

Figure 2.2 displays the MMMs for the three different single-forcing experiments: AA for **a**nthropogenic **a**erosols (pink, Figure 2.2b), NAT for **nat**ural forcing (brown, Figure 2.2c), and GHG for **g**reen**h**ouse **g**ases (green, Figure 2.2d); and compares them to observations (black). Figure 2.2a again displays the ALL MMM (blue). Note that the observations correspond to the black ordinates on the left, while forced and piC model outputs (colors, including yellow) correspond to the colored ordinates on the right, which have a scale a quarter the range to facilitate comparison. The blue, pink, brown, and green shaded areas are the 95% range of bootstrapped forced MMMs. They represent agreement in the forced signal between the institutions, even though—due to small sample size—they do not fully capture the magnitude of noise in the MMM caused by coincident simulated internal variability (see Section 2.2.4). The yellow shaded areas are also a 95% confidence interval, but they are obtained using randomly-

chosen continuous subsequences of the piC runs in place of the historical simulations, where the piC simulations are taken from the same set of research institutions which provided simulations for that historical forcing experiment. The yellow shading thus estimates the magnitude of noise in the MMM.

The variance of the forced MMMs over time (solid lines) and of the bootstrapped forced MMMs and randomized bootstrapped piC MMMs in a given year (shaded areas) vary from panel to panel inversely (though not proportionally) with the square root of the number of research institutions which simulated each forcing experiment (N). They are all roughly a quarter of observed variance – consistent with many precipitation fingerprinting studies, which often scale simulated precipitation up by a factor of 3-5 (Hegerl and Zwiers 2011; Polson et al. 2014; Undorf et al. 2018; Zhang et al. 2007). Aside from a few exceptions, the yearly magnitudes of the forced MMMs are not significantly different from zero, as they do not surpass the yellow zone consistent with noise in the MMM; this limits the detail with which we can examine the MMM directly. However, NAT (Figure 2.2c) and ALL (Figure 2.2a) are both significantly dry in 1982 (the year of the El Chichón eruption, near the driest observed year in 1984), and AA (Figure 2.2b) and ALL both display multi-decadal variability in the second half of the century (including a partial recovery) that is characteristic of the observations and uncharacteristic of NAT and GHG (Figure 2.2d).

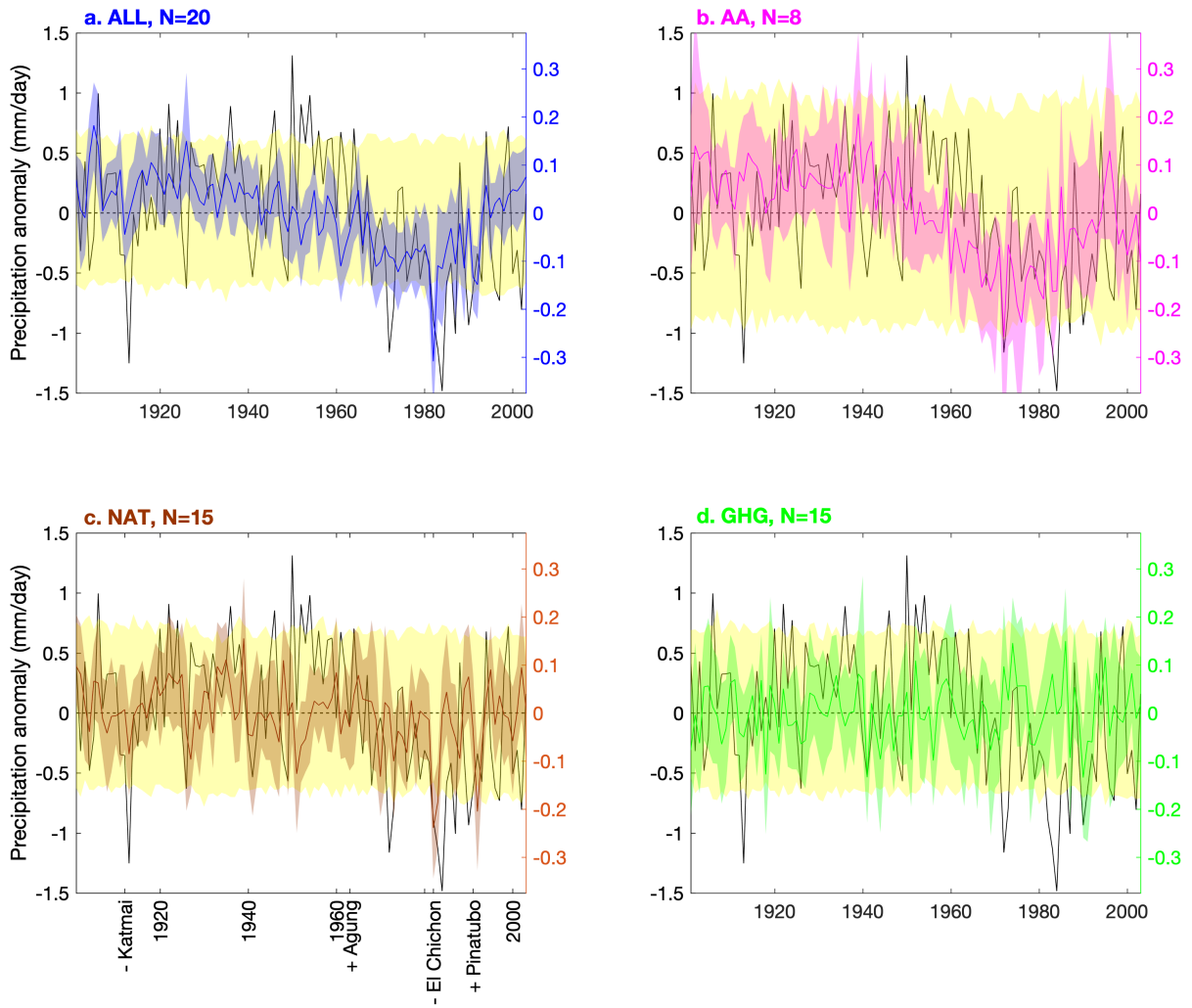

Figure 2.2: Forced MMMs: Forced MMM Sahel precipitation anomalies (colored lines; right, colored ordinates) and their yearly 95% confidence intervals from bootstrapping (colored shaded areas; right, colored ordinates) over observed Sahel precipitation anomalies (black lines; left, black ordinates) and the 95% confidence interval of the piC runs from randomized bootstrapping (yellow shaded areas; right, colored ordinates). N are the number of research institutions which performed each forcing experiment. Panel (c) additionally identifies the dates of large volcanic eruptions which had different effects on the aerosol optical depth in the northern and the southern hemispheres, as well as the sign of that difference (Haywood et al. 2013).

Figure 2.3 displays the mean padded power spectra (PS, lines) and 95% confidence intervals (shaded areas) of the bootstrapped forced MMMs (colors other than yellow), and compares them to that of the randomized bootstrapped piC MMMs (yellow). We calculate the piC MMM using the reduced set of models that contributed the AA experiment. With only 8

contributing research institutions, the AA MMMs filter out less noise from modelled internal variability—and thus have more power at all frequencies—than the MMMs associated with the other experiments. Thus, using this reduced set of models provides a conservative estimate of the spectral noise in all four forcing experiments. Figure 2.3 shows that the multi-decadal variability in AA (pink) and ALL (blue) is distinct from noise (yellow). It also confirms that the high-frequency variability in GHG is consistent with noise. Episodic volcanic forcing should not give rise, per se, to spectral peaks, though the observed pattern of large eruptions at the beginning and at the end of the century (see Figure 2.2c) may induce some spectral power at multidecadal timescales. Since we do not detect any meaningful spectral peak in the NAT PS (brown) associated with solar variability at 11 years, we interpret the NAT MMM to be mostly the result of volcanic aerosols.

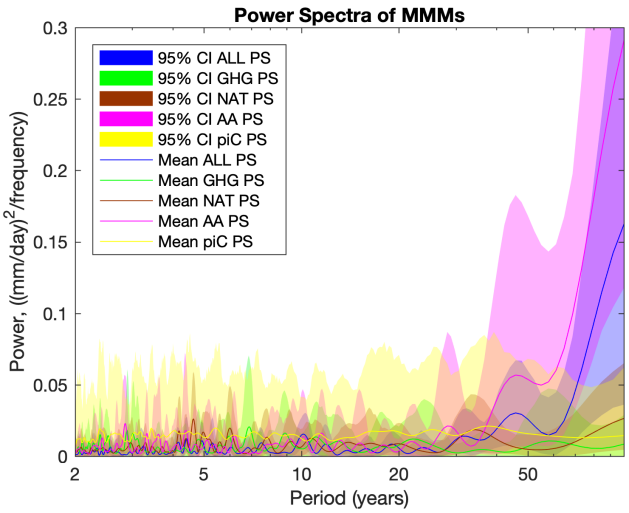

Figure 2.3: Forced MMM Power Spectra: Mean (lines) and 95% confidence intervals (shaded areas) of padded Power spectra (PS) of bootstrapped forced MMMs (ALL – blue, NAT – brown, AA – pink, GHG – green) and randomized bootstrapped AA piC MMMs (yellow).

Figure 2.4 again displays the values (dots) and PDFs (curves) of correlation (Figure 2.4a) and RMSE (Figure 2.4b) between observations and the bootstrapped ALL MMMs from Figure 2.1c and Figure 2.1d (blue) and compares them to the values (dots) and PDFs (curves) for

individual forcing experiments (solid curves distinguished by color) and the piC PDFs associated with the ALL experiment (dotted yellow curves). The piC PDFs corresponding to the three individual forcing experiments (which make use of only the models contributing to that experiment) are sufficiently similar to the ALL piC PDF that they are not plotted separately, with the exception of the AA piC sRMSE PDF (pink dotted curve in Figure 2.4b), which is wider and centered at a higher sRMSE than those of the other experiments, reflecting the high variance in the yearly values seen in the yellow shaded area in Figure 2.2b. Despite this difference, the p=.05 significance levels are still sufficiently similar for all four experiments for both correlation and sRMSE that they are represented by a single vertical grey dashed line at the p=.05 significance level of the ALL experiment. As the NAT and GHG MMMs contain mostly high-frequency variability – which is difficult to distinguish from noise remaining in the MMM (see Figure 2.3) – their PDFs are wider than the PDFs for the AA and ALL MMMs, which exhibit low-frequency variability uncharacteristic of noise in the MMM.

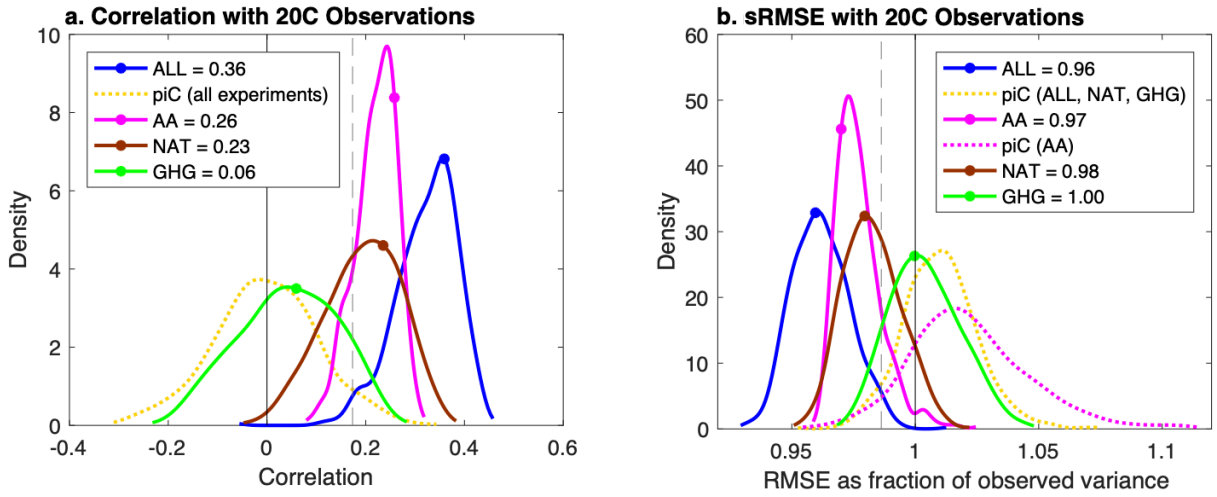

Figure 2.4: Performance of forced MMMs: Probability density function (PDF) of correlations (a) and sRMSE (b) of bootstrapped forced MMM 20th century Sahel precipitation (colored curves: blue = ALL, pink = AA, brown = NAT, green = GHG) and of randomized bootstrapped piC MMM Sahel precipitation corresponding to the ALL experiment (dotted yellow curves) and the AA experiment (dotted pink curve, b) with observed 20th century

Sahel precipitation. Actual forced MMM values are represented with colored dots on the PDFs. One-sided 95% confidence level represented with grey vertical dashed lines.

While the GHG MMM is not significantly better than noise at matching observations in both correlation (r=.06) and sRMSE (100% of observed variance), ALL (r=.36, sRMSE=.96), AA (r=.26, sRMSE=.97), and NAT (r=.23, sRMSE=.98) all achieve significance at p=.05. The discrepancy between the ALL MMM and NAT and AA individually under both metrics suggests that both anthropogenic and volcanic aerosols contribute substantially to the performance of the ALL MMM. Because the metrics for both AA and NAT fall within the other's bootstrapping confidence interval, according to this analysis, AA and NAT contribute roughly equally to the performance of the observed ALL MMM.

The ALL MMM has limited explanatory power as it is nearly constant, and, according to the RMSE, it leaves 96% of the variance unexplained. This unexplained variance could be due either to model deficiency or internal variability, since the MMM is designed to filter out internal variability—which will have similar characteristics but different phase across individual simulations—in favor of forced variability. Since observations include both internal and forced variability, no MMM would be able to match observations exactly. In this light, the ALL MMM correlation with observations of 0.36 is substantial. For comparison, we may liken this to simulations forced with observed SST, which reflect as best as possible observed internal climate variability as well as forced variability. As reported in Giannini et al. (2003), the correlation of the unsmoothed observations with the unsmoothed mean over version 1 of the atmospheric general circulation model developed at NASA's Goddard Space Flight Center in the framework of the Seasonal-to-Interannual Prediction Project (NSIPP1) from 1930-2000 is 0.60; the correlation of the ALL MMM with observations over the same period is not far behind at 0.47,

suggesting that a large fraction of the variability that SST-forced climate models can capture is externally forced.

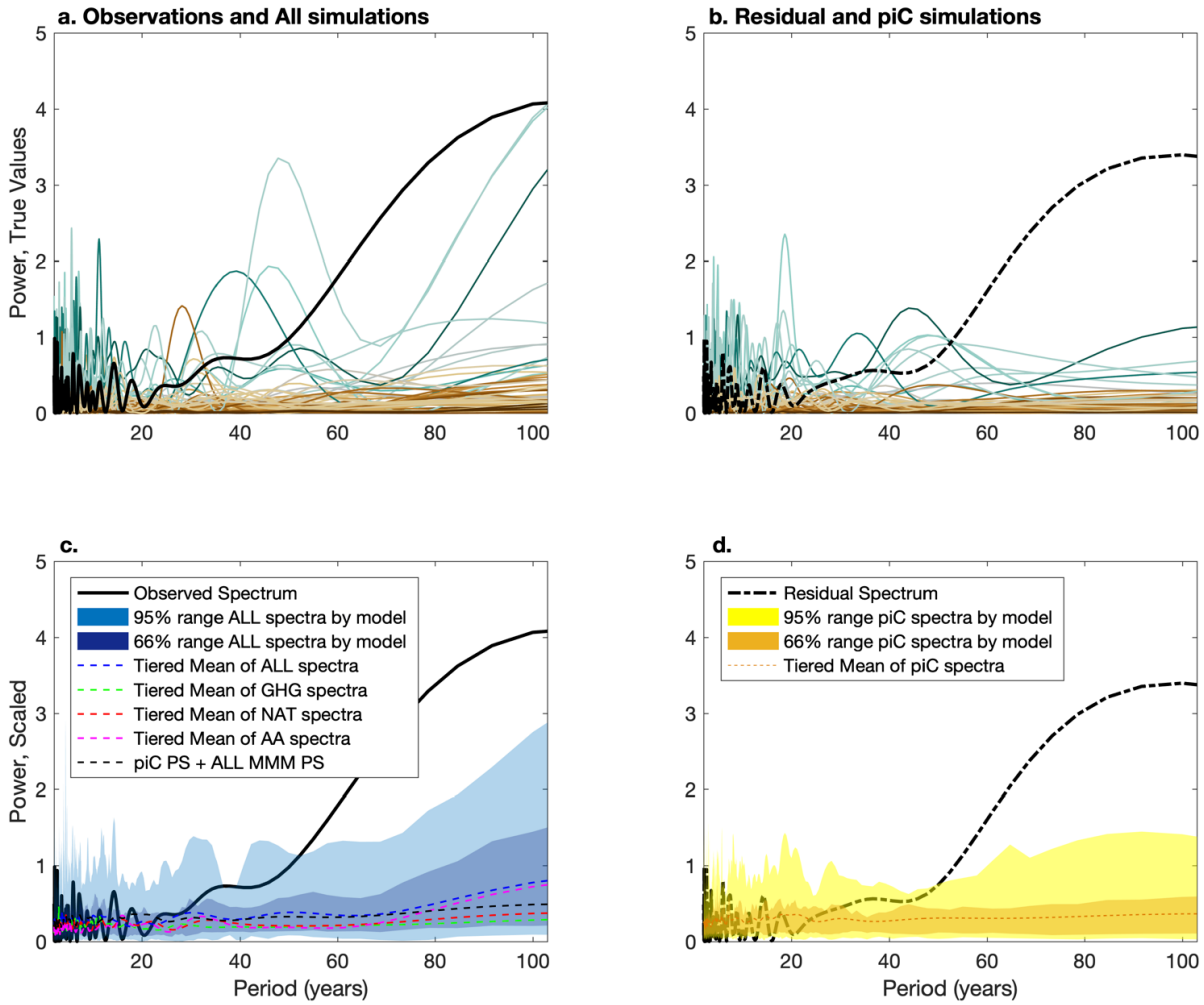

Figure 2.5: Residual Consistency: Power spectra (PS) of observed 20th century Sahel rainfall (solid black, a and c) and the residual after removing the ALL MMM (black dotted-dashed, b and d). (a) and (b): Mean PS by model of individual ALL (a) and piC (b) runs, colored by average JAS rainfall bias of the ALL runs compared to 20th century observations, where observed rainfall is grey, wet models are turquoise, and dry models are brown. piC PS (b) are additionally averaged over multiple subsections of the runs. (c): Tiered mean (blue dashed line) and 66% and 95% range (blue shading) of mean PS by model of individual ALL runs which were first rescaled to match 20th century observed JAS rainfall. Also displayed are the tiered means over PS of individual forced AA, NAT, and GHG runs (colored dashed lines). The black dashed line shows the sum of the tiered mean piC PS (from panel d) and the ALL MMM PS (i.e. Figure 2.3). (d): Tiered mean (orange dashed line) and 66% and 95% range (yellow shading) of mean PS by model of individual piC runs which were first rescaled so their corresponding ALL runs match 20th century observed yearly rainfall, as in (c).

2.3.3. Residual Consistency

To test the role of internal variability in the CMIP5 fully coupled models, we cannot use the MMM, because internal variability will have differing phase across different simulations. Instead, we examine power at different frequencies in individual coupled runs. Figure 2.5a compares the padded power spectrum (PS) of 20th century observed Sahel precipitation (solid black) to the padded PS of the ALL simulations, first estimated for individual runs, then averaged across ensemble members for each model. They are colored by the difference in the modelled and observed rainfall climatology from 1901 to 2003, where brown is used for models which are drier than observations, grey is used for models whose climatologies are near the observed climatology, and turquoise is used for models which are wetter than observations. As the individual ALL runs are single realizations that compound forced and internal variability like observations, they are directly comparable to observations.

While there are three models (MIROC-ESM p1, MIROC-ESM-CHEM p1, and GFDL-ESM2G p1) which nearly reach the high power of the observations at a period of 100 years, these models are biased wet, and also exhibit over-estimates of high-frequency variability. Figure 2.5b displays the PS of the estimate of observed internal variations implied by the MMM, calculated as the residual of observations with respect to the ALL MMM (black dashed-dotted line), and compares it to its modeled counterpart, estimated as the mean PS by model of the individual piC runs, colored by the same rainfall biases used in Figure 2.5a. Since there is often only one piC simulation per model, in order to reduce uncertainty in the PS, the long piC runs are divided into continuous, non-overlapping sections, and PS are taken separately for each section and then averaged together. We again see that wet models overestimate high-frequency variability, and no model matches the low-frequency power of the residual, pointing to

inconsistency between model simulations, their MMM, and observations. If the models underestimate forced variability, or if the MMM underestimates the magnitude of the modelled forced variability, this will cause the estimate of observed internal variability to be too large; so, while this comparison allows us to make a statement about consistency, it does not determine whether it is simulated internal variability or our estimate of forced variability that is incorrect. However, it is clear that modelled internal variability does not contribute substantial power at low frequencies.

The PS for both the forced and piC runs are clearly stratified by modelled precipitation climatology, or the mean JAS precipitation over the length of that simulation. To investigate whether any of the models capture the observed distribution of power across different frequencies, in Figure 2.5c and Figure 2.5d we rescale the simulations by model before taking the PS and the mean by model so that the climatology of each model's ALL simulations matches observed rainfall climatology. This mostly destroys the stratification in the previous panels (see Fig. A.1). The distribution of model-mean scaled ALL and piC PS are represented by blue and yellow shaded areas in Figure 2.5c and Figure 2.5d, respectively. The blue and orange dashed lines in Figure 2.5c and Figure 2.5d mark the centers of these distributions with 3-tiered, unweighted means over the PS of the ALL and piC runs, respectively. The other colored dashed lines in Figure 2.5c mark the tiered means over the PS of all runs in each of the three individual forcing experiments (magenta=AA, brown=NAT, green=GHG) for comparison.

The black dashed line in Figure 2.5c shows the sum of the tiered mean piC PS (orange dashed line from Figure 2.5d) and the PS of the ALL MMM (i.e. the blue line in Figure 2.3). If the MMM accurately represented the simulated forced power when scaled to the observed climatology, we would expect this sum to match the tiered mean ALL PS (blue dashed line).

Instead, it falls short at low frequency, suggesting that the ratio of the variance of the ALL MMM to observed climatology underestimates the ratios of simulated forced variance to climatological Sahelian precipitation in CMIP5 models. This may be because the ensemble is biased dry, or because differing responses to forcing between models cause the consensus forced response to have lower variance than exhibited in individual models. In addition to any implications for the sRMSE calculations displayed earlier, this means that the residual spectrum in Figure 2.5d is an overestimate of internal variability in observations as implied by the CMIP5 ensemble.

However, it is still clear that even scaled piC simulations do not exhibit any increase in power at low frequency (Figure 2.5d). Even though the inclusion of external forcing introduces low-frequency variance (Figure 2.5c), the CMIP5 models are unable to capture the scale of the increase in power at low frequency in the observed PS, which exceeds the 95th percentile of rescaled ALL PS at periods longer than 50 years. Of the different forced experiments, ALL and AA are the only ones that exhibit substantial multi-decadal variability. Thus, while the variance of the ALL MMM may be somewhat underestimated due to the dry bias of the ensemble or to attenuation from averaging, the vast majority of the discrepancy in low-frequency power between simulations and observations is due to model deficiency, whether in capturing the full magnitude of the forced response to AA, or in detailing the true character and magnitude of the other forced responses, low-frequency internal variability, and their interactions.

2.4. Discussion

The analysis in this study shows that the consensus response of Sahelian precipitation to 20th century external forcing in CMIP5 simulations, as defined by the 3-tiered multi-model mean (MMM), correlates significantly with observations. It further shows that *both* anthropogenic

aerosols (AA) and volcanic aerosols (NAT) contribute significantly and substantially to making CMIP5 MMM Sahel precipitation similar to observations, with AA mostly responsible for the multidecadal forced variability. Given that the performance of the ALL MMM can apparently be explained with AA and NAT alone, we conclude that GHG do not contribute to the consensus forced response of Sahel seasonal precipitation in CMIP5 models.

This does not mean that GHG do not influence Sahelian precipitation in any way, or that GHG will not play a significant role in the future as the magnitude of the forcing increases. While some individual models have indicated a role for GHG in the recovery since the mid 1990s (Dong and Sutton 2015), it is possible that the models as an ensemble do not yet capture the effects of GHG on Sahelian rainfall because the magnitude of the forcing is still too small over the historical period. Alternatively, competition between the mechanisms linking GHG forcing to Sahelian rainfall may have masked the effects of GHG by cancelling within individual simulations (Kawase et al. 2010) or between models (Biasutti 2013) in the MMM. Finally, it has been suggested that the response to GHG is inherently non-linear (e.g. different circulation responses to different magnitudes of warming in Neupane and Cook 2013), or interacts non-linearly with other forcings (e.g. the interaction of an "upped ante" and changing moisture supply, as suggested by Giannini and Kaplan 2019). These non-linearities are difficult to test without the ability to compare the ALL MMM to "all but GHG" simulations, which are not widely available in CMIP5.

While the standardized root mean squared error (sRMSE) of the ALL MMM with observations is also significantly different from noise, it is 96% of the observed rainfall variance, meaning that modelled forced variability can hardly account for observed variability since the ALL MMM is hardly better than a constant prediction. Our residual consistency test showed that

while the variance of the MMM may be too small relative to observed climatological seasonal-mean rainfall due to the dry-bias of simulations or to averaging, the discrepancy between total observed variability and total modelled variability is an order of magnitude larger than this difference, and modelled internal variability cannot account for the difference between the simulated forced response and observations.

Since modelled internal variability does not show substantial low-frequency variability while the AA MMM does, it is tempting to attribute the full magnitude of observed multi-decadal variability to AA, as many previous studies have done by focusing only on standardized trends (Held et al. 2005), correlations (Giannini and Kaplan 2019), or detectability in a fingerprinting framework (Polson et al. 2014; Undorf et al. 2018). However, such a claim would rely heavily on assumed grid-point linearity of the climate response to different forcings, which is disputed for tropical rainfall (i.e. Giannini and Kaplan 2019 on GHG and anthropogenic aerosols; Lohmann and Feichter 2005 on feedbacks involving the indirect aerosol effects; Meehl et al. 2003 on non-linear feedbacks between solar forcing and GHG; Neupane and Cook 2013 on GHG-induced circulation changes over Africa; and Polson et al. 2014 on the indirect aerosol effect and spatial trend patterns in the Asian Monsoon), as well as on the accuracy of simulated forced and internal variability. In fact, it is not possible to say without further investigation into the physical pathways influencing Sahelian precipitation whether the model deficiency is in the modelled response to forcing or in modelled internal variability. Given the strong link between Sahelian rainfall and North Atlantic SST (Ackerley et al. 2011; Giannini and Kaplan 2019; Giannini et al. 2013; Hoerling et al. 2006; Martin et al. 2014), it is perhaps not a coincidence that models lack strong low-frequency variability both in Sahel rainfall and in internally-generated Atlantic Multidecadal Variability in SST (AMV; Yan et al. 2018). The community is currently

still debating whether the observed AMV is forced by AA (Booth et al. 2012; Chang et al. 2011; Rotstayn and Lohmann 2002) or is an internal phenomenon which is linked to ocean circulation variability (Sutton and Hodson 2005; Yan et al. 2019; Zhang 2017; Zhang et al. 2016; Zhang et al. 2013) and is dramatically underestimated in most models (Yan et al. 2018).

Future work that focuses on characterizing and quantifying the mechanisms of influence on Sahelian precipitation in simulations and observations and using the next generation of climate models (Eyring et al. 2016) might shed new light on whether the model/observation discrepancy documented here is due to an underestimate in the strength of the precipitation response to AA or a failure of CMIP5-class climate models to capture low-frequency internal variability.

Chapter 3. Deficiencies in Simulated Low-Frequency Sahel Precipitation Variability from CMIP5 and CMIP6

Note: This chapter has been published in very near its present form as "Drivers of Low-Frequency Sahel Precipitation Variability: Comparing CMIP5 and CMIP6 Ensemble Means with Observations" in *Climate Dynamics* (2023), doi: 10.1007/s00382-023-06755-1.[7]. Minor edits have been made for clarity.

3.1. Introduction

Chapter 2 investigated the causes of observed Sahel precipitation variability using the Coupled Model Intercomparison Project phase 5 (CMIP5, Taylor et al. 2012), which is the first multi-model ensemble to include "Detection-Attribution" simulations aimed at separating the effects of the radiative forcing agents mentioned above. Rather than focusing on the specific behavior of any individual model—which may outperform other models in one metric but underperform in another—we employed a multi-model mean (MMM) to identify a more robust "central tendency" of the ensemble (Mote et al. 2011). We showed that, while the MMM variance might be somewhat damped due to the dry-bias of the ensemble, the MMM of historical Sahel precipitation matches observed historical precipitation better than simulated precipitation from any individual model in CMIP5. We found that AA and volcanic aerosols—but not GHG—are responsible for forcing simulated Sahelian precipitation that correlates well with

---

[7] AUTHORS: Rebecca Jean Herman[ab]*, Michela Biasutti[c], and Yochanan Kushnir[c]

[a] Deutsches Zentrum für Luft- und Raumfahrt, *Jena, Germany*
[b] Department of Earth and Environmental Sciences of Columbia University, *New York, NY*
[c] Lamont-Doherty Earth Observatory of Columbia University, *Palisades, NY*

* Corresponding author: Rebecca Jean Herman, rebecca.herman@columbia.edu

observations, with AA alone responsible for the low-frequency component of simulated variability. This conclusion appeared consistent with previous claims that global AA emissions—which originate mostly in the Northern Hemisphere, and which increased until the 1970s and then decreased until 2000 in response to clean air initiatives (Klimont et al. 2013; Smith et al. 2011)—caused multi-decadal variability in Sahel precipitation via changes in Northern Hemisphere surface temperature (Ackerley et al. 2011; Haywood et al. 2013; Hwang et al. 2013b; Undorf et al. 2018), or specifically via multidecadal variability in North Atlantic SST (the Atlantic Multidecadal Variability, AMV; Booth et al. 2012; Hua et al. 2019). However, we also found that, like previous simulations, the CMIP5-simulated rainfall response to forcing (given by the MMM) has too little low-frequency power relative to observations, and simulated internal variability is unable to account for this difference.

Because we did not examine the pathways through which any given forcing agent affects Sahel precipitation, we could not determine whether the discrepancy between the CMIP5 simulations and observations represents a *quantitative* underestimate of aerosol indirect effects and related feedbacks, or a *qualitative* inability of the models to simulate the observed climate response to forcing or observed modes of internal variability. The distinction is crucial if we are to trust these models' projections of future Sahelian rainfall.

A first investigation into the mechanisms of forced Sahel precipitation variability should focus on sea surface temperature (SST). Early Sahel climate variability research (Folland et al. 1986; Giannini et al. 2003; Knight et al. 2006; Palmer 1986; Zhang and Delworth 2006) and more recent studies (Okonkwo et al. 2015; Parhi et al. 2016; Park et al. 2016; Pomposi et al. 2015; Pomposi et al. 2016; Rodríguez-Fonseca et al. 2015 and references therein) used atmospheric simulations with prescribed global SST to demonstrate causal "teleconnections"

between remote SST in a number of ocean basins and Sahel precipitation, and used observational statistics to show the importance of these teleconnections. Unfortunately, like coupled simulations, atmospheric simulations have also historically failed to explain the full magnitude of observed Sahel precipitation variability (e.g. Scaife et al. 2009). Thus, while the dominant role of global SST over local land-use practices in driving the pacing of observed 20th century Sahel rainfall variability is unquestioned (Biasutti 2019), it has not yet been possible to rule out simultaneous (and potentially important) roles for solely-atmospheric responses to global radiative forcing that are in-phase with observed SST variability.

Since publishing Chapter 2, new coupled and atmosphere-only simulations became available from phase 6 of the Coupled Model Intercomparison Project (CMIP6, Eyring et al. 2016), in which many models have an improved representation of physical processes and newly implemented biogeochemical cycles as well as higher resolution (Masson-Delmotte et al. 2021). In coupled models, these changes have reportedly resulted in greater (and perhaps unrealistic) sensitivity to GHG (Forster et al. 2020; Zelinka et al. 2020) but little change in the representation of summer monsoons (Fiedler et al. 2020), though in this paper we identify and explore some notable differences in the historical evolution of Sahel rainfall. Atmosphere-only simulations in CMIP5 used both prescribed SST and remote radiative forcing simultaneously, and only covered the last couple decades of the 20th century, making it previously impossible to examine multi-decadal SST-forced variability using CMIP. CMIP6 for the first time includes an ensemble of atmosphere-only ("AMIP") simulations long enough to capture the observed period of multi-decadal Sahel precipitation variability, with and without simultaneous radiative forcing. This new ensemble will allow us to robustly analyze the simulated response to realistic SST in CMIP and in observations.

Here, we extend the methodology of Chapter 2 by using the well-established dependence of Sahel precipitation on global SST in combination with the CMIP6 atmosphere-only simulations to examine the mechanisms of Sahel precipitation variability in observations and in the CMIP ensembles in more depth. We decompose the simulated effects of individual external forcing agents (F) and internal climate variability on low-frequency Sahel precipitation variability (P) into five path components, presented in Figure 3.1: (1) teleconnections that communicate variations in SST to variations in precipitation (indicated by the arrow $\vec{t}$); (2) the "fast" atmospheric and land-mediated effect of external forcing (F) on precipitation ($\vec{f}$); (3) the expression of atmospheric internal variability (IV_a) as precipitation variability ($\vec{a}$); (4) the effect of forcing on SST ($\vec{s}$); and (5) the expression of internal variability in the coupled climate system (IV_o) as SST variability ($\vec{o}$). The path F → SST → P is the "slow," SST-mediated effect of forcing on precipitation. While the atmospheric response to forcing is not more rapid than teleconnections or atmospheric internal variability, we prefer to label this term the "fast response to forcing" rather than employing the other commonly used label—the "direct response"—in order to avoid terminology confusion with the "direct" effect of aerosols (which contributes to the "fast response" to forcing in combination with the "indirect effect" of aerosols).

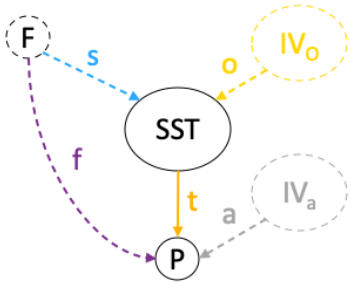

Figure 3.1: Causal diagram relating external forcings (F), atmospheric (IV_a) and oceanic internal variability (IV_o), sea surface temperatures (SST), and Sahel precipitation (P) via directional causal arrows. Unobserved variables and their causal effects are presented with dashed lines, while observed variables are presented with solid lines.

To this point, characterization of these path components in observations remains controversial. Firstly (top of diagram), separating the SST response to forcing ($\vec{s}$) from SST variability internal to the climate system ($\vec{o}$) has proven difficult. In particular, there is significant debate over whether observed AMV is a response to external forcing (Booth et al. 2012; Chang et al. 2011; Hua et al. 2019; Menary et al. 2020; Rotstayn and Lohmann 2002) or mainly an expression of internal variability in the Atlantic Meridional Overturning Circulation (AMOC, Han et al. 2016; Knight et al. 2005; Qin et al. 2020; Rahmstorf et al. 2015; Sutton and Hodson 2005; Ting et al. 2009; Yan et al. 2019; Zhang 2017; Zhang et al. 2016; Zhang et al. 2013) that is underestimated in models (Yan et al. 2018). This debate has been hard to resolve partially because internal variability in AMOC and aerosol forcing may have coincided by chance in the 20th century (Qin et al. 2020). Next, examine the bottom of the diagram. The effect of the observed SST field on Sahel precipitation ($\vec{t}$) can be directly estimated using atmosphere-only simulations, but while these simulations capture the pattern of observed Sahel precipitation variability, many fail to capture its full magnitude (Biasutti 2019; e.g. Hoerling et al. 2006; Scaife et al. 2009). This could reflect an underestimate in climate models of the strength of SST teleconnections ($\vec{t}$), which could be resolution dependent (Vellinga et al. 2016); or of land-climate feedbacks that amplify the teleconnections, such as vegetation changes (Kucharski et al. 2013). But it could also reflect a significant additional role in the observations for a fast response to forcing ($\vec{f}$) that drives (Bretherton and Battisti 2000) and confounds the SST-forced signal [P ← F → SST → P; see Pearl et al. (2016) for notation] or coincides with it by chance.

We will not be able to resolve all of these uncertainties, but if coupled simulations or the new atmospheric simulations from CMIP6 can explain either precipitation or SST variability in observations, then this exercise should simultaneously shed light on the drivers of observed

variability (where the forced response and simulated internal variability jointly explain the observations) and the realms of common shortcomings of CMIP models (where simulations fail to explain observations). We continue to use the MMM to understand the forced response in both observations and simulations. While model improvement efforts also benefit from model-specific analysis, we believe it is most urgent to identify and address model biases that are common across Model Intercomparison Projects because they prevent the ensemble from giving accurate mean and uncertainty estimates in future projections.

To examine the path components in coupled simulations, we need a parsimonious characterization of the relationship between SST and Sahel precipitation. The scientific literature has linked Sahelian precipitation to many ocean basins, including the equatorial Pacific Ocean, the Indian Ocean, the North and South Atlantic Oceans, and the Mediterranean Sea (Rodríguez-Fonseca et al. 2015). Observed Mediterranean SST correlates with observed Sahel precipitation mainly at high frequencies (not shown), which are not the focus of this study. Giannini et al. (2013) summarize the influence of the other relevant ocean basins in a single index, defined as the difference between average SST in the North Atlantic (NA) and in the Global Tropics (GT), and termed the North Atlantic Relative Index (NARI). Warming of NA is argued to increase the moisture supply to the Sahel, destabilizing the column from the bottom up (Giannini and Kaplan 2019), while warming of GT is expected to stabilize the column by causing upper tropospheric warming throughout the tropics (Chou and Neelin 2004; Giannini 2010). Thus, an increase in NARI is expected to wet the Sahel while a decrease is associated with drying. Giannini and Kaplan (2019, hereafter GK19) identify NARI as the dominant SST indicator of 20th century Sahel rainfall in observations and CMIP5 simulations. Others have preferred to summarize the teleconnected SST as an Interhemispheric Temperature Difference, linking Sahel rainfall to

energetically-driven shifts in the Intertropical Convergence Zone (Donohoe et al. 2013; Kang et al. 2009; Kang et al. 2008; Knight et al. 2006; Schneider et al. 2014). The two indices are sufficiently correlated at the low frequencies of interest here (r=0.89 in observations, .79 in CMIP5, and .98 in CMIP6 for periods of >20 years; see Figure 3.4 and relevant discussion for motivation of this choice) that it is difficult to separate their potential effects on Sahel precipitation. We prefer NARI because the mechanisms associated with it are predictive rather than diagnostic (but see Section 3.5 for a discussion of the choice). Here, we choose to use NARI together with the assumption of linearity to approximate the full slow response as the product of the NARI response to external forcing and the strength of the NARI-Sahel teleconnection.

This paper is organized as follows: Section 3.2 provides details on the simulations and observational data used in this analysis while Section 3.3 discusses the methods. In Section 3.4.1, we update Chapter 2's analysis to CMIP6, examining the total response to forcing (all paths from F to P) and internal variability (all paths from IV to P). We then evaluate the performance of the CMIP6 AMIP simulations, decomposing them into the path components from the bottom half of Figure 3.1 ($\vec{t}$, $\vec{f}$, and $\vec{a}$) in Section 3.4.2, and focusing on the NARI teleconnection in Section 3.4.3. Section 3.4.4 decomposes coupled simulations of NARI into the path components from the top half of Figure 3.1 ($\vec{s}$ and $\vec{o}$), while Section 3.4.5 evaluates the consistency of the NARI teleconnection established in Section 3.4.3 with coupled simulations. Finally, in Section 3.4.6, we use simulated NARI and the simulated NARI teleconnection to decompose the total response of Sahel precipitation to external forcing in coupled simulations (examined in Section 3.4.1) into NARI-mediated and residual components. We discuss whether the residual is consistent with a fast response in Section 3.5 before concluding in Section 3.6.

3.2. Data

We examine 20th century Coupled Model Intercomparison Project simulations from CMIP5 (Taylor et al. 2012) and CMIP6 (Eyring et al. 2016), including "historical" simulations forced with all sources of external radiative forcing (ALL) and "pre-Industrial control" (piC) simulations in which all external forcing agents are held constant at pre-Industrial levels. We additionally examine "Detection and Attribution" simulations (Gillett et al. 2016) forced with AA alone, natural forcing alone (NAT, which includes volcanic aerosols as well as solar and orbital forcings), and GHG alone. Finally, we also examine CMIP6 amip-piForcing (amip-piF) simulations (Webb et al. 2017), in which atmospheric models are forced solely with observed SST, and CMIP6 amip-hist simulations (Zhou et al. 2016), which are forced with observed SST and historical ALL radiative forcing. Our calculations begin in 1901 and extend to the end of the simulated period: 2003 for CMIP5 and 2014 for CMIP6.

In Chapter 2, we used all models available through the International Research Institute (IRI) and Lamont-Doherty Earth Observatory (LDEO) Climate Data Library for each forcing subset. Here, in order to provide a more stringent comparison of the effects of different forcing agents, we exclude models from the coupled ensemble if no AA, GHG, or ALL simulations for that model (or a related model from the same institution) were available on the IRI/LDEO Climate Data Library (for CMIP5) or on Pangeo's CMIP6 Google Cloud Collection (for CMIP6) as of May 4, 2022. For the AMIP ensemble, all existing CMIP6 amip-piForcing simulations were downloaded from the Earth System Grid Federation website, and models were excluded if there was not at least one available amip-piForcing simulation or if there was no amip-hist simulation available on Pangeo's CMIP6 Google Cloud Collection. Tbl. B.1, Tbl. B.2, and Tbl. B.3 enumerate the simulations used in this analysis.

Precipitation observations are from the Global Precipitation Climatology Center (GPCC, Becker et al. 2013) version2018, and SST observations are from the National Oceanic and Atmospheric Administration's (NOAA) Extended Reconstructed Sea Surface Temperature, Version 5 (ERSSTv5, Huang et al. 2017).

We analyze precipitation over the Sahel (12°-18°N and 20°W-40°E), and the SST indices of GK19: the North Atlantic (NA, 10°-40°N and 75°-15°W), the Global Tropics (GT, ocean surface in the latitude band 20°S-20°N), and the North Atlantic Relative index (NARI, the difference between NA and GT). All indices are spatially- and seasonally-averaged for July-September.

3.3. Methods

In defining the causal diagram in Figure 3.1, we have explicitly assumed that external forcing does not affect internal variability. This may not be completely justified (e.g. Ham 2017), but is standard practice in model-based attribution methods that go beyond simple comparison. Such methods typically further assume (implicitly or explicitly) that observations and simulations are the sum of externally forced climate signals and independent internal climate variability. This allows the researcher to define the simulated response to forcing and to internal variability using an ensemble of simulations with differing initial conditions (Hegerl and Zwiers 2011). Our formulation additionally assumes that precipitation variability does not affect simultaneous or following atmospheric or oceanic internal or forced variability – a simplification that precludes the possibility of vegetation feedbacks. While these might be relevant for the real world, we assume they are of secondary importance in the CMIP ensembles analyzed here, which do not include dynamic vegetation.

We define the response of a climate variable to a set of radiative forcing agents as the multi-model mean (MMM) of that variable over a set of historical simulations which prescribe that forcing agent. The MMM filters out intermodel differences and internal variability (which is independent across simulations), leaving an ensemble consensus response to forcing (which is dependent on the common radiative forcing applied to every model). We calculate the MMM as a 3-tiered weighted average over (1) individual simulations (runs) from each model, (2) models from each research institution, forming the Institution Mean (IM), and (3) institutions in that ensemble. Details of the weighting are provided in Chapter 2; the results are robust to differences in weighting. Time series are not detrended, and anomalies are calculated relative to the period 1901-1920 unless otherwise noted.

To evaluate the performance of the simulations relative to observations, we compute correlations (r), which capture similarity in frequency and phase, and root mean squared errors standardized by the standard deviation of observations (sRMSE), which measure yearly differences in magnitude between the simulated MMM and observations. The MMM time series are usually smoothed with a 20-year lowpass filter before calculating these statistics, and this is notated $(\bullet)_{LF}$. We chose to divide high- and low-frequencies at a period of 20 years because that is the border between low- and high-power variance in observations (see Figure 3.4). An sRMSE of 0 represents a perfect match between the simulated MMM and observations, and 1 would result from comparing the observations with a constant time series.

To gauge uncertainty in the smoothed MMM estimates of the forced signals and associated metrics, we resample the institution means (output of tier two) with replacement to calculate a set of perturbed MMMs (details in Chapter 2), yielding probability distribution functions (PDF) of the MMMs and of the values of each metric. Due to the finite number of

simulations, these PDFs underestimate the true magnitude of the uncertainty. We evaluate significance by applying a randomized bootstrapping technique, which increases the effective sample size, to the smoothed piC simulations with one significant improvement over Chapter 2: instead of using just one subset of each piC simulation at a random offset to calculate the model means (tier 1 of the MMM) in each bootstrapping iteration, we take enough subsets to match the number of that model's historical runs. Done this way, the confidence intervals calculated using piC simulations accurately represent noise in the forced MMMs. PiC PDFs from the same ensemble associated with different experiments differ slightly because a different number of simulations are available for different subsets of forcing agents (see Tbl. B.2).

We perform a residual consistency test, which compares the power spectra (PS) of individual simulations (not smoothed) to that of observations, with one significant modification over Chapter 2: we calculate the PS using the multi-taper method. Confidence intervals for the PS for observations and MMMs are given by the multi-taper method, without accounting for the uncertainty in the MMMs themselves. Mean PS by model are pictured with colors according to climatological rainfall bias given by those simulations. The multi-model mean of these PS is calculated using the three tiers from the definition of the MMM, but without weights, since spectral power is not attenuated when averaging PS. The power spectrum of observed Sahel rainfall is used to justify the choice of 20 years for the low-pass filter.

3.4. Results

3.4.1. Changes in CMIP6: Total Precipitation Response to Forcing and Internal Variability

In this section, we compare observed Sahelian precipitation to simulated forced and internal precipitation variability.

We begin by examining forced variability. The Multi-Model Mean (MMM) over coupled simulations filters out atmospheric and oceanic internal variability ($\vec{a}$ and $\vec{o}$), leaving the fast and slow precipitation responses to external radiative forcing ($\vec{f}$ and F → SST → P; see Section 3.3 for a discussion of the MMM). Figure 3.2 compares observed Sahelian precipitation anomalies to the simulated response to four sets of forcing agents (colors) in CMIP5 (dotted curves) and CMIP6 (solid curves). To highlight the primarily low-frequency simulated precipitation response to slowly-varying anthropogenic emissions, the anthropogenic aerosols b, "AA," magenta) and greenhouse gases (d, "GHG," green) MMMs are smoothed with a 20-year lowpass filter and contrasted with smoothed observations (black). Natural forcing (c, "NAT," brown and red), on the other hand, is dominated by sporadic and non-periodic short-lived volcanic episodes, some of which are highlighted on the x ordinates. Spectral decomposition of any kind assumes that all components of the time series are periodic, so the simulated and observed responses to volcanic eruptions will necessarily be split between the apparent "high-frequency" and "low-frequency" components. Because the episodes are short-lived, we find it visually more helpful to compare the NAT MMMs to high frequency observed precipitation variability (grey), calculated by subtracting the low-pass filtered time series from the full time series. ALL simulations with all three forcing agents (a, blue) include both episodic and low-frequency forced variability, and so their MMMs are visually compared to the full observed precipitation variability (light grey) in addition to the smoothed version. The figure also presents the bootstrapping 95% confidence intervals of the forced MMMs (blue, magenta, red/brown, and green shaded areas) and of MMMs over randomly-shifted CMIP5 and CMIP6 preindustrial control (piC) simulations (dotted and solid black lines, respectively). These dotted and dashed lines represent the magnitude of noise deriving from coincident internal variability in the MMMs; differences between panels

arise from varying numbers of simulations for the different forcing subsets (see Section 3.3: Methods and Tbl. B.2) and from smoothing in some panels.

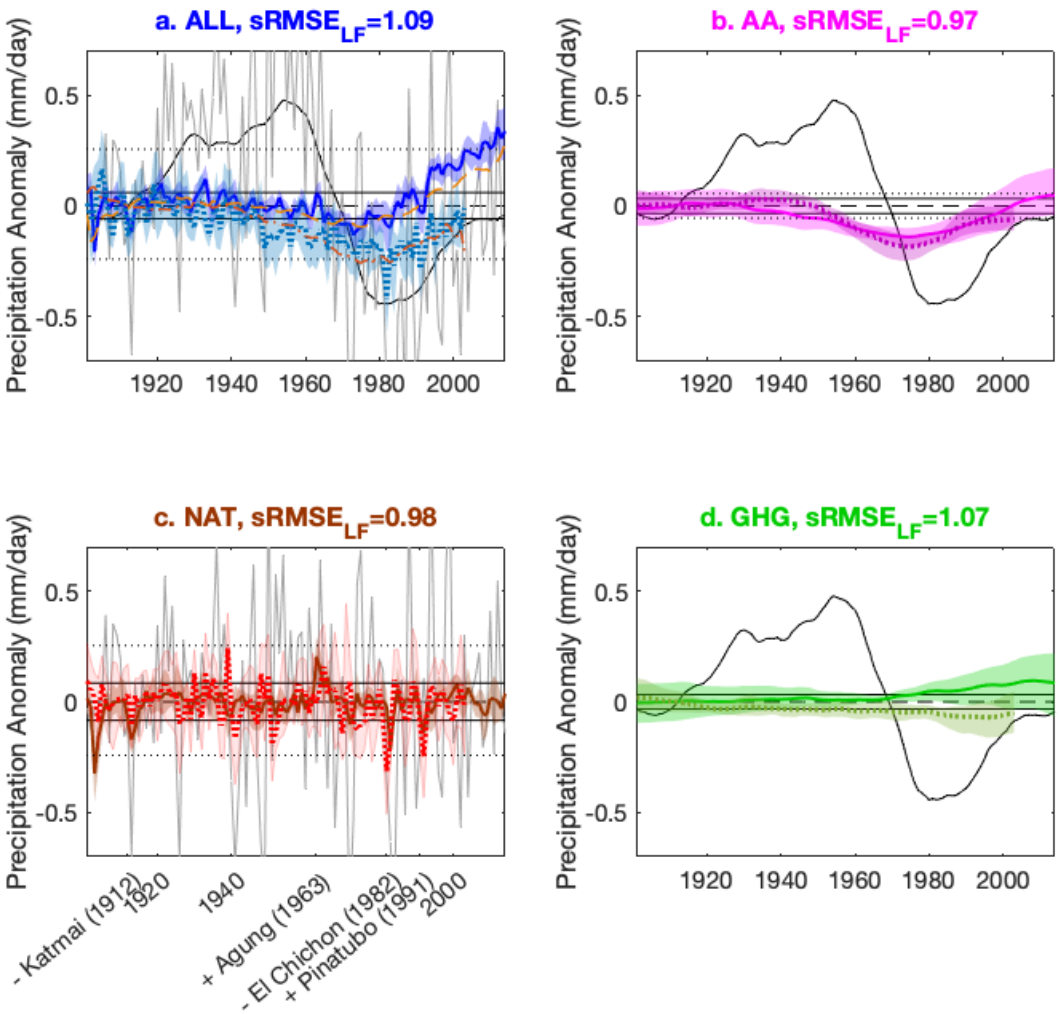

Figure 3.2: Observed high (grey) and low-frequency (black) Sahelian precipitation anomalies and MMM simulated precipitation anomalies (colored) from CMIP5 (dotted bold curves) and CMIP6 (solid bold curves) forced with ALL (a, blue), AA (b, magenta), NAT (c, brown/red), and GHG (d, green). The colored shaded areas surrounding the MMMs denote the bootstrapping confidence intervals, and the horizontal black lines mark the confidence intervals of randomized bootstrapped MMMs from CMIP5 (dotted) and CMIP6 (solid) piC simulations, representing the magnitude of noise in the MMMs. For AA (b) and GHG (d), which cause low-frequency precipitation variability, simulations are smoothed over 20 years before taking the MMM and are visually compared to smoothed observations. Because volcanic forcing in NAT (c) causes short-lived episodic precipitation variability, we present observed high-frequency precipitation variability in grey. We also make note of hemispherically asymmetric volcanic forcing from Haywood et al on the x ordinates, where a negative sign denotes an eruption that cooled the northern hemisphere more than the

southern hemisphere while a positive sign denotes the opposite, aligning with the sign of the expected Sahel precipitation response to the eruption. The ALL MMM (a) includes both episodic and low-frequency forced variability, and so we present the full observed precipitation variability (light grey) in addition to the smoothed version (black). This panel additionally shows the sum of the smoothed AA and GHG MMMs for CMIP5 (auburn dotted-dashed curve) and CMIP6 (amber dashed curve). The label shows the standardized root mean squared error of the CMIP6 MMM with observations at low-frequency ($sRMSE_{LF}$).

Though we sometimes visually present "high-frequency" variability to clarify the impact of volcanic eruptions on total as well as apparent "low-frequency" variability, only low-frequency (LF) correlation and sRMSE statistics are presented. This means that the statistic was calculated after smoothing the MMM and observations, and the confidence interval and significance level are calculated by repeating the statistic on the smoothed bootstrapped MMMs of forced and piC simulations, respectively. Significance in the context of this paper means that the performance of the forced response is separable from unforced noise in the performance statistic.

The CMIP6 MMM is less successful than CMIP5 at producing a forced dry interval in the second half of the century that compares to the observed drought in the 1970s and 80s (panel a). To understand why, we examine the individual forcing experiments. In the AA experiments (b), CMIP6 looks similar to CMIP5—precipitation declines in the mid-century and then recovers in response to clean air initiatives, preceding the timing of observed variability by about 10 years—but the drought is not quite as strong. There are some differences in the NAT experiments between CMIP5 and CMIP6 (c), but the largest MMM variations in both ensembles are interannual episodes that are clearly associated with volcanic eruptions, most notably following El Chichón in 1982. In the GHG experiments (d), CMIP6 shows anomalous wetting after 1970 that wasn't present in CMIP5.

The changes in the single forcing experiments are reflected in the ALL MMM (a): while CMIP5 reaches peak drought in 1982 – close to the observed precipitation minimum – CMIP6 dries very little and only until 1970, after which it displays an anomalously wetter climate than CMIP5 through the end of the century. But while the precipitation responses to different forcing agents appear to add linearly in CMIP5 (compare the auburn dotted-dashed curve to the blue dotted curve), the late century wetting in CMIP6 is larger than the sum of GHG and AA wetting (amber dashed curve; including NAT does not help). This effect is robust to differences in model availability for the different sets of forcing agents. Thus, in the ALL MMM, CMIP6 displays slightly less drying from AA compared to CMIP5, more wetting from GHG, and additional wetting after 1990 from a non-linear interaction between forcings.

As a result of these changes, the response to forcing in CMIP6 is a poor match to observations. At low frequencies, the correlation between the CMIP6 ALL MMM and observations is statistically indistinguishable from 0, and, while the $sRMSE_{LF}$ between observations and the ALL MMM in CMIP5 (0.88 ± 0.04) is significantly better than a constant prediction (1) or noise in the MMM (not shown), the sRMSEs for all CMIP6 MMMs are equivalent to or worse than noise ($\geq$0.96, see panel titles). The poor correlation for CMIP6 makes it clear that amplifying the simulated precipitation response to forcing will not help explain observed precipitation.

In Chapter 2 we showed that the CMIP5 ALL MMM (dotted blue curve) better reproduces observed Sahelian precipitation than individual ALL institution means (IMs). This is not true for CMIP6, whose MMM performance is a closer match to the median of the Institution Mean (IM) performances (not pictured). Though it does not perform better than the IMs, the MMM seems an accurate representation of the behavior of the individual IMs and the ensemble

as a whole. In Figure 3.3, smoothed individual simulations are presented in grey, with 95% of the yearly precipitation anomalies falling within the dotted black curves. IMs are presented in cyan, and the institution that performs best in correlation and sRMSE (MRI) is presented in green. The smoothed MMM is presented in blue, and compared to smoothed observations (Obs), in black. The major characteristics of the MMM generally apply to the IMs: none of the IMs capture the observed pluvial, all of the IMs show wetting at the end of the century despite almost none producing any drought at all, and those IMs that do produce a late-century rainfall minimum reach minimum precipitation earlier than observations. MRI appears to outperform the other IMs in part because the drought is later than the others (but still earlier than observed), and also in part because the difference in magnitude between the late-century drying and wetting is smaller than the other IMs. But its performance statistics are not distinct from the performance distributions over all IMs in CMIP6 (not pictured), and it still differs strongly from observations, with no simulated pluvial and with a drought and a recovery that are weaker than observations and some (or most) of the other IMs. Thus, the discrepancy between the MMM and observations is much larger than the difference between models, and we continue to use the MMM to represent the response to forcing in CMIP6.

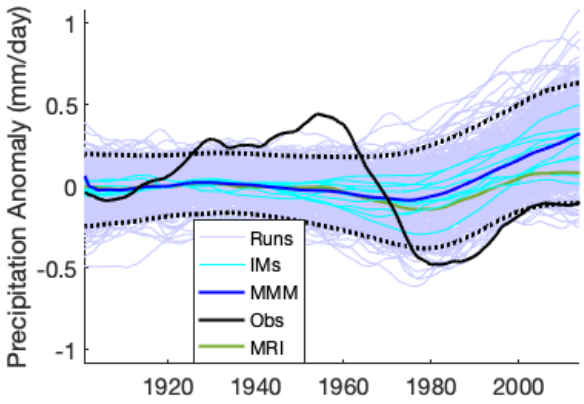

Figure 3.3: Low-frequency Sahel precipitation anomalies from individual ALL simulations ("runs", grey), Institution Means (IMs, cyan), and the MMM (blue) compared to observations

(black). The best-performing IM (MRI) is highlighted in green. The 95% range of runs is outlined in dotted black curves.

Atmospheric and oceanic internal variability ($\vec{a}$ and $\vec{o}$) are not anchored in time, so they do not survive in the MMM, and will generally not align between the observed time series and any given individual simulation; thus, we must compare observations to the ensemble of individual simulations. One way to target low-frequency internal variability would be to compare the observed time series to the 95% range of smoothed individual simulations, outlined in dotted black curves in Figure 3.3. This span clearly does not encompass the observed pluvial between 1925 and 1965 nor the precipitation minimum in the 1980s. A more quantitative approach is presented in Figure 3.4, which compares the power spectra (PS) of individual piC simulations (colored brown to turquoise by model climatological rainfall) to the observed PS (solid black) and the PS of the ALL-residual (observations minus the ALL MMM, dotted-dashed black) to determine whether simulated internal variability can explain observed precipitation variability on its own or in combination with the simulated response to forcing, respectively. In the observations and the residual, variance at periods longer than about 20 years (low-frequency) is roughly 5 times as large as the high-frequency variance. Low-frequency variability in the piC simulations is smaller than, and inconsistent with, either observed or residual variability. Moreover, it is similar in magnitude to simulated high frequency variability. If the observed low-frequency variability is internal, then this suggests that internal variability in simulated Sahel rainfall derives mostly from atmospheric ($\vec{a}$), rather than oceanic ($\vec{o}$), internal variability, or that simulated oceanic internal variability is too white (Eade et al. 2021). Because the shape of the spectrum is wrong, even a bias correction that inflates simulated internal variability would not bring simulations and observations into alignment.

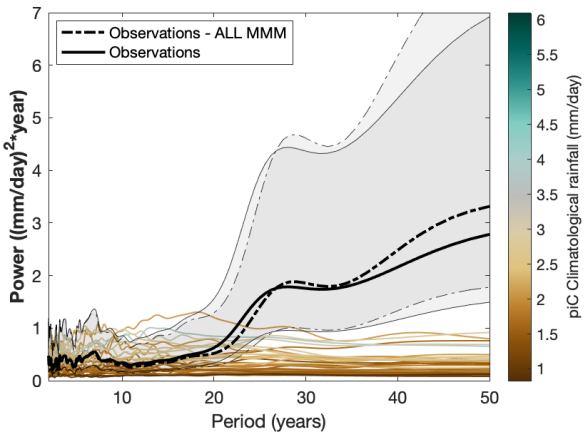

Figure 3.4: PS of observed Sahelian precipitation (solid black curve) and the residual of observations and the ALL MMM (dotted-dashed black curve) and associated 95% confidence intervals (grey shading), compared to the model-average PS of individual piC simulations. Mean piC PS are colored by the average yearly piC precipitation by model, where brown simulations are drier than observed, grey simulations match observed yearly precipitation, and turquoise simulations are wetter than observed.

We must conclude that no linear combination of the simulated forced MMM (which is a poor match to observations at low frequencies) and simulated internal variability (which has insufficient low-frequency variance) in the coupled CMIP6 ensemble can explain observed low-frequency Sahel variability during the 20th century. Thus, model deficiency in the CMIP ensemble cannot be limited to the simulation of climate feedbacks which amplify—but do not otherwise change—the simulated response to forcing: the CMIP6 ensemble displays a fundamental inability to simulate the evolution of the observed Sahelian precipitation response to forcing, the magnitude of observed low-frequency internal variability, or both. To identify the proximate cause of this failure, in the next three sections we examine each causal path component identified in Figure 3.1.

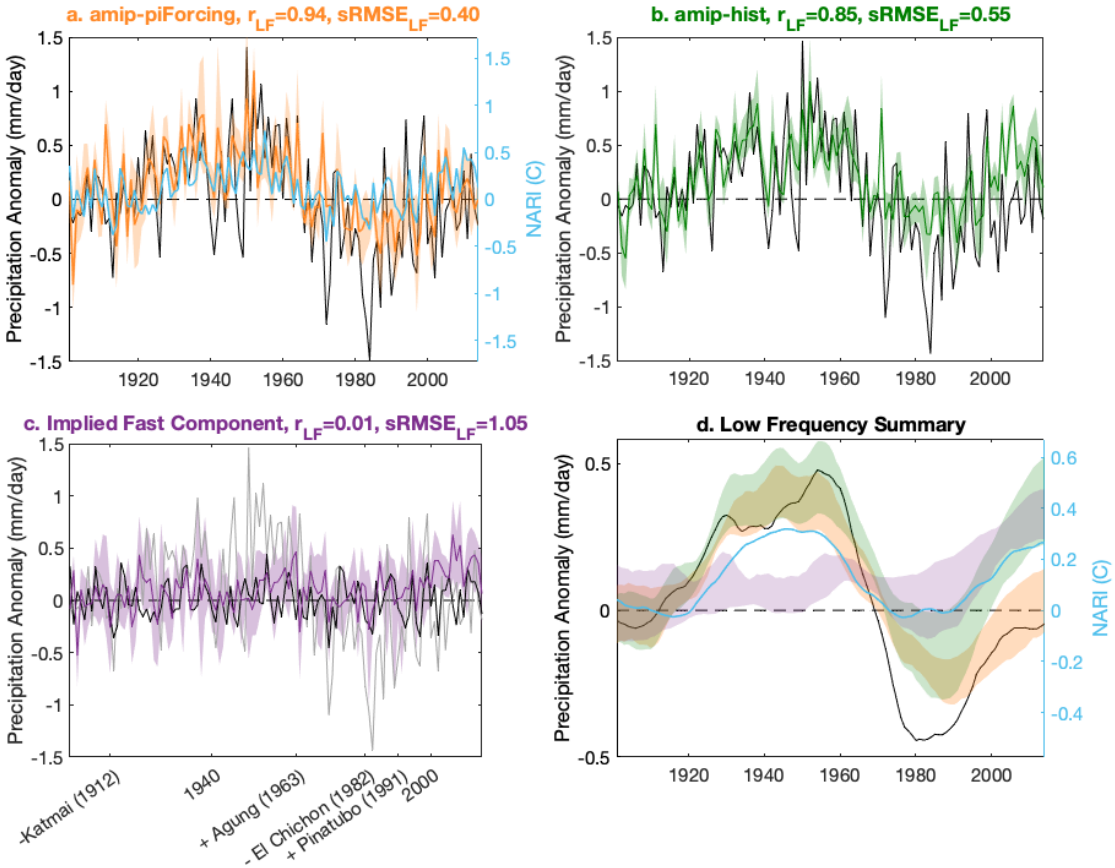

Figure 3.5: Sahelian precipitation anomalies in observations (black) and simulations (colored) from CMIP6 atmospheric models forced with observed historical SST alone (a, amip-piF, orange) and with observed historical SST and all historical external forcing agents (b, amip-hist, dark green). The shaded areas denote the bootstrapping confidence intervals about the simulated MMMs. Panel (a) additionally displays observed NARI anomalies (light blue, right ordinates). The right ordinates for panel (a) are scaled by the inverse of the simulated amip-piF teleconnection strength (see Section 3.4.3) so that, when read on the left ordinates, NARI represents its predicted impact on precipitation. Panel (c) compares observed precipitation at all frequencies (grey) and at high frequencies (black) to the mean implied simulated fast component in AMIP simulations (amip-hist – amip-piF, purple). As in Figure 3.2, panel (c) denotes hemispherically asymmetric volcanic eruptions, where the pictured sign denotes the sign of the expected Sahelian precipitation response to the eruption. Panel titles show the correlation (r_{LF}) and $sRMSE_{LF}$ of simulated MMMs with observations at low frequencies. Panel (d) shows the low-pass filtered versions of panels (a)-(c).

3.4.2. AMIP simulations: the Response to SST, Atmospheric Internal Variability, and the Fast Response to Forcing ($\vec{t}$, $\vec{a}$, and $\vec{f}$)

To isolate the effect of SST on the Sahel ($\vec{t}$), we examine precipitation in the CMIP6 amip-piForcing simulations, which force atmosphere-only models with the observed SST history (containing both internal, $\vec{o}$, and forced, $\vec{s}$, oceanic variability) and constant preindustrial external radiative forcing (no $\vec{f}$). The MMM of simulated Sahel precipitation filters out atmospheric internal variability ($\vec{a}$), leaving the precipitation response to the entire observed SST field (although the reliability of the prescribed SST is still investigated; see Chan and Huybers 2021; Chan et al. 2019). At all frequencies (Figure 3.5a), the amip-piF MMM (orange) is a much better match to observations (black) than the coupled MMMs are. Though the amip-piF MMM still doesn't accurately capture many observed interannual episodes (notably including the precipitation minimum in 1984 and the local minimum in 1972), it captures much of the magnitude of observed low-frequency variability, even including wetting between the 1920s and the early 1960s, which is missing from the coupled MMM. A smoothed low-frequency comparison can be seen in panel (d). When smoothed, the amip-piF MMM achieves very high correlation (r_{LF} = 0.94, CI = [0.90, 0.95]) and low $sRMSE_{LF}$ (0.40, CI = [0.37, 0.52]) with observations. The main source of low-frequency sRMSE (which would ideally be 0) appears to be due to the influence of the dry episodes of 1972 and 1984 on the smoothed observed time series in the 70s and 80s.

Could the difference between the MMM and observations be explained by simulated random atmospheric internal variability? Figure 3.6a compares the PS of observations (black) to that of the amip-piF MMM (dashed orange) and to the mean over the individual-simulation PS (solid orange). The 95% confidence intervals for the PS of the observations and of the amip-piF

MMM are displayed in grey and orange shading. The PS of the amip-piF MMM contains only SST-forced variability ($\vec{t}$), while the mean over the individual-simulation PS contains atmospheric internal variability in addition ($\vec{a}$). Atmospheric white noise gives the mean of PS slightly more power at all frequencies, and thus it falls within the grey shaded area and is not statistically distinguishable from the observed PS (black). Unlike previous generations of AMIP experiments (e.g. Scaife et al. 2009), when combined with atmospheric internal variability, global SST causes Sahel precipitation variability consistent with the full magnitude of observed variability at all frequencies. This substantial low-frequency power can only be achieved when atmospheric internal variability aligns with the high-performing MMM, and thus would place observed variability within the range of individual simulations. This suggests that CMIP6 atmospheric models successfully capture observed teleconnections to SST that are important at low frequencies.

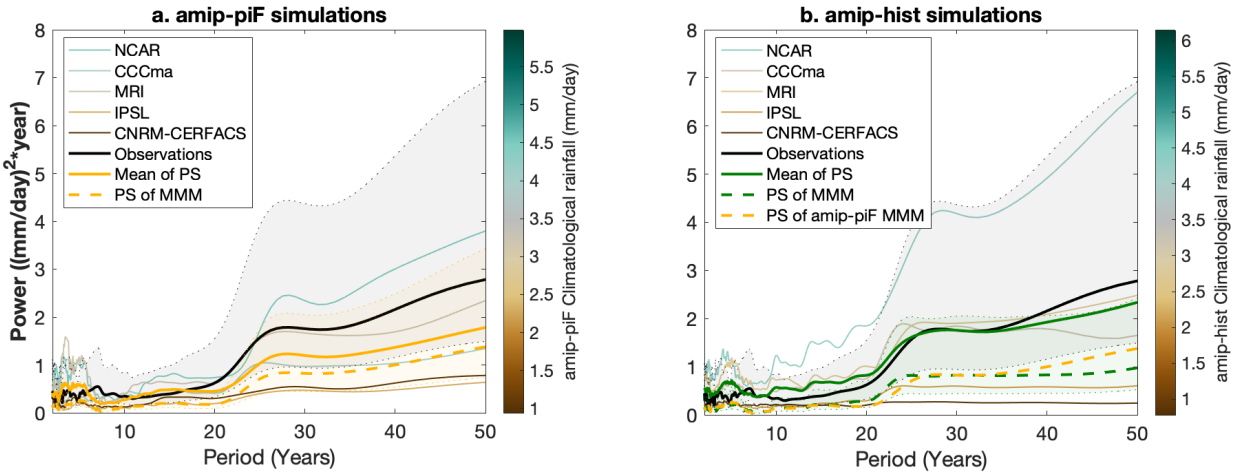

Figure 3.6: PS of observed Sahelian precipitation (bold black) and associated 95% confidence interval (grey shading) compared to the PS of amip-piF simulations (a) and amip-hist simulations (b). As in Figure 3.5, mean PS by model are colored by average yearly precipitation, where brown is drier than observed, grey is observed, and turquoise is wetter than observed. The mean of the model PS is displayed in bold orange for amip-piF (a) and in bold green for amip-hist (b, the MMM PS is below the observed PS in both cases). The bold dashed lines show the PS of the MMMs with associated 95% confidence intervals (colored shaded areas). The amip-piF MMM is repeated in (b) to facilitate comparison.

Could the discrepancy be better explained by the atmospheric response to forcing? The "fast" atmospheric and land-mediated precipitation response to ALL in the CMIP6 AMIP ensemble ($\vec{f}$) can be estimated by subtracting the MMM of amip-piF simulations (Figure 3.5a, orange) from that of amip-hist simulations (Figure 3.5b, green), the latter of which are forced with historical SST and historical external radiative forcing. This estimate (panel c, purple) gives us the *natural direct effect* (as opposed to the 'controlled direct effect'; Pearl 2022) of radiative forcing on Sahel precipitation given observed SST. (If the atmospheric responses to radiative forcing and SST do not add linearly, then the natural direct effect might not be independent of SST, and thus may be different in coupled models.)

Low frequency variability in the AMIP "fast" MMM (panel d, purple) is likely mostly due to anthropogenic emissions. It displays a small wet undulation centered at 1960 and a wetting trend after 1985. It does not perform well relative to observed low-frequency variability on its own (r = 0.01, sRMSE = 1.05), suggesting that it is not the dominant driver of observed variability. Additionally, it only hurts the performance of the amip-hist MMM (green) relative to amip-piF (orange) by keeping simulated precipitation too wet in recent decades—low-frequency correlation reduces from 0.94 (CI=[0.90, 0.95]) to 0.85 (CI=[0.66,0.96]), sRMSE is increased from 0.40 (CI=[0.37, 0.52]) to 0.55 (CI=[0.31,0.77]), and the spectral properties of the precipitation MMM are virtually unchanged (Figure 3.6b, green)—suggesting that the AMIP "fast" MMM may be an inaccurate representation of the observed fast response to radiative forcing.

Short-lived episodic variability in the AMIP "fast" MMM (Figure 3.5b) is likely to be associated with volcanic eruptions, noted on the x ordinates, obscured by noise from both sets of AMIP simulations. Though part of the observed response to these volcanic eruptions will exist in

the “low-frequency” component, most of the low-frequency variability that dominates total observed precipitation variability (grey) is not volcanic and would inhibit comparison with the mostly-flat “fast” MMM. Instead, we compare the “fast” response to observed “high-frequency” precipitation variability (black), which will clarify the timing (though not the full magnitude) of observed episodes of precipitation variability that may be a response to volcanic eruptions. Indeed, the total fast response appears to match the sign and timing of observed episodes near the marked eruptions, notably including the observed precipitation minimum in 1984. This suggests that the actual observed precipitation minimum in 1984—which is not well-captured by the amip-piF MMM (panel a)—is at least partially a fast response to the eruption of El Chichón in 1982. The “low-frequency” component of the fast response to the eruption of El Chichón might have helped lessen the gap between the amip-piF low-frequency MMM (panel d, orange) and observations (black) around 1980, but the benefit of this fast volcanic drying is overwhelmed in the fast (purple) and amip-hist (green) MMMs by the unrealistic anthropogenic fast wetting trend.

The high performance of the AMIP MMMs suggests that the principal deficiency in reproducing observed low-frequency Sahelian precipitation variability in coupled models stems from a failure to simulate the observed combination of forced and internal variability in SST or from corruption of the simulated teleconnections in coupled models (due to e.g., variations in model basic state and patterns of SST variability). The reduced performance of the amip-hist MMM relative to the amip-piF MMM at low frequencies suggests that a secondary deficiency might arise from an overzealous fast wetting response to anthropogenic emissions.

3.4.3. The NARI Teleconnection: AMIP Simulations and Observations ($\vec{t}$)

We next examine Sahel teleconnections to SST in more depth. Observed NARI is presented in Figure 3.5a in light blue on the right ordinates. NARI correlates reasonably with SST-forced Sahelian precipitation at all frequencies in the amip-piF MMM (orange, left ordinates; $\mathrm{r} = 0.60, \mathrm{CI} = [0.42, 0.62]$), but this still leaves 64% of total variance unexplained, suggesting influences from SST variations elsewhere or non-linear or non-stationary effects (Losada et al. 2012). A low-frequency comparison of NARI with simulated precipitation can be seen in panel d ($r_{LF} = 0.72$, $\mathrm{CI} = [0.54, 0.81]$). It is clear that NARI matches the timing and sign of observed (black) and simulated (orange and green) low-frequency Sahel precipitation variability, but no linear teleconnection with NARI can explain the full magnitude of the simulated drought (orange), because NARI in the 1980s is in the same range as in the beginning of the century.

Let's assume that the NARI teleconnection is linear at all frequencies (as claimed by GK19) and unconfounded by the influence of SST changes in other ocean basins (correlating NARI and amip-piF MMM precipitation with globally-varying MMM SST, not shown, suggests that candidates for confounders are limited to ocean basins north of the area covered by NARI or the Interhemispheric Temperature Difference; we will revisit this assumption in Section 3.5). Then we can measure the strength of the NARI teleconnection by the regression coefficient of the amip-piF precipitation MMM, which contains only SST-forced variability, on NARI. This calculation yields a regression slope of $0.87 \pm 0.26 \frac{\mathrm{mm}}{\mathrm{day} * ^{\circ}\mathrm{C}}$. This value is affected by both high- and low-frequency variability, which is appropriate if the teleconnection is, indeed, linear. The left ordinates in Figures 3.5a and 3.5d are scaled relative to the right ordinates by this teleconnection strength so that, when read on the left ordinates, the light blue curve represents

the expected precipitation response to NARI. This view highlights how NARI captures the timing of simulated low-frequency variability, even though it fails to explain the full magnitude of simulated dry anomalies after 1975. In the rest of this paper, we use the NARI teleconnection as the preferred linear representative of the simulated influence of SST on Sahel precipitation in the 20th century.

The teleconnection strength that we calculated from the amip-piF MMM is not directly comparable to observations, because the latter includes the fast precipitation response to forcing, which can in theory confound estimates of the teleconnection. A more direct comparison can be drawn between the apparent teleconnection strength in the amip-hist MMM (0.93 ± 0.41) and in observations (1.04). The consistency lends credence to our previous suggestion that simulated SST teleconnections to Sahel rainfall appear to have the appropriate strength in CMIP6, at least in the AMIP MMM.

3.4.4. Forced and Internal SST Variability in Coupled Simulations ($\vec{s}$ and $\vec{o}$)

We now examine simulation of forced ($\vec{s}$) and internal ($\vec{o}$) SST variability in the coupled ensembles. Figure 3.7 compares observations (black and grey) to the simulated coupled SST response to forcing ($\vec{s}$)—represented by MMM anomalies (colors)—for NARI (right column) and its constituent ocean basins: the North Atlantic (NA, left column) and the Global Tropics (GT, middle column). The horizontal dotted and solid black lines show the bootstrapping 95% confidence intervals of the piC simulations for statistical significance for CMIP5 and CMIP6, respectively, while the colored shaded areas denote uncertainty in the CMIP5 and CMIP6 MMMs. As above, CMIP5 MMM anomalies are presented in dotted curves and CMIP6 in solid curves, color-coded according to their forcing. We observed that the simulated North Atlantic and tropical SST responses to external forcing are quite similar in CMIP5 and CMIP6.

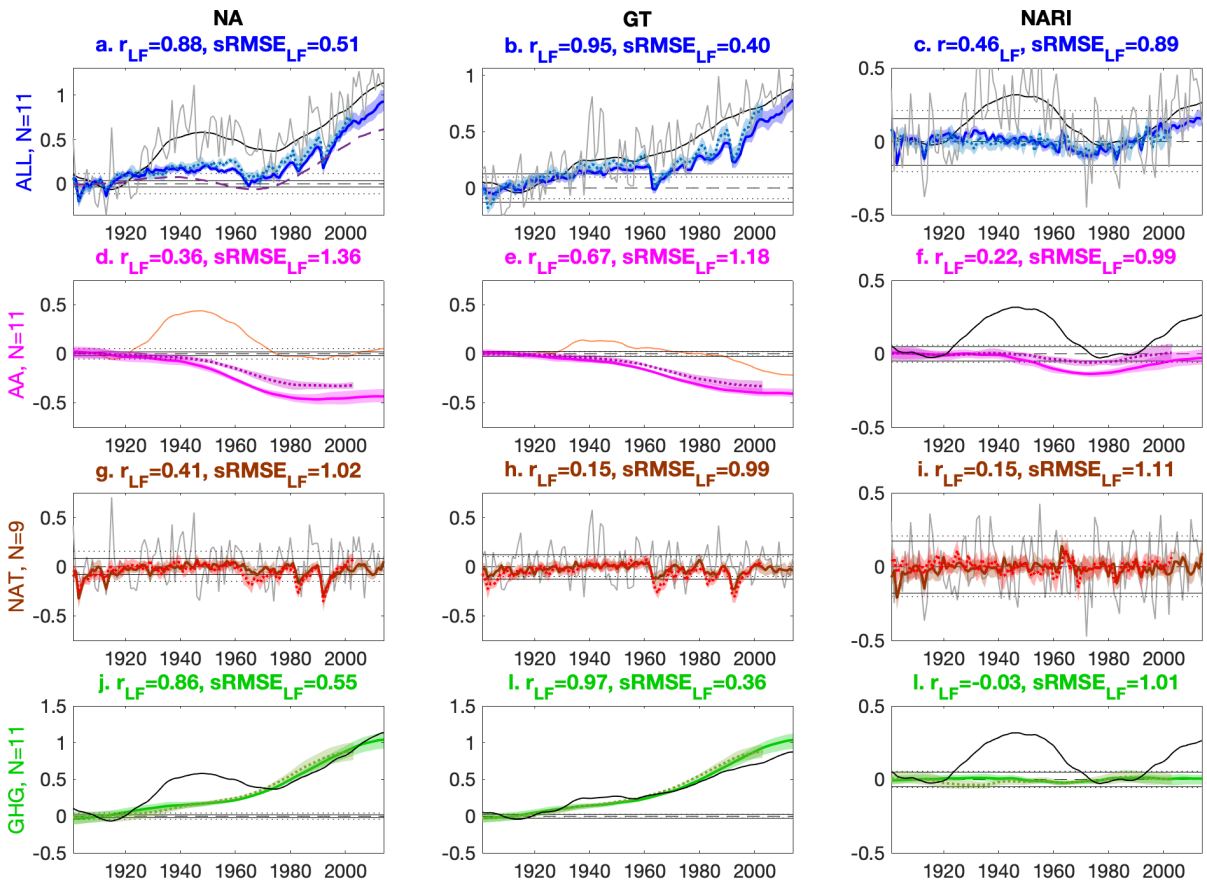

Figure 3.7: Observed high- (grey) and low-frequency (black) SST anomalies (°C relative to 1901-1920) and simulated SST anomalies from CMIP5 (dotted curves) and CMIP6 (solid curves) for the North Atlantic (NA, left column), the Global Tropics (GT, middle column), and the North Atlantic Relative Index (NARI, right column) when forced with ALL (blue, top row), AA (magenta, second row), NAT (brown/red, third row), and GHG (green, bottom row). As in Figure 3.2, the shaded areas mark the bootstrapping confidence intervals, and the horizontal black lines mark the confidence intervals of randomized bootstrapped MMMs from CMIP5 (dotted) and CMIP6 (solid) piC simulations, representing the magnitude of noise in the MMMs. AA (second row) and GHG (last row) simulations are smoothed over 20 years and compared to smoothed observations (black) or a smoothed residual (orange) between observed SST (black) and simulated GHG-forced SST (green, bottom row) in that basin. NAT simulations are compared to high-frequency observed variability and are presented relative to the 1920-1960 mean, between volcanic eruptions. The y labels show the number of institutions that were used for each subset of forcing agents in CMIP6 (N, see Tbl. B.2), and for all panels, the subplot titles display the correlation (r_{LF}) and $sRMSE_{LF}$ between the smoothed MMM and smoothed observations (or GHG residual) for CMIP6. Panel (a) additionally displays the sum of simulated NA forced with AA and GT (burgundy dashed curve).

CMIP5 and CMIP6 historical MMMs are unable to capture observed NARI (panel c, grey), which shows strong multi-decadal variability (black) throughout the century. In the ALL MMM (top row, blue), the temporal evolution of NARI (c) matches the observations at low frequencies with some skill, and significantly better than noise in the MMM (r_{LF} = 0.46, CI = [0.39, 0.51]; $sRMSE_{LF} = 0.89 \pm 0.03$ for CMIP6), but fails to capture the observed multi-decadal NARI warm period between 1925 and 1970. Moreover, its NA (a) and GT (b) components are a poor match to the observed marked multi-decadal variability in NA and the roughly linear warming trend in GT. For NA, the match between observations and the ALL-forced response is better in the later part of the record, but worse in the first half. During the period prior to 1960, according to the MMMs from both CMIP ensembles, GHG warming (j, green) masks AA cooling (d, magenta) to produce a roughly constant temperature in the ALL MMM (a, blue). The simulated cold episode in 1964 is due to the eruption of Agung in 1963 (g, brown and red), and it is only after the mid 1960's that increased GHG warming overtakes stagnating AA cooling to produce pronounced warming in fairly good accord with observations. The simulated response to external radiative forcing cannot explain much of the observed low-frequency variability in NA (a, black), especially the anomalously warm period between 1930 and 1960. In both CMIP5 and CMIP6 ALL simulations, the MMMs of GT (b, blue) are anomalously colder than observations between 1960 and 2000. This behavior is reminiscent of the inaccurate simulation of global SST in HadGEM2-ES, as pointed out by Zhang et al. (2013), who questioned the claims of Booth et al. (2012) that observed AMV was externally forced. This period coincides with a cluster of volcanic eruptions (h) that possibly affect GT too strongly in simulations. It is also the period when simulated AA cooling (e, magenta) is the strongest and not yet compensated by GHG warming (k, green). These correspondences lead us to question whether the match of simulated

and observed NARI in this period happens due to compensating errors across different ocean basins and between the responses to different radiative forcing agents.

Examining all types of experiments (all rows), we see that coupled simulations cannot reproduce realistic low-frequency mean variability in the North Atlantic (left column) or in NARI (right column) no matter what forcing is applied. GHG (l) and NAT (i) both produce little variability in NARI. AA forcing (second row, magenta)—which had appeared to explain the timing of observed low-frequency Sahel precipitation variability in Chapter 2—does produce some low-frequency NARI variability (f), but it does not capture the warm period between 1925 and 1970 or correlate significantly with observations. Moreover, it derives from simulated NA and GT that do not match observations. Specifically, AA do not cause oscillatory variability in NA. Instead, they cause nearly monotonic cooling throughout the century in NA (d) and also in GT (e) with especially steep declines in SST between ~1940 and 1980. Though legislation to curb pollution reduced global AA loading in the northern hemisphere after 1970 (Hirasawa et al. 2020; Klimont et al. 2013; Smith et al. 2011), its effect in both CMIP ensembles is to halt the cooling of NA, not to cause actual warming. This is consistent with estimates of the hemispheric difference in total absorbed solar radiation in AA simulations in CMIP6, which level off—but do not decrease—after 1970 (Menary et al. 2020). We summarize the discrepancy between observations and simulations by noting that, while observed multidecadal variability in NARI derives from multidecadal variability in NA, simulated multidecadal variability in the AA-forced NARI MMM (f) comes from the fact that AA-forced cooling in NA precedes the cooling in GT.

In observations, the warming response to GHG dominates regional averages such as NA and GT, so the AA-forced MMMs are best compared to an observed "GHG-residual" (that is, the observations minus the GHG-forced MMM, presented in orange instead of black), which

represents our best estimate of the sum of observed oceanic internal variability and the observed responses to aerosols (the response to natural forcing is assumed to be small). These estimates (panels d and e, orange) differ greatly from the simulated response to AA, especially in the mid-century warm anomalies, which maximize as simulated AA (magenta) begin to cool NA and GT. Observational estimates may be too high between 1940 and 1945 (Chan and Huybers 2021), but while correcting this bias would address the warm period in GT and reduce the magnitude of the warm period in NA, it could not explain the magnitude or duration of the warm period in NA. Inaccuracies in the simulated response to GHG forcing – which is small in the first half of the century – are unlikely to be the cause of this discrepancy. If the observed Atlantic Multidecadal Variability is externally forced, then volcanic aerosols (g)—which cool NA during the reference period for our anomaly calculations—would have to play a much larger role in forcing observed NA variability than suggested by the CMIP ensembles.

Could internal SST variability ($\vec{o}$) instead explain the difference between the simulated response to forcing and observations in these ocean basins? In Figure 3.8, we present the mean PS of SST indices for piC simulations from each CMIP6 model (colder than observed models are in blue and warmer than observed models are in red). We compare these PS to the PS for observed SST (solid black), the GHG-residual (dotted-dashed black), and/or the ALL-residual (dotted black), omitting PS for time series with dramatic trends. Simulated internal variability in most of the CMIP6 models used in this study does not match residual or observed low-frequency variability in NA (a), GT (b), or NARI (c). In CMIP5, climatological NA and GT are colder and internal variability at all frequencies is larger than in CMIP6, but no model shows an increase in spectral power at low frequencies for any of these SST indices (not shown). There are, however, three CMIP6 models that produce low-frequency internal variability in NA consistent with

observations—in order of decreasing power at 50 years: CNRM-ESM2-1 p1 (pink), IPSL-CM6A-LR p1 (blue), and CNRM-CM6-1 p1 (grey)—suggesting that observed variability may be internal. Certainly, either the simulated SST response to forcing, simulated oceanic internal variability, or both, are not well represented in the CMIP ensembles.

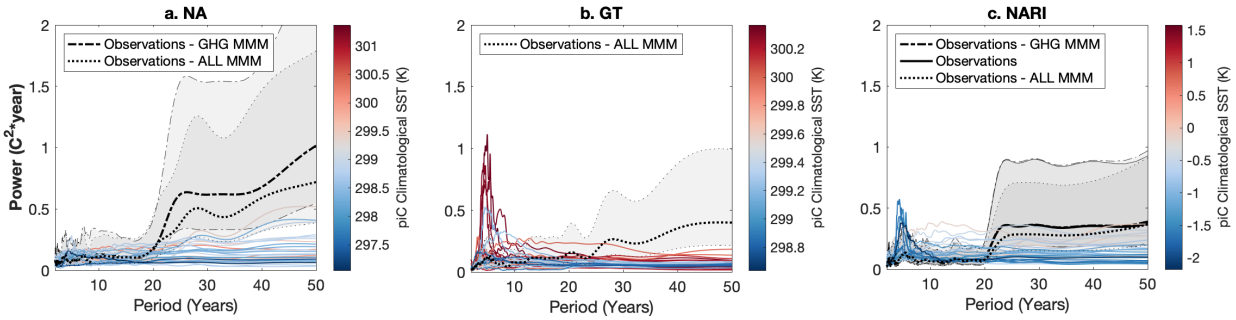

Figure 3.8: PS of observed SST (bold solid black), observed SST – GHG MMM (dotted-dashed black), observed SST – ALL MMM (dotted black) and associated 95% confidence intervals (black shading) in NA (a), GT (b), and NARI (c), compared to the PS of piC simulations. Similar to Figure 3.5, mean PS by model are colored by average SST, where blue is colder than observed, grey is observed, and red is warmer than observed.

Though simulated MMM SST is inconsistent with observations, it may still contribute to the relatively high correlation of the CMIP5 AA MMM with observations. While the correlation of the AA NARI MMM with observations is not significant on its own, it is still positive (r_{LF} = 0.27, CI = [-0.09, 0.55] for CMIP5), and it shares many of its shortcomings with simulated precipitation: neither capture the observed warm/wet period from the 1920s to the 1960s and both reach their minimum in temperature/rainfall about 10 years before the observed minimum. Though we have argued that SST is a more important driver of Sahel rainfall variability than the fast response to radiative forcing in observations and AMIP simulations, the fast response in coupled simulations may mask the differences between the NARI-mediated precipitation response to AA in CMIP5 and observed rainfall changes, giving the appearance of a partially right result, but for the wrong reasons. Additionally, because there are only minor differences in the ALL NARI MMM between CMIP5 and CMIP6 (Figure 3.7c), the difference in mean

simulated Sahel rainfall between CMIP5 and CMIP6 must derive from some combination of changes in the fast response to forcing, the nature of the NARI-Sahel teleconnection, and variability in other SST basins.

3.4.5. The NARI teleconnection in Coupled Simulations

Now that we have examined NARI in the coupled ensemble, we may investigate the ensemble's representation of the NARI teleconnection with Sahel precipitation. Though the estimate of the teleconnection strength from the amip-piF MMM is much more likely to be causal than estimates from coupled simulations that are confounded by external radiative forcing, we cannot automatically assume that it represents teleconnections in the coupled models. First, differences between observed and simulated mean state global SST (Richter and Tokinaga 2020; Wang et al. 2014) may affect the mechanism and strength of the atmospheric teleconnection. Second, while GK19 already showed NARI is dominant in CMIP5, changes in simulated global SST variability may mean that NARI is not the dominant SST driver of precipitation variability in CMIP6. Finally, changes in the atmospheric models themselves between CMIP5 and CMIP6 could cause differences in the teleconnection unrelated to SST.

Do the teleconnections in coupled simulations appear consistent with that in the amip-piF MMM? Calculating the teleconnection strength in coupled simulations is difficult: the relationship between SST and precipitation in historical simulations is confounded by radiative forcing, and SST variability in the (unforced) piC simulations is small and varies from simulation to simulation. Thus, there is no guarantee that NARI is the prominent driver of precipitation variability in the piC simulations. Furthermore, piC simulations must be treated individually, leaving the teleconnection obscured by atmospheric internal variability. The teleconnection strengths calculated from individual piC simulations are, therefore, generally

smaller and less certain than the amip-piF teleconnection strength. Nevertheless, the amip-piF teleconnection strength does fall within the estimated range for CMIP5 ($0.5 \pm 0.6 \frac{\text{mm}}{\text{day}*^\circ\text{C}}$) and CMIP6 ($0.3 \frac{\text{mm}}{\text{day}*^\circ\text{C}}$, $\text{CI} = [-0.1, 0.9]$) piC simulations. (These estimates of the teleconnection strength are unrelated to model performance, not shown.) As a second test, we compare the confounded teleconnection strength from the amip-hist MMM ($0.9 \frac{\text{mm}}{\text{day}*^\circ\text{C}}$, CI=[0.6,1.4]) to that of bootstrapped MMMs over the coupled ALL simulations in CMIP5 ($0.96 \pm 0.56 \frac{\text{mm}}{\text{day}*^\circ\text{C}}$) and CMIP6 ($1.5 \pm 0.3 \frac{\text{mm}}{\text{day}*^\circ\text{C}}$). The confounded teleconnection strength in the CMIP6 amip-hist MMM is consistent with that in CMIP5, but it is smaller than, and inconsistent with, that in CMIP6, falling outside the 95% confidence interval. This may be because NARI variability in the coupled ensemble is smaller relative to the magnitude of external radiative forcing than it is in the amip-hist ensemble. If this is the cause for the apparent inconsistency in confounded teleconnection strengths between the two ensembles, we may still confirm the NARI teleconnection strength indirectly in CMIP6 by showing that the fast response to forcing implied by the NARI teleconnection is consistent with the mean fast response from the amip-hist simulations.

3.4.6. Fast and Slow Responses to Forcing in Coupled Simulations ($\vec{f}$ and F → SST → P)

Under the assumption that the dominant simulated path of SST influence on the Sahel is captured by a linear relationship with NARI, we estimate the slow response to forcing in coupled simulations as the simulated NARI MMM scaled by the teleconnection strength derived from amip-piF MMM ($0.87 \frac{\text{mm}}{\text{day}\,^\circ\text{C}}$, Section 3.4.3), so that a warm (cold) NARI predicts a wet (dry) Sahel. In Figure 3.9, simulated NARI (as in Figure 3.7, right column) is displayed on the left ordinates in light blue (CMIP6, left) and turquoise (CMIP5, right). On the right ordinates are the

total simulated precipitation responses to forcing (as in Figure 3.2), colored by forcing agents. The right ordinates are scaled by the teleconnection strength so that, when read on the right ordinates, simulated NARI represents the estimated slow component of the precipitation response to forcing.

Simulated precipitation matches the estimated slow response in CMIP6 relatively well: The phase of simulated precipitation anomalies in the CMIP6 ALL (a), AA (b), and NAT (c) MMMs match the estimated NARI-driven slow responses throughout the century, and the magnitudes of the anomalies also match throughout the century in the NAT MMM (c), and until 1980 in the AA (b) and ALL (a) MMMs. But after 1980, the ALL (a), AA (b), and GHG (d) MMMs all experience increases in precipitation beyond our estimate of the slow response. In CMIP5, the timing of simulated precipitation anomalies still mostly matches anomalies in the estimated slow response, but the ALL (e) and AA (f) MMMs experience drying after 1950 that is larger than can be explained by our NARI-based estimated slow response to forcing. The discrepancies between the total and NARI-mediated responses, which we'll denote $P_{nonNARI}$, account for most of the difference in simulated forced precipitation between CMIP5 and CMIP6.

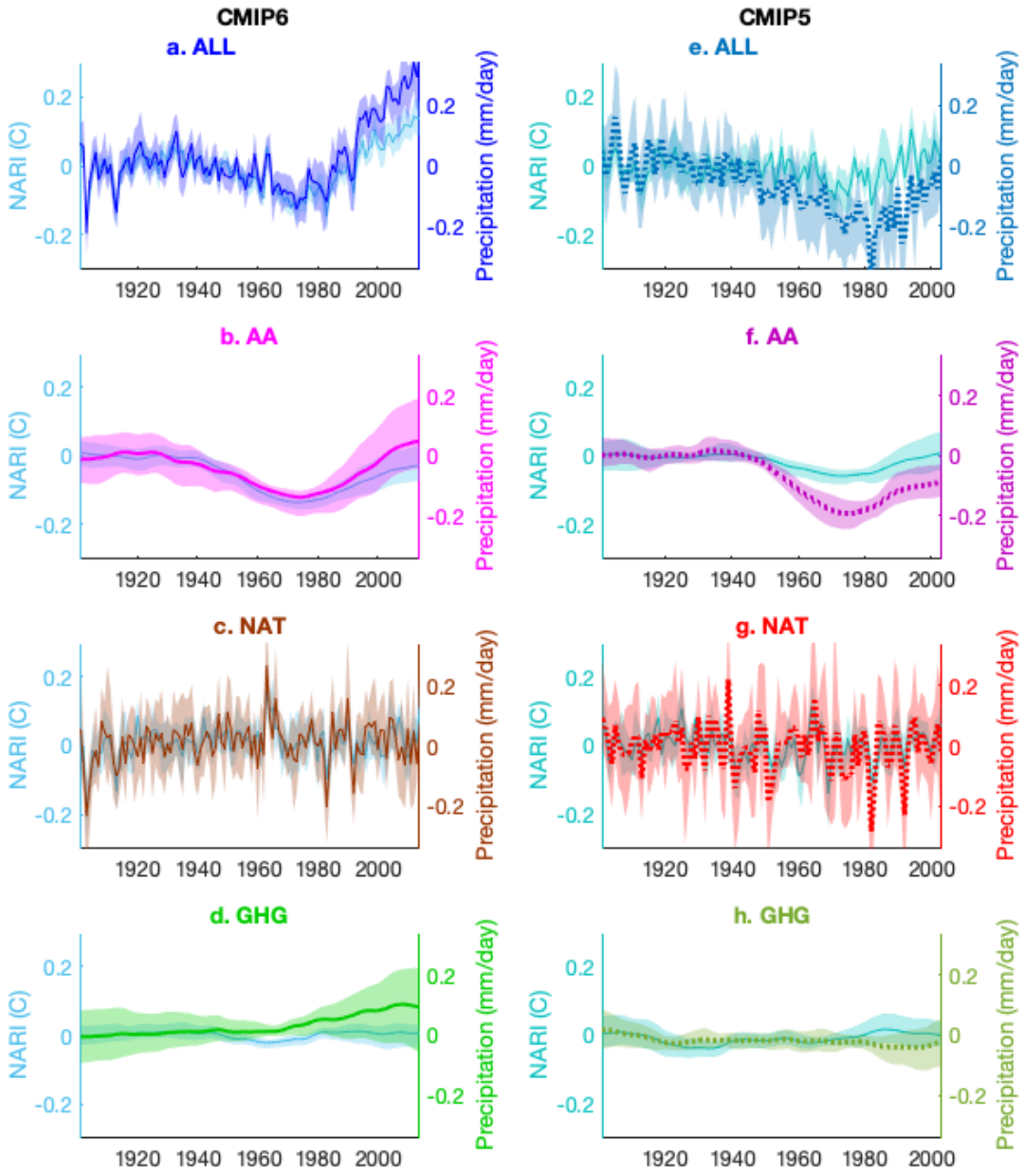

Figure 3.9: Simulated Sahel precipitation (right ordinates, same as Figure 3.2) MMMs (bold solid and dotted curves) at various frequencies and associated 95% confidence intervals (shaded areas) in CMIP5 (right column) and CMIP6 (left column) when forced with ALL (blue, top row), AA (magenta, second row), NAT (brown/red, third row), and GHG (green, bottom row), compared to simulated NARI (left ordinates, thin light blue and turquoise curves, same as Figure 3.7). The right ordinates are scaled such that a 1°C change in NARI corresponds to a 0.87 mm/day change in precipitation, given by the teleconnection strength in the CMIP6 amip-piF MMM (see Section 3.4.3).

Is $P_{nonNARI}$ consistent with the mean simulated fast response in the AMIP simulations? Time series for these differences are displayed in Figure 3.10 in a fashion similar to Figure 3.2, and are compared to the fast response obtained as the difference between amip-hist and amip-piF simulations (purple, as in Figure 3.5c). Mean $P_{nonNARI}$ in the CMIP6 (solid curves) ALL ensemble (Figure 3.10a, blue) is consistent with the sum of wetting from $P_{nonNARI}$ in the AA (b) and GHG (d) MMMs. It doesn't capture the simulated fast wetting in between 1950 and 1970, but does capture the fast wetting after 1980, and matches the AMIP fast response significantly better than noise ($sRMSE_{LF}$ = 0.46, CI = [0.40, 0.77]). These consistencies give us confidence that $P_{nonNARI}$ describes the mean fast response in CMIP6 coupled simulations with all subsets of radiative forcing agents, and that NARI is sufficient to summarize the effect of SST on Sahel rainfall in the MMM from the CMIP6 coupled ensemble.

$P_{nonNARI}$ in CMIP5 (dotted curves), on the other hand, doesn't resemble the mean fast response from CMIP6 AMIP simulations at all. In the ALL MMM (a, turquoise), $P_{nonNARI}$ dries from the 1940s to the end of the record, and is composed of drying centered at 1970 from the AA MMM (b, magenta) and drying after 1980 from the GHG MMM (d, green). Whether $P_{nonNARI}$ in CMIP5 is also consistent with a fast response to forcing in CMIP5 (which we cannot directly estimate) or is a response mediated by SST in ocean basins other than the Atlantic cannot be firmly established by this analysis, but we offer our perspective below.

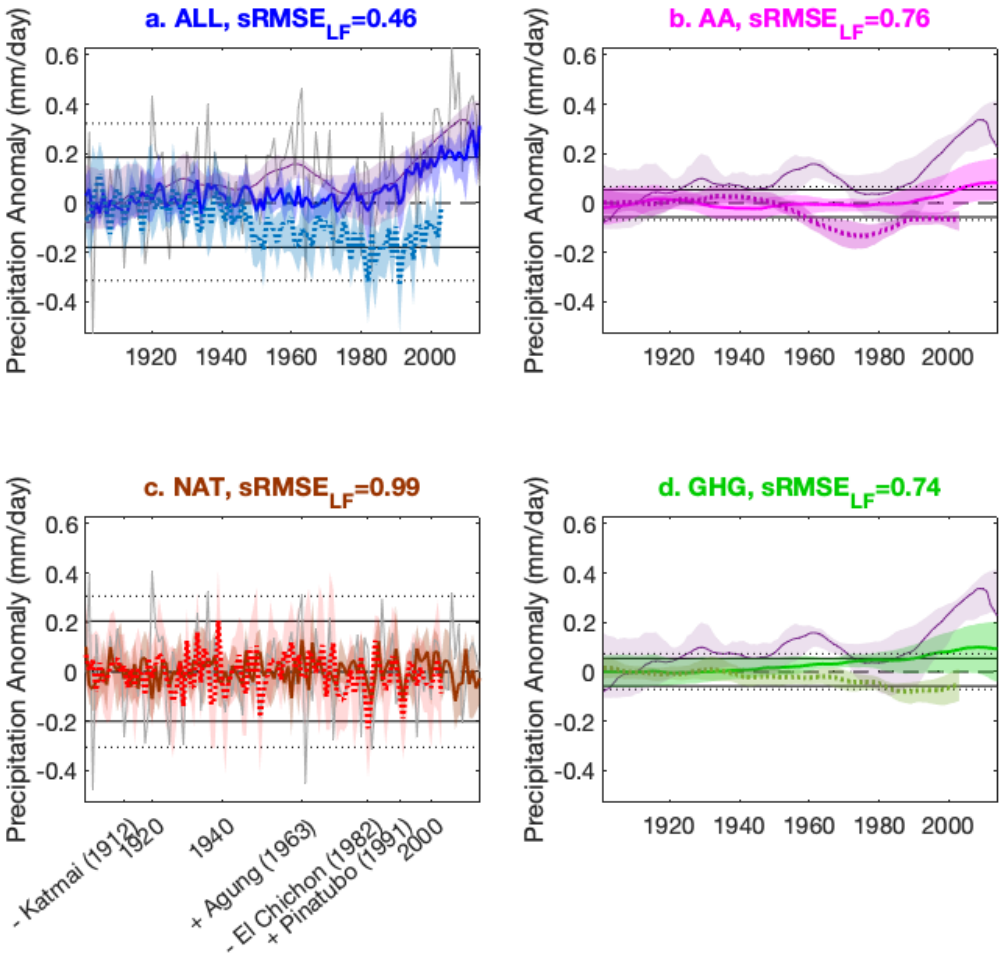

Figure 3.10: Compares the MMM fast Sahelian precipitation response to forcing in AMIP simulations (thin purple curve, as in Figure 3.5c) to $P_{nonNARI}$ MMMs (precipitation – 0.87*NARI; the difference between the colored and light blue curves in Figure 3.9) in coupled CMIP5 (bold dotted curves) and CMIP6 (bold solid curves) simulations forced with ALL (a, blue), AA (b, magenta), NAT (c, brown/red), and GHG (d, green), displayed as in Figure 3.2.

3.5. Discussion: $P_{nonNARI}$ in CMIP5 and CMIP6

The $P_{nonNARI}$ response to GHG is more readily interpreted as a fast response in CMIP6 than in CMIP5, as it is generally accepted that the fast response of the Sahel to GHG is wetting (Biasutti 2013; Gaetani et al. 2017; Giannini 2010; Haarsma et al. 2005; Mutton et al. 2022). Though basic physical theory links increased (reduced) aerosol concentrations to decreasing (increasing) rainfall via fast surface cooling (warming) and increasing (decreasing) optical depth of the atmosphere (Allen and Ingram 2002; Rosenfeld et al. 2008), the observed and simulated

fast responses are difficult to predict because reflective and absorbing aerosols may have different fast effects, and because the concentrations of aerosols vary greatly in space and time and are affected by circulation patterns which differ between models and observations (Hirasawa et al. 2022; Liu et al. 2018). One could propose storylines for the $P_{nonNARI}$ responses to AA in CMIP5 and CMIP6, but these storylines cannot be confirmed without a thorough analysis of the ensemble in the style of Hirasawa et al. (2022), who found evidence of late-century fast drying due to African black carbon emissions in the Community Atmosphere Model 5 – a phenomenon which does not appear to be prominent in the MMM for either CMIP ensemble. But even if the simulated $P_{nonNARI}$ MMM does not successfully capture the fast response to forcing in the observed climate system, the good match between $P_{nonNARI}$ in the coupled CMIP6 ensemble and the estimated mean fast response in amip-hist simulations increases our confidence that $P_{nonNARI}$ in CMIP6 reflects a simulated fast response to forcing, that NARI is a sufficient linear representative of the effect of SST on Sahel rainfall in CMIP6, and that the strength of the teleconnection in the coupled MMM is the same as in the AMIP MMM.

The same cannot be said for CMIP5. We calculated the correlation of $P_{nonNARI}$ with SST in CMIP5 and CMIP6 (not shown) and found it to be uniformly negative in CMIP5 (consistent with the result of GK19 for the 20th century, but not the 21st) aside from the subpolar North Pacific and Atlantic basins (which are most strongly correlated with NARI at low frequencies), giving us confidence that $P_{nonNARI}$ in CMIP5 is not an artifact of using the wrong strength for the NARI teleconnection. The same calculation is uniformly positive in CMIP6, reflecting the prominence of global warming in simulated SST and opposite trends in $P_{nonNARI}$. Uniform warming would not appear as part of NARI, and Sahel drying in response to uniform warming is strong in models that simulate a deeper ascent profile, but weak otherwise (Hill et al. 2017), so it

is possible that newer parameterizations and higher resolution have changed the sensitivity to this forcing in the latest generation of models. Nevertheless, the inconsistency in the sign of this relationship means further investigation is needed to reach a conclusion.

Additionally, we correlated the difference between CMIP6 and CMIP5 in MMM SST with the corresponding difference in MMM Sahel rainfall. The pattern (emerging at low frequencies) is also reminiscent of the simulated global warming pattern, but now includes anomalies of the opposite sign in the subpolar North Atlantic – in the region where models simulate a so-called warming hole in response to GHG and AA (i.e. Baek et al. 2022). Martin et al. (2014) suggest that, unlike the North Atlantic south of 40°N, variability in the subpolar North Atlantic is not well-connected to Sahel precipitation variability, so the connection to $P_{nonNARI}$ might be confounded by the fast effect of the imposed anthropogenic forcings or by its relationship to NA. Insignificant correlations between SST in this region and Sahel precipitation at all frequencies in the amip-piF MMM would support this view, at least when observed SST variability is involved. Nevertheless, it is intriguing that an index summarizing the warming hole pattern (the difference between the North Atlantic box 60-15W by 50-65N and the area 60S-60N) is well correlated at low frequencies not only with the CMIP5-CMIP6 Sahel rainfall difference ($r_{LF} = 0.71$), but also with $P_{nonNARI}$ in each ensemble ($r_{LF} = 0.78$ in CMIP5 and $r_{LF} = 0.86$ in CMIP6). Whether there is a causal connection between the strength of the warming hole in the coupled models and $P_{nonNARI}$ could be tested by targeted simulations, and this is left for future work. In any case, the simulated forced warming hole pattern differs from observed in a manner similar to NARI: it does not capture the observed warm period in the first half of the century, and then cools too little and too early followed immediately by warming in the second

half of the century (not shown); thus, our assessment that a deficient simulation of low-frequency SST variability is the primary reason for CMIP's poor Sahel precipitation variability still stands.

3.6. Summary and Conclusion

In this Chapter, we decompose Sahelian precipitation from atmospheric ("AMIP") simulations into (1) teleconnections from sea surface temperature (SST), (2) atmospheric noise, and a (3) fast, atmospheric- and land-mediated response to forcing. We then calculate a linear regression coefficient between the North Atlantic Relative Index (NARI)—an index of the warming of the North Atlantic relative to the Global Tropics—and the AMIP Multi-Model-Mean (MMM) and employ it to decompose coupled simulations into (1) forced NARI variability, (2) internal NARI variability, and (3) forced precipitation variability not explained by a linear relationship with NARI ($P_{nonNARI}$). Treating NARI as a representative for Sahel teleconnections from global SST allows comparison of the components from AMIP and coupled simulations. We examine these components in order to determine which are responsible for differences in the simulation of Sahel rainfall between the 5th and 6th generations of the Coupled Model Intercomparison Project (CMIP) and which are to blame for the inconsistency of both ensembles with observed Sahel precipitation variability.

When forced with observed SST alone, mean precipitation from CMIP6 atmospheric simulations captures the evolution of observed low-frequency variability quite well (r_{LF} = 0.94, CI = [0.90, 0.95]; $sRMSE_{LF}$ = 0.40, CI = [0.37, 0.52]), and—when combined with atmospheric white noise—these simulations are also able to explain the full spectral power of observed low-frequency variability. This is a welcome improvement from previous generations of climate models, and allows us to use the simulations to make claims about observed variability as well as the shortcomings of the simulations.

Including radiative forcing alongside observed SST in the AMIP simulations mostly acts to increase mean wetting in the second half of the century and notably worsens the MMM sRMSE (0.55, CI = [0.31,0.77]), suggesting that the mean simulated fast response is incorrect while the observed fast component is small and plays a secondary role to SST-forced precipitation variability. Nevertheless, even with the inclusion of the fast response, atmospheric models forced by observed SST capture much of the observed Sahel rainfall variability. This leads us to conclude that observed precipitation variability is mostly a response to global SST, and that the failure of coupled simulations to explain observed low-frequency Sahel precipitation variability must be due mainly to an incorrect response to SST, either because of incorrect teleconnections or deficiencies in reproducing observed forced and internal SST variability.

Using the AMIP ensemble, we summarize simulated Sahel teleconnections from global SST as a linear relationship of $0.87 \pm 0.26 \frac{\text{mm}}{\text{day}*^{\circ}\text{C}}$ with NARI (a relationship that explains about 48% of low-frequency, and 36% of the total, variance in the simulated precipitation response to observed global SST given by the AMIP MMM) and verify that this simulated teleconnection strength is consistent with observations. This estimate is consistent with the coupled CMIP ensembles, meaning that—to the extent that NARI is representative of simulated teleconnections in the AMIP and coupled simulations—the main deficiency of CMIP coupled simulations for explaining observed low-frequency precipitation variability is likely in explaining observed SST variability.

Indeed, simulated mean low-frequency NARI variability is different and much smaller than observed, with the latter mostly coming from low-frequency variability in North Atlantic SST (NA). In both CMIP5 and CMIP6, anthropogenic aerosols (AA) cause a cooling trend and GHG cause a warming trend, but no combination of forcing agents produces a decadal-scale

oscillation in NA. The resulting NARI-mediated slow response to external radiative forcing is to dry the Sahel slightly in the 60s and to wet it immediately afterwards; this does not, in isolation, explain the timing or magnitude of the observed drought or recovery. Only three CMIP6 models (out of 25 CMIP5 and 30 CMIP6 models) are able to generate internal SST variability commensurate to the residual (the difference between total and radiatively forced) low-frequency variability. If we trust these three models, one possible interpretation is that observed SST variability is mostly an expression of internal climate variability that is poorly simulated in the other CMIP models. However, even these three models cannot explain observed *precipitation* variability; and it is not clear physically that we should trust them more than the other models. We must conclude that the CMIP coupled ensembles as a whole are inconsistent with observed low-frequency SST and precipitation variability, and that it is impossible to determine from these ensembles whether the latter is mainly the expression of a response to forcing, oceanic internal variability, or both.

How do we reconcile our results with those using coupled simulations to support claims that the observed Atlantic Multidecadal Variability (AMV) is externally forced (Bellomo et al. 2018; Booth et al. 2012; Hua et al. 2019; Murphy et al. 2017)? The discrepancy can be explained because these studies examine only one or two models (Booth et al. 2012) or subtract a linear trend from simulated NA before comparing to observations (Bellomo et al. 2018; Hua et al. 2019; Murphy et al. 2017), thus aliasing non-linear simulated global warming and inducing artificial low-frequency variability in the simulated AA-induced monotonic decreasing step function (Baek et al. 2022 demonstrated this effect in CESM1). Furthermore, our results for the CMIP5 and CMIP6 ensembles are consistent with the findings of Zhang et al. (2013) for the model used by Booth et al. (2012), showing that simulated multidecadal variability in the North

Atlantic is often associated with unrealistic multidecadal variability in global (tropical) SST. This does not preclude a role for external forcing in observed AMV: observational evidence may still suggest an important role for volcanic aerosols (Birkel et al. 2018) that is poorly captured in simulations, and AA and GHG may still contribute some forced cooling and warming, respectively, to observed AMV in the second half of the century. But a prominent role for internal variability cannot yet be dismissed, as suggested by Yan et al. (2018), who—consistent with our analysis—find that most models do not capture observed AMOC variability. These results cast doubt on claims that AA caused observed AMV based on attribution studies employing CMIP ensembles.

With nearly identical forced NARI variability and similar teleconnection strengths, the difference in Sahel precipitation between CMIP5 and CMIP6 is likely due to $P_{nonNARI}$, which contains the fast response to forcing and slow responses mediated by other SST patterns. Because forced NARI variability in the coupled simulations is small relative to observations throughout the century, $P_{nonNARI}$ has a larger relative effect on the evolution and performance of the MMMs from coupled simulations than on those from AMIP simulations. The CMIP6 MMMs underperform relative to CMIP5 because $P_{nonNARI}$ includes substantial fast wetting responses to increasing GHG and decreasing AA that are comparable in magnitude to the NARI-related component and consistent with the detrimental fast response in that ensemble's AMIP MMM. In contrast, $P_{nonNARI}$ in CMIP5 is drying that may be a fast response to anthropogenic emissions or a slow response to global warming and forced changes in the subpolar North Atlantic.

$P_{nonNARI}$ plays an important role in CMIP5 because it delays the drought from the NARI-mediated slow response to forcing and increases its strength, causing simulated precipitation to correlate well with observed precipitation even though simulated SST—which is the primary

driver of observed precipitation variability—is incorrect. We conclude that CMIP5 MMM Sahel precipitation correlates with historical observations despite not reproducing the physical phenomena that were most important for the Sahel in the 20th century. Thus, while both CMIP5 and CMIP6 suggest that historical AA emissions had a larger effect on observed Sahel precipitation variability than GHG, it is premature to conclude that AA were the dominant influence on observed variability based on attribution studies using CMIP (e.g. Hua et al. 2019; Polson et al. 2014; Undorf et al. 2018).

This work has shown that, while there has been progress in the simulation of the Sahel's response to global SST, much remains uncertain in our understanding and simulation of the pathways of Sahel multi-decadal variability, especially in the characterization of Sahel teleconnections and the simulation of low-frequency variability in North Atlantic SST. Differing mechanisms can lead to similar time evolutions in observations and simulations; to avoid this pitfall, future work should focus on evaluating in more detail the hypothesized pathways of the Sahel response to anthropogenic emissions and oceanic internal variability in individual models as well as the CMIP ensembles in order to further categorize model performance and improve predictions of the future. As a first step toward this goal, Chapter 4 will focus on the role of global SST in driving Sahel rainfall change.

Chapter 4. SST Influence on Sahel Precipitation in CMIP6

Note: This chapter was researched in large part as a project for Elias Barenboim's causal inference class and under the lasting guidance of Adele Ribeiro.

4.1. Introduction

The dominant role of global sea surface temperature (SST) in driving the pacing (though not necessarily the full magnitude) of 20th century Sahel rainfall variability was demonstrated in the early stages of Sahel climate variability research (Folland et al. 1986; Giannini et al. 2003; Knight et al. 2006; Palmer 1986; Zhang and Delworth 2006), and has been further reinforced in more recent studies (Okonkwo et al. 2015; Parhi et al. 2016; Park et al. 2016; Pomposi et al. 2015; Pomposi et al. 2016; Rodríguez-Fonseca et al. 2015 and references therein). A dry Sahel during the West African Monsoon from July to September (JAS) has generally been associated with positive SST anomalies in the global tropics (including individual contributions from the Pacific, Indian, and South Atlantic Oceans), negative SST anomalies in the Mediterranean, and subtropical or extratropical North Atlantic SST that is positive relative to the South Atlantic or to the global tropics. However, the relative importance of these basins and the mechanisms by which they affect the Sahel are still debated due to a number of complications. First, observations relationships appear to have changed over time (Losada et al. 2012; Rodríguez-Fonseca et al. 2011) and are confounded by global anthropogenic emissions. Furthermore, simulated teleconnections differ strongly between models (Joly et al. 2007) and are often a poor match to observations (Joly and Voldoire 2009; Joly et al. 2007).

Giannini et al. (2013, hereafter G13) argued that the influence of global SST on Sahel rainfall in observations and simulations could be summarized with a single index, known as the North Atlantic Relative Index (NARI) and defined as the difference between subtropical North

Atlantic SST (10°-40°N and 75°-15°W) and SST in the Global Tropics (20°S-20°N). Their argument, which is based in column thermodynamics, is that SST in the global tropics sets the minimum surface moist static energy required for convection to occur—called the "ante" for convection (see Section 1.5)—by setting upper tropospheric temperature around the tropics via convective quasi-equilibrium (CQE; Arakawa and Schubert 1974; Emanuel et al. 1994) and the weak temperature gradient constraint (Sobel et al. 2001), while the subtropical North Atlantic supplies critical moisture to the Sahel in quantities proportional to temperature-driven evaporation over the subtropical North Atlantic (see Section 1.6). They gain confidence that this index captures an important aspect of the underlying physics by showing that the index helps explain future projections and inter-model differences in addition to past change, and later work shows that versions of this index also perform well in the Coupled Model Intercomparison Project phase 5 (Giannini and Kaplan 2019). However, the proposed mechanism has not yet been verified, and in Chapter 3, we find that there is more to be understood about the simulated Sahel precipitation response to observed global SST variability than can be explained by NARI. While we find that atmospheric simulations from the Coupled Model Intercomparison Project phase 6 (CMIP6) with prescribed observed *global* SST alone reproduce the full magnitude of observed low-frequency 20th century variability in Sahel rainfall for the first time, we also show that a linear teleconnection with NARI can only account for 50% of the simulated SST-forced precipitation variability over the Sahel.

Because global SST has such a dominant effect on Sahel rainfall, characterizing the true dependence of Sahel pluvials and droughts on global SST is invaluable for many reasons. First, this understanding is necessary for attribution of past change and identifying the global actors responsible for it, which may have economic implications regarding climate reparations (Naylor

and Ford 2023). It is also necessary for process-based climate model validation and improvement efforts (Nowack et al. 2020), which go beyond statistically comparing historical simulations with observations, and may lead to more effective and trustworthy changes to the climate models. Given that state-of-the-art climate simulations struggle to reproduce observed rainfall patterns, such an understanding is also required for prediction of future rainfall change (which also requires accurate projections of global SST under future global warming) that can inform long- and short-term famine mitigation efforts aimed at keeping food production high despite changes in rainfall patterns, such as planting crops that will thrive in the expected conditions (Vignaroli 2017) and building coordinated water-management systems that can accommodate and address future change (Bruins 2019). Finally, such an understanding is essential for determining the efficacy of proposed climate engineering initiatives aimed at controlling the future of Sahel rainfall directly (i.e. by changing the land-surface albedo; Taylor et al. 2002b), and the incidental effects of climate engineering efforts focused on other goals, such reducing global warming via solar dimming (Izrael et al. 2014; Storelvmo et al. 2014).

All of these goals inherently depend on *causal* knowledge, and statistical associations are not enough to address them. Non-causal statistical associations can be useful for short-term prediction if the joint distributions of all relevant variables remain the same. However, statistical relationships cannot be expected to hold in the future as the world warms and aerosol emissions shift, leading to SST and land-sea contrast patterns not typical of the historical record, among other changes (Kamae et al. 2014; Ma and Xie 2013). Neither can they be expected to hold under interventions such as future climate engineering efforts or hypothetical past interventions that might have changed the evolution of Sahel rainfall in the 20th century.

Observed statistical relationships between SST in a given ocean basin and Sahel rainfall may be non-causal for a number of reasons. For instance, the apparently-strong relationship between mean SST over the global tropics and Sahel rainfall could be partially a by-product of the fact that both are affected by greenhouse gases, while in theory Sahel rainfall should respond to tropospheric temperature, which, in the tropical mean, responds instead to a precipitation-weighted mean of SST over the global tropics (Sobel et al. 2002). In this case, we would say that the statistical relationship is *confounded* by a third variable. (The confounding variable could also be SST in another ocean basin). In an observational study, such confounding could lead the researcher to attribute an atmospheric effect of anthropogenic emissions on Sahel rainfall to SST or vice versa, or to attribute SST-driven variability to the wrong ocean basin. Furthermore, autocorrelation in any time-series dataset can bias cross-correlation metrics to be significantly large even for unrelated variables, and to maximize at the wrong time lag for causally-related variables – even sometimes leading to inverted inferred causal relationships (Runge et al. 2014; Zhang et al. 2021). Thus, regression and correlation analyses are not an appropriate way to determine causal relationships, especially in time-series data.

The classic approach for truly measuring a causal effect is through a randomized controlled trial. But, as with many large-scale climate phenomena, performing a true randomized controlled experiment on the relationship between global SST and widespread drought and famine in the Sahel is not only unethical, but also infeasible: we cannot simply intervene with perfect control on the temperature of entire ocean basins, and even if we could, there is only one realization of the state of the monsoon each year, so we cannot control for other varying factors. Researchers often escape the ethical and practical concerns of the real world by performing controlled experiments in simulated environments; but, as previously mentioned, simulated

environments often differ dramatically from the observed environment. The results of such a study should hold in the general circulation model that was used for the experiment, but they do not necessarily reflect the behavior of the true climate system.

Model Intercomparison Projects (MIPs) such as the Coupled Model Intercomparison Project phase 6 (CMIP6; Eyring et al. 2016) attempt to address our lack of trust in individual climate models by soliciting semi-standardized simulations from all institutions maintaining applicable climate models worldwide, allowing researchers to determine the robustness of their conclusions to model parameterization choices. Unfortunately, global climate simulations are computationally expensive, and this plurality in model parameterization comes at the cost of the freedom to perform needed targeted experiments. For instance, in Chapter 3 we used the atmospheric "amip-piForcing" simulations forced with observed global historical SST that are available in CMIP6 to directly measure the combined effect of observed global SST on Sahel precipitation, without confounding between SST and rainfall due to external radiative forcing or feedbacks between rainfall and SST. However, we were not able to reliably measure the individual effects of different ocean basins or SST indices on the Sahel because SSTs in different basins are confounded in the observed record by historical external radiative forcing and interactions between basins. We could theoretically address this with simulations that prescribe SST that varies independently in different ocean basins; but, unfortunately, the focus of CMIP6 targeted SST experiments is limited to low-frequency variability in the North Atlantic and North Pacific (Zhou et al. 2016). Thus, though simulations have some comparative advantages over observations in that they provide multiple realizations of the climate response to certain prescribed conditions and they have complete records of a wide variety of climate variables, they cannot provide robust controlled experiments targeting every variable of interest. In order to

examine robust roles for variables not explicitly targeted in a MIP, we must have a way of measuring true causal effects 'observationally'—or in the absence of a randomized controlled experiment—whether the dataset is composed of observations or simulations.

An early attempt of the scientific community to extract causal information from 'observational' data is Granger causality (Granger 1969), which states that X is a *Granger-cause* of Y if knowledge of X improves predictions of Y that are already based on the past of Y. Granger causality reflects usefulness for predictability, but doesn't truly learn the underlying causal structure because it assumes that all past variables are causal if they improve predictions without considering the possibility of confounding by variables not included in the analysis. Unfortunately, the researcher is prevented from simply adding every potential confounder to the analysis because Granger causality methods do not perform well in high-dimensional problems with more than a couple variables (Runge et al. 2019a). Furthermore, Granger causality cannot accommodate causal relationships that occur faster than the frequency of observations, and while it reduces the impact of autocorrelation effects, it fails to completely address the problem.

To address these limitations, the scientific community has begun to explore the practical efficacy of an alternative framework called *causal inference* (Kretschmer et al. 2021), which provides a theoretically-complete set of rules for discovering causal dependencies and making causal claims from non-interventional data (Pearl 2009; Runge 2018a; Runge et al. In Review; Runge et al. 2019b). *Causal discovery* can help the researcher distinguish between direct causes, indirect causes (that are mediated by another variable), and confounded covariates of a chosen target variable. It can also help the researcher construct a set of sound causal assumptions that can support *causal effect estimation,* which determines whether and how a causal effect can be quantified given the discovered or assumed knowledge about the causal structure underlying the

data generation (see Kretschmer et al. 2021 for practical simple examples in climate science). In its simplest application, causal effect estimation formalizes the choice of covariates to control for in observational regressions (a decision which is non-obvious), but it can also be used to determine more complicated formulas for estimating causal effects. The last decade has seen rapid development of causal discovery algorithms for time series data, which have already been applied at various stages of development to climate problems with some apparent success (e.g. Ebert-Uphoff and Deng 2012; Kretschmer et al. 2017; Kretschmer et al. 2016; Nowack et al. 2020). The methods have continued to improve since then, with higher performance and fewer restrictive assumptions.

In this chapter, we employ state-of-the-art causal discovery methods for time series to learn about the qualitative causal structure relating SST in different ocean basins to each other and to Sahel rainfall in simulations, and attempt to evaluate the efficacy of these methods in practice. We begin by focusing on the long, freely-varying pre-Industrial control simulations because they are not confounded by anthropogenic emissions. These methods should also work on 20th century observations, but further work first must be done to determine how best to represent anthropogenic emissions in the analysis. It is not obvious how best to represent anthropogenic emissions because anthropogenic aerosols are not evenly distributed and greenhouse gas concentrations increase monotonically with very little random variability, yet doing so is necessary because the effects of anthropogenic emissions are so pervasive in the observed record that they will likely cause problems for the causal discovery algorithm if not included.

Our immediate goal is to test the G13's claim that the North Atlantic and Global Tropics (or tropical tropospheric temperature) fully mediate the causal effects of all ocean basins on

Sahel rainfall, which would mean that other ocean basins are not identified as direct causes of Sahel precipitation. Our eventual goal is to use causal effect estimation to quantify the effects of SST in individual ocean basins on Sahel rainfall, giving a complete simple model for Sahel rainfall variability in each climate model. The simple models can be validated by comparing their predicted responses to observed historical global SST variability with precipitation produced by atmospheric simulations from the same climate model driven by this same observed global SST. Because the model is causal, it is much more likely to generalize across changing background climate states than associative statistical models. This pursuit requires full knowledge of the causal structure relating SST in the relevant ocean basins to each other, and because this structure may differ between climate models or between observations and climate models, we also seek to discover the full causal structure relating of modes of SST variability to each other in climate simulations and to characterize the differences between various simulations and between simulations and observations. If the underlying causal structures for different climate models are consistent with each other, then future work will be able to leverage the rules of causal inference to define a single expression for the effect of SST in each ocean basin on Sahel rainfall in every climate simulation. Otherwise, climate simulations may need to be analyzed separately or be selected for process-based consistency with observations (Nowack et al. 2020).

Section 4.2 gives an introduction to causal inference. We return throughout the section to an illustrative example (introduced in Section 4.2.4) to clarify the concepts. Readers already familiar with the concepts of causal inference should feel free to skip most of this section, but are still advised to read Section 4.2.10, which reviews the complications of time series causal discovery, and Sections 4.2.9 and 4.2.11, which introduce the statistical methods and causal

discovery algorithms used in this chapter. The reader can also consult Ebert-Uphoff and Deng (2012) and Kretschmer et al. (2021) for other helpful introductions to the concepts of causal inference. In Section 4.3, we motivate and introduce the climate indices used in this chapter and review the existing literature on their interactions to form a causal hypothesis we expect to hold for the observed climate system in the absence of external radiative forcing. Section 4.4 details the methods used in this Chapter. We apply causal discovery to CMIP6 coupled simulations in Section 4.5. Section 4.5.1 focuses on tuning adjustable algorithm parameters, and Section 0 discusses how we arrive at our results. Section 4.5.3 examines in detail the discovered causal relationships, and Section 4.5.4 evaluates the performance of LPCMCI. Finally, in Section 4.5.5, we synthesize our results and discuss how much trust we can place in different physical conclusions. In Section 4.6, we discuss whether the inconsistent performance of causal discovery algorithms on our dataset is likely to generalize to other datasets, and we conclude in Section 4.7.

4.2. Introduction to Causal Inference

4.2.1. Foundations in Probability Theory

A *joint probability distribution* $P(\boldsymbol{X})$ over a set of random variables $\boldsymbol{X} = \{X_i | i = 1 \dots N\}$ defines the likelihood of observing any possible combination of outcomes, or values for the random variables. We can obtain informative observational *marginal distributions* such as $P\left(\boldsymbol{X} \middle| X_j = x\right) = P(\boldsymbol{X} \cap X_j = x)/P(X_j = x)$ by selecting only the portion of the joint probability distribution that satisfies the condition $X_j = x$, a process we call *conditioning* on X_j. (For continuous variables, we might use a range rather than an individual value as our condition.) A marginal distribution is how scientists might express the probability that there is a drought in the Sahel given that we observe La Niña that year. Two variables X_i and X_j in $\boldsymbol{X}$ are considered *independent* (written $X_i \perp\!\!\!\perp X_j$) when $P(X_i|X_j) = P(X_i)$ (or, equivalently, when $P\left(X_i X_j\right) =$

$P(X_i)P(X_j)$). They are *conditionally independent* given some conditioning set $Z \in \boldsymbol{X}$ (written $X_i \perp\!\!\!\perp X_j|Z$) if $P(X_i|X_jZ) = P(X_i|Z)$. In plain language, this means that knowledge of X_j does not change our estimation of the likelihood of X_i if we already know Z.

In theory, there are multiple ways to test the equivalence of the probability distributions $P(X_i|X_j)$ and $P(X_i)$. One could measure the difference between these distributions using a conditional *distance correlation* (DC) proportional to $\left(\iiint w(X,Y,Z)\big(P(XY|Z) - P(X|Z)P(Y|Z)\big)^2 dXdYdZ\right)^{1/2}$ for some weight function w (Edelmann et al. 2019). It has the formulation of a classic Euclidian distance metric, so it is non-negative and equals 0 if and only if $X \perp\!\!\!\perp Y|Z$. Alternately, one could use *conditional mutual information* (CMI) as the metric, which is defined $I(X;Y|Z) = \iiint P(XYZ)\log\left(\frac{P(XY|Z)}{P(X|Z)P(Y|Z)}\right) dXdYdZ$, and measures the reduction of uncertainty in X due to knowledge of Y given knowledge of Z. It is easy to see that $I(X;Y|Z) = 0$ if $X \perp\!\!\!\perp Y|Z$, and it has been shown that $I(X;Y|Z)$ is non-negative and equals 0 only if $X \perp\!\!\!\perp Y|Z$ (Cover and Thomas 1991).

4.2.2. Structural Causal Models

Causal inference assumes that data is generated from a quasi-deterministic set of structural equations

$$X_i \coloneqq f_i(pa_i, U_i), U_i \sim P_i(u)$$

that define each examined *endogenous* variable $X_i \in \boldsymbol{X}$ as some function f_i of $pa_i \subseteq \boldsymbol{X}\backslash\{X_i\}$ – a subset of the other endogenous variables which we call the *parents* of X_i – and of unmeasured *exogenous* stochastic noise U_i such that each instance u_i of a given noise variable U_i is drawn independent and identically distributed (iid) from a specified probability distribution $P_i(u)$ and

$U_i \perp\!\!\!\perp U_j \; \forall i \neq j$. The ordered set $\langle \boldsymbol{X}, \boldsymbol{U} = \{U_i\}, \boldsymbol{f} = \{f_i\}, \boldsymbol{P} = \{P_i\} \rangle$ is called a *Structural Causal Model* (SCM), and it generates an *observational joint probability distribution* $P(\boldsymbol{X})$.

The symbol $:=$ in our definition of the SCM is used to emphasize that the structural assignments are not reversible equations, but rather represent a causal one-way flow of information from right to left. That means that if we *intervene* on X_j, the data must still satisfy f_i for $i \neq j$, but no longer must satisfy the original functional relationship given by f_j. If we perform a *hard intervention* and set X_j to x regardless of the values of other variables, then the data must satisfy the SCM where $X_j = f_j'( ) = x$ and f_i for $i \neq j$ remain unchanged. This new SCM generates an *interventional joint probability distribution* that we denote $P(\boldsymbol{X}|do(X_j = x))$. In general, $P(\boldsymbol{X}|do(X_j = x)) \neq P(\boldsymbol{X}|X_j = x)$; in plain language: correlation does not imply causation.

In the new SCM, X_j has no parents; any variable X_i that is still dependent on X_j in the intervened SCM is called a *descendant* of X_j in the original SCM. Formally, $X_i \not\perp\!\!\!\perp X_j | do(X_j) \rightarrow X_i \in de_j \; \forall ij$ (note that j may equal i, so a variable is its own descendant). To extend the family relationships metaphor further, X_i is a *child* of X_j if $X_j \in pa_i$, and $X_i \in an_j$ is an *ancestor* of X_j if $X_j \in de_i$. All ancestors of X_i are considered to be *causes* of X_i, but only pa_i are *direct causes*.

The requirement that $U_i \perp\!\!\!\perp U_j \; \forall i \neq j$ means that a variable X_i can never be made independent of its parent $X_j \in pa_i$ by conditioning on other variables: only descendants of X_j contain information about U_j, and they also contain independent variability unrelated to the response of X_i to X_j, so while they can change the nature of the relationship between the variables, they cannot destroy it. It also means that if there exists some Y such that intervening on Y affects two distinct variables $X_i, X_j \in \boldsymbol{X} | i \neq j$ even when conditioning on all their other

parents ($Y \not\perp\!\!\!\perp X_i X_j | pa_i \cup pa_j \backslash Y, do(Y)$), then $Y \in pa_i \cap pa_j \subseteq \boldsymbol{X}$ must be included in the set of endogenous variables and appear as a parent of both X_i and X_j; this requirement is called *causal sufficiency*. Though there exists technically involved work on cyclic SCMs (Bongers et al. 2021; Forré and Mooij 2018, 2020; Rubenstein et al. 2017), for the purposes of this chapter we will focus on acyclic SCMs, in which we can define a *topological ordering* for i such that the parents of every variable X_i are a subset of its predecessors: $pa_i \subseteq \{X_j | j < i\}$. In an acyclic SCM satisfying causal sufficiency, a variable is independent of all of its other predecessors when conditioned on its parents: $X_i \perp\!\!\!\perp X_j | pa_i \; \forall j < i | X_j \notin pa_i$, because the parents contain all information about $U_j, j < i$ that gets passed on to X_i.

4.2.3. Causal Diagrams

For causal effect estimation (Section 4.2.6), it is usually assumed that we do not know the functional relationships $\boldsymbol{f}$, or estimating the causal effects would be trivial. But we must know or assume the set of parents associated with each node X_i, and from this we can determine all independencies in the underlying SCM. This knowledge can be more concisely represented in the form of a directed graph called a *causal diagram*.

A *graph* is a set of variables called *nodes* that may be disconnected or connected with a line called an *edge* or an *adjacency*, and the set of nodes adjacent to a given node are called its *neighbors*. A *directed graph* is a graph where every adjacency is *oriented* to be an arrow → with a tail and a head; the graph's unoriented adjacencies are called its *skeleton*. A directed graph is a *causal diagram* if there is a node for every variable, and we connect $X_j \rightarrow X_i$ if and only if $X_j \in pa_i$. The researcher often does not have access to $\boldsymbol{U}$, but since $U_i \perp\!\!\!\perp U_j \; \forall i \neq j$ and each U_i only affects one variable, the exogenous variables are implied and are not represented directly. For acyclic SCMs, a causal diagram is a *directed acyclic graph* or DAG.

We say that two nodes $X \in \boldsymbol{X}$ and $Y \in \boldsymbol{X}$ in the graph are *d-separated* given a conditioning set $\boldsymbol{Z} \subseteq \boldsymbol{X} \setminus \{X, Y\}$ if all paths connecting X and Y are *blocked*. A path is considered *blocked* if it contains a *chain* $a \to b \to c$ or a *fork* $a \leftarrow b \to c$ with $b \in Z$, or a *collider* $a \to b \leftarrow c$ such that $\boldsymbol{Z}$ contains no descendant of b. By construction, if X and Y are d-separated given $\boldsymbol{Z}$ in the graph, then $X \perp\!\!\!\perp Y | \boldsymbol{Z}$ in the underlying SCM, and this property is called the *Causal Markov Condition*.

If we perform an intervention on X_j, we remove all arrows pointing to X_j in the graph, just as we would replace the structural equation f_j and make X_j independent of its original parents. Any variable X_i is in the set de_j of descendants of X_j if there exists a *directed path* (with all arrows pointing the same direction along the path) of any length (including 0) $X_j \to \cdots \to X_i$. Such a path is not blocked when $Z = \emptyset$, and is not disconnected when we intervene on X_j.

4.2.4. Example and Causal Reasoning

Let's work through a simple example, which we will return to throughout the remaining subsections of this Introduction to Causal Inference. In the summer, people are more likely to go to the beach, eat ice cream, and get attacked by sharks. Let B, I, and S represent monthly July and January aggregates of the number of people at the beach, ice cream sales, and shark attacks, respectively. (We choose July and January monthly aggregates in an attempt to roughly satisfy the requirement that sequential samples are drawn iid, but we would still need to examine our data to make sure there are no trends over time.) Though all of these variables are correlated, to understand which variables cause the others, we must reason about interventions. We know that forcing someone to eat ice cream does not increase his likelihood of encountering a dangerous shark or cause him to go to the beach; that means that $I \notin an_S \cup an_B$, and we draw no arrows coming out of I. On the other hand, while we will assume that spending time at the beach doesn't

change a person's likelihood of eating ice cream, it does directly increase the likelihood of an encounter with a shark; so $B \in pa_S$, and we draw an arrow $B \rightarrow S$. For now, let's assume that shark attacks don't cause people to eat ice cream or inhibit them from going to the beach, so $S \notin an_I \cup an_B$ and we add no further arrows. Our diagram so far is $I, B \rightarrow S$.

We have neglected some factors that affect variables in our diagram. For instance, the likelihood of a shark attack increases when the prey sharks eat move toward the shore (P). The location of shark's prey should not affect ice cream sales, and—assuming that the public is not informed of the movement of shark's prey—it shouldn't affect the number of people at the beach; so, while $P \in pa_S$, $P \notin an_I \cup an_B$. Because P only affects one node in the graph, we can exclude P from the diagram without violating the causal sufficiency assumption, in which case P is considered part of U_S. If we want to include it, we must also consider how it is affected by the other parents of S (if P affects S, then any parent of P would be a parent of S when P is excluded). If we claim that the movement of sharks' prey does not depend on the number of people at the beach, then we simply add the arrow $P \rightarrow S$, and our graph is now $I, B \rightarrow S \leftarrow P$. We have also neglected the temperature (T). We will assume that temperature does not directly affect the movement of sharks' prey or the prevalence of shark attacks, but high temperatures do encourage people to go to the beach and to eat ice cream, so $T \in pa_B \cap pa_I$. Because temperature affects more than one variable in the diagram, we must include it for our graph to qualify as a causal diagram. We assert that temperature is not affected by any of the other variables, and our causal diagram is $I \leftarrow T \rightarrow B \rightarrow S \leftarrow P$. The causal diagram contains one fork ($I \leftarrow T \rightarrow B$), one chain ($T \rightarrow B \rightarrow S$), and one collider ($B \rightarrow S \leftarrow P$). Valid topological orderings for this graph include the 6 orderings where T is first and S is last.

The Causal Markov Condition means that we can read conditional independencies from the diagram using the rules of d-separation. In the observational dataset, if we do not condition on any variables, the only path between ice cream sales and shark attacks ($I \leftarrow T \rightarrow B \rightarrow S$) is unblocked, and we may see a non-causal correlation between ice cream sales and shark attacks because both are more likely when it's warmer out: $I \not\perp\!\!\!\perp S$. Our graph shows us that we can destroy the correlation between ice cream sales and shark attacks by conditioning on temperature, on the number of people at the beach, or both: $I \perp\!\!\!\perp S|T$, $I \perp\!\!\!\perp S|B$, $I \perp\!\!\!\perp S|TB$. However, one should not always condition on *covariates*, or additional variables that correlate with the investigated cause and/or effect variables. The only path between the number of people at the beach and the movement of the sharks' prey ($B \rightarrow S \leftarrow P$) is naturally blocked ($B \perp\!\!\!\perp P$), but if we condition on the number of shark attacks (for example, by examining all months in which there were 2 total shark attacks) then we will un-block this non-causal path, and will likely see that months with more crowding at the beach coincide with months where sharks' prey is further from the shore ($B \not\perp\!\!\!\perp P|S$). One might be tempted to conclude that the presence of people at the beach scares away sharks' prey, but this is inconsistent with the assertions of our causal diagram. Instead, this correlation arises because the number of people at the beach and the proximity of sharks' prey to the shore both increase the number of shark attacks, so a change in one implies an opposite change in the other if we are to maintain exactly 2 shark attacks.

4.2.5. Latent Confounding and Mixed Graphs

In practice, a researcher may not have observations of all the variables $\boldsymbol{X}$ in the underlying SCM needed to satisfy causal sufficiency (Section 4.2.2). In our example (Section 4.2.4), if we do not have access to the temperature record, then T would become exogenous (external to the system). Practically, it could be absorbed by U_B and U_I, causing them to covary,

or it could equivalently be represented as its own variable U_{BI} that also appears in the structural equations for both B and I. We call it *latent confounding* or *unmeasured confounding* when U_i and U_j for $i \neq j$ covary no matter how it is expressed.

We can represent latent confounding between X_i and X_j graphically in an *acyclic directed mixed graph* (ADMG) by connecting $X_i \leftrightarrow X_j$ with a bidirected arrow. We would represent our example system with the mixed graph $I \leftrightarrow B \rightarrow S \leftarrow P$ over the remaining endogenous nodes $\{I, B, S, P\}$. Variables connected with a bidirected edge are called *spouses*, and bidirected arrows are also removed in the case of intervention on either spouse. In graphs with unmeasured confounding, a node may not be independent of all of its predecessors solely by conditioning on its parents (unlike in Section 4.2.2). However, the rules of d-separation still apply (Section 4.2.3), where b behaves as the middle node of a chain in $a \leftrightarrow b \rightarrow c$ and the middle node of a collider in $a \leftrightarrow b \leftarrow c$, and $a \leftrightarrow c$ is a fork that cannot be blocked. In the causally-sufficient graph $I \leftarrow T \rightarrow B \rightarrow S \leftarrow P$, I and B were d-separated given T, but they are not separable in the mixed graph $I \leftrightarrow B \rightarrow S \leftarrow P$. We can still separate I and S in the mixed graph by conditioning on B.

4.2.6. Causal Effect Estimation

Given a correct causal graph for an observed system, *do-calculus* relies on the Causal Markov Condition to leverage the d-separations in the graph (see Sections 4.2.3 and 4.2.5) to express target interventional distributions such as $P\left(Y \middle| do(X = x)\right) | \, X, Y \in \boldsymbol{X}$ in terms of observational distributions such as $P(Y|X = x)$ (see Sections 4.2.1 and 4.2.2). To learn about the rules of do calculus, please consult Pearl (2009).

The simplest application of these rules is called the *Backdoor Criterion*. In order to measure $P(Y|do(X))$ from $P(\boldsymbol{X})$, the researcher must block all noncausal "backdoor"

confounding paths $X \leftarrow \cdots \rightarrow Y$ and backwards paths $X \leftarrow \cdots \leftarrow Y$ between X and Y without unblocking any non-causal collider paths $X \rightarrow \cdots \leftarrow Y$, all of which would otherwise induce a non-causal or reversed-causal observational correlation between these variables. The Backdoor Criterion states that $P(Y|do(X)) = P(Y|XZ)$ if the researcher can find a conditioning set $Z \in \boldsymbol{X}$ that (1) blocks every path between X and Y that has an arrowhead pointing to X and (2) does not include any descendants of X. The second condition prevents us from accidently unblocking collider paths. It also prevents us from trying to measure the effect of temperature on the prevalence of shark attacks by controlling for the number of people at the beach: $P(S|do(T)) \neq P(S|TB)$, which would destroy part of the causal relationship; and from trying to measure the effect of temperature on the number of people at the beach by controlling for the number of shark attacks: $P\big(B\big|do(T)\big) \neq P(B|TS)$, which would induce a selection bias into the calculation. In this case, there are no backdoor paths between T and S or between T and B, so $P(S|do(T)) = P(S|T)$ and $P(B|do(T)) = P(B|T)$. However, if we believe the number of people at the beach may affect ice cream sales and want to measure that effect, we must control for temperature: $P(I|do(B)) = P(I|BT)$.

When the system is not causally sufficient (Section 4.2.5), it is sometimes not possible to convert an interventional query into an observational expression. For instance, if we do not have access to the temperature record ($I \leftrightarrows B \rightarrow S \leftarrow P$), then the rules of do-calculus would tell us that we *cannot* measure $P(I|do(B))$ from the observational distribution $P(I, B, S, P)$. In this case, we say that the causal effect is not *identifiable*.

Causal effect estimation goes beyond the backdoor criterion to generally determine exactly when a causal effect $P(Y|do(X), W)$ can be estimated in terms of the available observational distributions given the assumed causal diagram, and gives an expression for the

effect when possible. A related task is *mechanism estimation*, which attempts to recover the f_i in the SCM based on additional functional assumptions. For continuous, linear variables, the derivative over X of the *expected value* $E[\cdot]$ of the query—$\partial E[Y|do(X), W]/\partial X$—can be considered the causal mechanism.

The output of causal effect estimation will only be correct if the assumed causal diagram is correct, but this is true for standard regression-analysis as well, which implicitly assumes that the regressors are a complete set of parents for the regressand or a backdoor-criterion separating set for some cause and effect variables. Causal effect estimation has the clear advantage because it can fairly interact with a much wider range of assumptions and because it clearly states the assumptions and their implications.

4.2.7. Equivalence Classes and Partial Ancestral Graphs

If a causal diagram represents a hypothesis or a set of beliefs from a domain expert rather than ground-truth knowledge, it can be partially validated against an observational distribution $P(\boldsymbol{X})$ before causal effect estimation (Section 4.2.6) by testing the conditional independencies that are implied by the d-separations in the graph (Sections 4.2.3 and 4.2.5). In our example (Section 4.2.4), when temperature is not measured, our graph $I \leftrightarrow B \rightarrow S \leftarrow P$ implies that $B \perp\!\!\!\perp P$, $I \perp\!\!\!\perp P$, $I \perp\!\!\!\perp P|B$, $I \perp\!\!\!\perp S|B$, and $I \perp\!\!\!\perp S|BP$ in $P(\boldsymbol{X})$. But if we find that the movement of sharks' prey is unconditionally dependent on the number of people at the beach $P \not\perp\!\!\!\perp B$ in $P(\boldsymbol{X})$, then our diagram is not an accurate representation of the observed system. If we had instead hypothesized that crowding at the beach drives sharks' prey away from the beach (Figure 4.1a), then we would find no inconsistency between the observational distribution $P(\boldsymbol{X})$ and our diagram, which implies only that $I \perp\!\!\!\perp P|B$, $I \perp\!\!\!\perp S|BP$, and $I \perp\!\!\!\perp S|B$ (B blocks both paths connecting I to S: $I \leftrightarrow B \rightarrow S$ and $I \leftrightarrow B \rightarrow P \rightarrow S$). Though I and P are not d-separated in this graph, the graph is still

consistent with the observed distribution even if $I \perp\!\!\!\perp P$ in $P(\boldsymbol{X})$: while the Causal Markov Condition states that every independence implied by the graph must hold in $P(\boldsymbol{X})$, there may be additional independences in $P(\boldsymbol{X})$ that are not implied by the graph (Section 4.2.3). However, if we additionally found that ice cream sales are not independent of shark attacks when we condition on the number of people at the beach $I \not\perp\!\!\!\perp S|B$, then Figure 4.1a would again fail to explain the observed distribution. An alternate hypothesis that the public is informed about the whereabouts of the sharks' prey and avoid going to the beach when sharks are likely to be present, as in Figure 4.1b, would imply only that $I \perp\!\!\!\perp P$ and $I \perp\!\!\!\perp S|BP$, and would thus be consistent with the observational distribution.

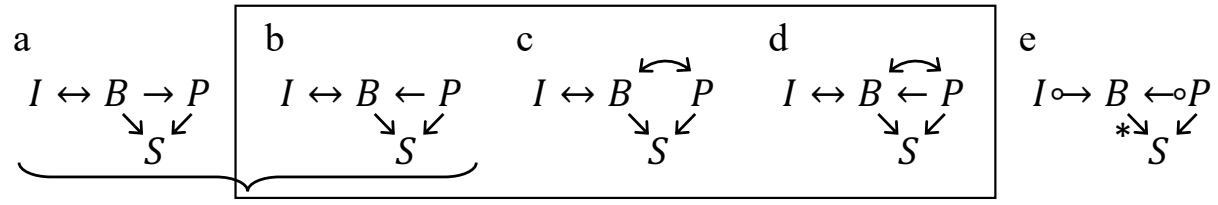

Figure 4.1: Alternate hypotheses for a causal relationship between the number of people at the beach and the location of sharks' prey (a and b, noted with a bracket). Some directed acyclic mixed graphs (b-d, outlined with a box) that belong to the same equivalence class, presented as a partial ancestral graph (e).

We cannot completely evaluate graphs this way because there are other graphs that would lead to the same set of d-separations. In this example, we would find the same set of d-separations if we had assumed there is an unmeasured confounder between the number of people at the beach and the movement of sharks' prey—instead of (panel c) or in addition to (panel d) an effect of the movement of sharks' prey on the number of people at the beach.

When multiple diagrams yield the same d-separations, they belong to the same *equivalence class.* A non-Markovian equivalence class for ADMGs (Section 4.2.5) can be represented with a *partial ancestral graph* (PAG). In any *ancestral graph*, an arrow $A \rightarrow B$

implies ancestorship rather than parenthood, so two nodes that are adjacent in the ancestral graph may not be directly adjacent in an ADMG consistent with it. For instance, the ADMG $T \rightarrow B \leftrightarrows S$ could be re-expressed as an ancestral graph $T \overset{\frown}{\rightarrow B \rightarrow} S$, which states that T is an ancestor of S that is still related to S even when conditioning on B (without conditions, T is connected to S by the path $T \rightarrow B \rightarrow S$, but if we condition on B, then we unblock the path $T \rightarrow B \leftrightarrow S$). PAGs additionally admit edges with circles at their ends, representing multiple types of relationships between the variables in its component ADMGs. Table 4.1 details the possible PAG edges and lists the ancestral and/or confounded relationships that can exist between those same variables in the members of the equivalence class, with the one complication that edges that form a tripe with neighboring circles, such as $A \leftarrow\circ B \circ\rightarrow C$, cannot be fully-oriented independently, because forming a collider would remove the DAG from the equivalence class (Zhang 2008).

Table 4.1: Edges that may connect two nodes in an Acyclic Directed Mixed Graph (ADMG) that belongs to an equivalence class represented by a Partial Ancestral Graph (PAG) are marked with an X based on the type of edge that connects those nodes in the PAG.

PAG edges \ ADMG edges	$\leftrightarrows$	$\leftarrow$	$\rightleftarrows$	$\rightarrow$	$\leftrightarrow$
$\circ\!-\!\circ$	X	X	X	X	X
$\circ\!\rightarrow$			X	X	X
$\rightarrow$			X	X	
$\leftrightarrow$					X
$\overset{*}{\rightarrow}$				X	

The equivalence class for Figure 4.1 panels (b)-(d) is presented in panel (e). The arrow $B \leftarrow\circ P$ in panel (e) represents a causal effect (panel b), confounding (panel c), or both (panel d). Though we have not pictured any causal diagrams with a causal relationship $I \rightarrow B$, such a

relationship would not change the colliders or d-separations implied by the diagram, so this is allowed in the equivalence class. However, $B \rightarrow S$ cannot be confounded without changing the d-separations of the graph; this edge is called *visible* and we denote it with a star or a 'v' (these marks are often omitted).

4.2.8. Causal Discovery

Sometimes the domain expert cannot formulate a causal diagram or equivalence class that is consistent with the data from first principles alone. If we can validate or reject a hypothesized graph by checking conditional independencies in the data (Section 4.2.7), we can also "discover" the equivalence class of the causal diagram underlying a dataset by listing every possible equivalence class over a set of nodes and validating each one. Such a naïve approach quickly becomes computationally intractable. But under the *faithfulness* assumption—which is the reverse of the Causal Markov Condition and requires that independence in the observational distribution $P(\boldsymbol{X})$ implies d-separation in the causal graph—it is possible to choose a subset of independence conditions to check so that each one limits the set of possible equivalence classes.[8] Theoretically, faithfulness should not always hold, since there may be competing unblocked causal paths between two variables that counteract each other. It has been argued that faithfulness violations are incredibly unlikely when causal effects are randomly distributed (Spirtes et al. 2000), but real systems may not be randomly distributed, and constraints arising from physical conservation laws or from the persistence of a stationary dynamic system may make faithfulness violations more likely in the real world (Andersen 2013).

[8] Some algorithms rely on the somewhat-weaker *adjacency faithfulness* assumption, which only requires that adjacency in the graph (rather than d-connectedness, which is the opposite of d-separation) implies conditional dependence in the observational distribution (Ramsey et al. 2012).

Algorithms that use conditional independencies to learn the equivalence class associated with a dataset are called *constraint-based causal discovery* algorithms. The classic algorithm was formulated with a reasonable runtime by Peter Spirtes and Clark Glymour (1991), and termed the "PC" algorithm after their first names. Here, we use the language of PAGs (Section 4.2.7) to present a slight modification of the PC algorithm that allows for latent confounding (Section 4.2.5). Begin by assuming a partial ancestral graph that is maximally connected with edges with circles at both sides. First, learn the skeleton of the graph (adjacencies without orientations) by searching for pairs of variables $\{X, Y\}$ that are not directly related, for which we can find a set $Z_{XY} \subseteq \boldsymbol{X} \backslash \{X, Y\}$ such that $X \perp\!\!\!\perp Y | Z_{XY}$. Search through potential separating sets (where each element of the set is a potential ancestor of Y) of increasing size, starting with 0. If such a set is found, remove the edge between X and Y. Next, begin to orient edges by searching for colliders, which induce relationships between pairs of variables that have already been separated by some conditioning set Z_{XY}. When a pair of non-adjacent variables $\{X, Y\}$ share a common neighbor: $X \circ\!\!-\!\!\circ N \circ\!\!-\!\!\circ Y$, if $N \notin Z_{XY}$ (N is not in the set that separated X and Y), partially orient the edges to show that N is a collider: $X \circ\!\!\rightarrow N \leftarrow\!\!\circ Y$. Finally, replace remaining circles with heads or tails if the opposite choice would create a new collider or a cycle.

In practice, a researcher does not have access to the observational probability distribution function $P(\boldsymbol{X})$, and must rely on some statistical estimator of conditional dependence (see Section 4.2.9). Sampling and measurement errors mean that no test statistic will ever be exactly 0 even when the underlying processes are truly independent, regardless of our choice of metric. So, we must define some (significance) threshold α under which statistical conditional dependence is considered equivalent to 0, and face a tradeoff between identifying false independencies and missing true independencies. Thus, in practice, causal discovery relies not only on the

faithfulness assumption, but also on some version of the even stronger assumption of *strong faithfulness*: that all statistical dependence test results below some threshold correspond to d-separations in the graph. This assumption is often violated in real data (Uhler et al. 2013), and can be sensitive to the choice of metric for conditional independence and the way it is estimated relative to the properties of the underlying SCM.

4.2.9. Statistical Conditional Independence Testing

In Section 4.2.1 we discussed some metrics for measuring dependence $X \not\perp\!\!\!\perp Y|Z$ by quantifying the difference between the distributions $P(XY|Z) = P(XYZ)/P(Z)$ and $P(X|Z)P(Y|Z) = P(XZ)P(YZ)/P(Z)^2$ including CMI and distance correlation. In practice, we cannot use these metrics without estimating $P(\boldsymbol{X})$. Given samples drawn iid from $P(\boldsymbol{X})$, one can estimate $P(\boldsymbol{X})$ by dividing the domain of the observations into subranges and measuring the prevalence of observations within each subrange.

"Fixed global bandwidth" estimators divide the total range of the observed variables into even subranges, as is typical when making a histogram, for instance. The researcher must choose the number of bins, and faces a tradeoff: if the bin size is too large, the histogram doesn't have enough resolution to shed light on the true probability distribution, but if the bin size is too small, the accuracy of the estimates decreases. For instance, a small bin size can lead to unfortunate coincidences such as having no observations in some bin near the center of a normal distribution, preventing the estimated distribution from appearing smooth, let alone normal.

"Data-adaptive" approaches instead adjust the limits of the subranges to maintain the same amount of data in each range, with smaller ranges where data are dense, and larger ranges where data are sparse, avoiding this strange behavior. But the researcher still adjusts the number and sizes of the bins by choosing the amount of data in each bin, and still faces a tradeoff

between accuracy and precision. On the one hand, with smaller bins and less data in each bin, the variance of the estimate for each bin increases and so the accuracy decreases. On the other hand, in the extreme case of having one large bin, there is no variance in the probability density estimate (it must equal 1), but there is also no resolution in the estimated probability distribution. This lack of resolution makes two distributions more likely to appear equivalent. Thus, when estimating independence, a large bin size causes a bias toward zero whereas a small bin size increases variance due to chance.

CMI can be estimated using a data-adaptive k-nearest-neighbor estimate (CMIknn), and Runge (2018b) implements a computation significance test for of CMI by constructing a new distribution $P(\tilde{X}YZ)$ where $\tilde{X} \perp\!\!\!\perp Y|Z$, but the relationship between X and Z is preserved. He accomplishes this by shuffling values of X within a "shuffle neighborhood" of observations with the most similar Z values. Thus, CMIknn has two additional important tunable parameters: k nearest neighbors (knn) and shuffle neighbors (SN).[9] If SN is too small, then the relationship between X and Z will be destroyed and the null distribution will be overly similar to the alternate distribution, so increasing SN increases sensitivity (true dependence results) while decreasing SN increases fidelity (reduces false dependence results). Knn is expressed as a fraction of the sample size (N, the length of the time series), and Runge (2018b) recommends values between 0.1 and 0.2. SN is a small integer, and values between 5 and 10 are recommended. These recommendations are based on computational analysis of the conditional independence tests with sample sizes from 50 to 2000 and dimensionalities (number of variables) of 3 and 10. Within the context of causal discovery, these recommendations are still considered preliminary, and the

authors request further analysis to determine the best parameter choices. Unfortunately, this CMI estimator results in low detection power in the context of causal discovery with multiple variables, meaning it often finds independence where there is none and leads to incorrectly removing edges, especially when conditioning sets are large (see Figure 5e of Runge et al. 2019a).

Another approach to estimating the probability distribution function from independent samples of $P(\boldsymbol{X})$ is calculating the empirical characteristic function, which is an unbiased and consistent estimator of the probability distribution that does not rely on dividing the data into bins. Székely et al. (2007) use empirical characteristic functions to estimate distance correlation, but only when $Z = \emptyset$, and so some parametric assumption must be made to regress out Z from X and Y, and this approach can only be used when the functional relationships satisfy the parametric assumptions.

Another approach is to avoid estimating $P(\boldsymbol{X})$ altogether. The *partial correlation coefficient* $\rho_{YX \cdot Z} = (\rho_{YX} - \rho_{YZ}\rho_{XZ})/[(1 - \rho_{YZ}^2)(1 - \rho_{XZ}^2)]$^.5 is not a distance metric for the probability distributions—it only examines the mean of the distributions, and it can be negative—but it suffices as a theoretical proxy for conditional independence for linear dependencies with additive Gaussian noise. When $Z = \emptyset$, it is easy to see that the true unconditional correlation of X and Y—$\rho_{XY} = (E[XY] - E[X]E[Y])/\sqrt{\sigma_X \sigma_Y} \propto \iint XY[P(XY) - P(X)P(Y)]dXdY$, where σ_X is the standard deviation of X—will always equal 0 when X and Y are independent even when the underlying SCM is not linear. However, the conditional partial correlation coefficient $\rho_{XY \cdot Z}$ implicitly uses linear regression to remove the effect of Z on X and Y, and doesn't generally equal 0 when $X \perp\!\!\!\perp Y|Z$ if the relationship between Z and the other variables is non-linear. Furthermore, the reverse implication that $\rho_{YX \cdot Z} = 0 \rightarrow X \perp\!\!\!\perp Y|Z$—which

is necessary to satisfy the faithfulness assumption when using partial correlation as the conditional independence indicator—also only holds when the functional assignments are linear and the noise is Gaussian. Thus, we can only use this indicator for conditional independence for causal discovery when we believe our SCM to be linear with additive Gaussian noise. We do not have to know $P(\boldsymbol{X})$ to calculate the *sample partial correlation coefficient* $r_{YX \cdot Z} \propto (N \sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i)$, and partial correlation results in higher detection power for causal discovery in the linear case (Runge et al. 2019a).

4.2.10. Time Series

When the variables $X_i = X_i(t)$ are time series, their values at different times are not independent, violating our definition of an SCM (see Section 4.2.2). We can address this by dedicating a node $X_i(t)$ for every variable X_i at every time t. Since the observations of different variables at the same time are likely to be confounded by the past of all variables in the system, we allow the parents of $X_i(t)$ to include past observations of all variables including itself $(X_j(t - \tau), \tau > 0)$ in addition to simultaneous observations of other variables $(X_j(t), j \neq i)$. In practice, $\tau \in 0 \dots \tau_{max}$ is also limited to be less than or equal to some finite maximum time lag $\tau_{max} \in \mathbb{R}$.

Simultaneous feedbacks still cannot be represented under the assumption of acyclicity; but, given a high enough sampling rate, feedbacks loops can be "unrolled" over time such that at least one of the causal relationships in the feedback mechanism has a lag. In our example, for instance, we already recognize that increasing the number of people at the beach increases the expected number of shark attacks when sharks are present, but it may also be that people are less likely to go to the beach the days after a shark attack occurs. We would not be able to represent this feedback when looking at monthly aggregates because it would create a 'simultaneous' cycle

in the graph. But if we examine a time series with a timestep of one day, we can unroll the feedback over time, and we might obtain a graph such as Figure 4.2a.

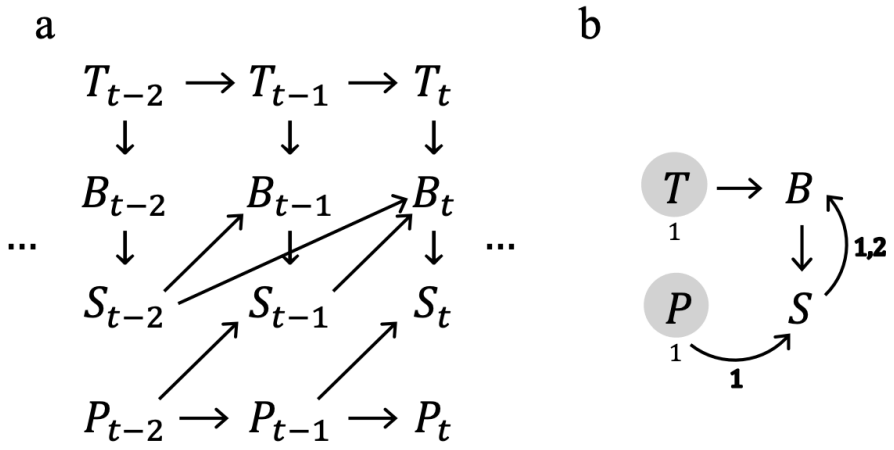

Figure 4.2: Time series causal graph (a) and corresponding summary graph (b).

Without further assumptions, it would be impossible to evaluate conditional independence for such a causal diagram, because we would only have one observation for each node in the graph. Under the assumption that the causal structure is *stationary*, a single functional relationship f_i is needed to define the behavior of $X_i(t)$ for all t. This means the causal graph will have an infinitely repeating structure, and can be represented more compactly as a summary graph in which each variable appears only once, auto-dependence is represented by a circle about the node's name with causal lags listed below, and simultaneous dependencies are marked with straight arrows while lagged dependencies are marked with curved arrows labelled with the length of the lag in outlined numbers (Figure 4.2b)[10]. More importantly, it

[10] This is method doesn't clarify the direction of the lag for confounded adjacencies. A convention for this case should be determined.

allows us to use observations at different times as a sample for testing conditional independence between the variables.

Unfortunately, these samples $X_i(t)$ are still not independent of one another, and statistical conditional independence tests (see Section 4.2.9) also rely on the assumption that samples are independent and identically distributed. Runge et al. (2014) showed that, in the linear case, the partial correlation $\rho_{YX \cdot pa_Y \backslash X}$ (Section 4.2.9) is affected by autocorrelation in both X and Y in addition to the causal effect between them even though the other parents of Y are included in the conditioning set. This causes theoretical problems for effect estimation, suggesting that $\rho_{YX \cdot Z}$ can give a biased estimate of the causal effect of X on Y even when the conditioning set Z satisfies the backdoor criterion. It does not pose a theoretical problem for causal discovery, but it does pose a practical one: while the true partial correlation coefficient should still equal zero when variables are independent, the dependence on the autocorrelation of X and Y makes it more likely that the sample partial correlation coefficient $r_{YX \cdot Z}$ (Section 4.2.9) will cross the defined threshold for dependence (Section 4.2.8) even for d-separated variables when samples are finite, resulting in detection of false edges and false orientations. Runge et al. (2014) showed that, for effect estimation, the bias in $r_{YX \cdot Z}$ can be addressed by conditioning on the parents of both X and Y.

Generally, it has been argued that iid violations reduce the number of degrees of freedom of the data and the width of the test statistic distributions for any conditional independence test, but theoretical results on the implications of iid violations on true conditional independence metrics (including CMI and distance correlation) and in turn on the soundness of time-series causal discovery algorithms when data are finite are still missing. Nevertheless, Runge (2018a) showed computationally that eliminating autocorrelation by conditioning on the parents of *both*

the cause and effect variables improves the performance of time-series causal discovery algorithms even when using CMI to measure independence. This test $X \perp\!\!\!\perp Y | pa_X \cup pa_Y \backslash X$ is called *momentary conditional independence* (MCI).

4.2.11. Time-Series Causal Discovery Algorithms

Including past values of all variables as separate nodes (Section 4.2.10) dramatically increases the dimensionality of the dataset, reducing the detection power of conditional independence test and thus the performance of causal discovery algorithms in the time series case. The PCMCI algorithm (Runge et al. 2019a) generalizes the PC algorithm to the time-series case (see Section 4.2.8). At the same time, it attempts to reduce false negatives (missing edges) by avoiding unnecessary conditions that reduce the detection power of the conditional independence tests and might result in strong faithfulness violations (see Section 4.2.8). It also partially addresses autocorrelation biases by adding an additional stage at the end of the algorithm where each pair of neighbors are tested for MCI (Section 4.2.10), which helps reduce false positives (extra edges) and gives the algorithm its name.

PCMCI assumes causal sufficiency (Section 4.2.2), so τ_{max} must be large enough to include the longest causal time lag affecting more than one variable in underlying SCM, and should be physically and statistically motivated. PCMCI additionally assumes no simultaneous dependencies, and so the orientation of edges is entirely determined by the time lag, and no orientation step is required in the algorithm. PCMCI was followed by PCMCI+ (Runge 2020), which relaxes the second assumption and allows simultaneous dependencies as long as there are no simultaneous cycles; and Latent PCMCI (LPCMCI; Gerhardus and Runge 2021), which additionally relaxes the first assumption and allows for latent confounders (Section 4.2.5). Because LPCMCI can learn confounded relationships, while a τ_{max} shorter than the longest

causal lag in the underlying SCM may in practice reduce the algorithm's effectiveness at removing autocorrelation effects and may also limit its ability to orient certain lagged adjacencies, it will not in theory produce an incorrect graph in the absence of (strong) faithfulness violations – it will simply return a graph with more confounded adjacencies and fewer lagged causal adjacencies; thus, τ_{max} can be considered an analysis choice. PCMCI+ requires orienting only simultaneous edges, but LPCMCI must orient all edges. Both PCMCI+ and LPCMCI return a PAG which is augmented in that it may contain edges with an x at one or both ends. These edges mean that the orientation rules led to conflicting orientations for that edge, betraying an inappropriate conditional independence test or a violation one of LPCMCI's assumptions – acyclicity, stationarity, or (strong) faithfulness for the tested conditions.

The possibility of latent confounding in LPCMCI complicates the algorithm because even lagged adjacencies may be confounded rather than causal, and so conditioning on lagged adjacencies may unblock colliders and cause confounding. LPCMCI thus performs preliminary iterations of the algorithm which begin to learn the diagram but then restore all removed edges while retaining orientations, allowing LPCMCI to re-test the skeleton while leveraging orientation information to avoid low detection power and reduce autocorrelation effects. This reduces the consequences of premature incorrect conditional independence tests, and increases the chances of detecting important edges once some orientations are known. The number of preliminary iterations used in the LPCMCI algorithm (p) is tunable by the user: a larger p is more computationally expensive, but could yield better results. Gerhardus and Runge (2021) test LPCMCI with 0-4 preliminary iterations[11], and find that the majority of the improvement in

[11] "k" in (Gerhardus and Runge 2021)

performance is gained with the first preliminary iteration, and further preliminary iterations yield smaller gains.

4.2.12. Evaluation of Causal Discovery Algorithms

Causal discovery algorithms are typically evaluated by how successful they are in recovering to the ground-truth DAG (Section 4.2.3) or ADMG (Section 4.2.5) that generated the examined data, both in theory and in practice.

On the theoretical side, the algorithm must be proven to be *consistent*, meaning it makes correct decisions and converges on the correct graph in the limit of infinite data. These proofs often rely on faithfulness (and all other assumptions of the algorithm) and assume an 'oracle' conditional independence test that is not constrained by the statistics of finite samples and always tells the algorithm truthfully whether the probability distributions of two variables are conditionally independent (see Section 4.2.2 for a definition of independence). Theoretical consistency results given real conditional independence tests and finite data are desired, but often difficult to prove and not provided.

On the practical side, an algorithm can be evaluated by repeatedly applying the algorithm to real data and comparing the discovered causal graph to the true underlying data-generating mechanism. Since the ground-truth causal structure for observed systems is not known, the standard is to use simulated data. The comparison can be carried out in two stages, where first the skeleton is evaluated, and then the orientation of the edges. Classical metrics for comparing an undirected graph to ground-truth knowledge include recall and precision scores. A classical recall score rewards a graph for the fraction of "true adjacencies" it contains. This measure is maximized not only for a graph that perfectly matches the ground-truth, but also for a fully-connected graph where every pair of nodes is adjacent regardless of the ground-truth. Thus, the

recall score is always balanced with precision, which evaluates a graph based on the fraction of its adjacencies that also appear in the ground-truth, and these scores can be combined into a single F1 score. Recall can be calculated without bias from partial knowledge of the positive adjacencies in the ground-truth, but precision requires full knowledge of true adjacencies.

4.3. Relevant Climate Variables and their Interactions

This chapter is motivated by previous work (Giannini and Kaplan 2019; Giannini et al. 2013; Chapter 3 of this dissertation) that defines the North Atlantic as the area-weighted mean of SST within 10°-40°N and 75°-15°W, and the Global Tropics as area-weighted mean of temperature ocean surface in the latitude band 20°S-20°N (outlined in black in Figure 4.3) from July to September. We continue using area-weighted means, but in order to approach causal sufficiency (Section 4.2.2), the STT indices should not only capture the teleconnections to the Sahel, but should also capture individual modes of climate variability and should minimize overlap with each other. In Figure 4.3, it is clear that the North Atlantic and Global Tropics overlap with each other. Furthermore, the Global Tropics includes a number of distinct modes of internal climate variability, including the El Niño Southern Oscillation in the tropical Pacific, the Atlantic Niño in the tropical Atlantic, the Atlantic Meridional Mode in the North Tropical Atlantic, and the Indian Ocean Basin Mode and Indian Ocean Dipole. Thus, we construct a new set of climate indices inspired by the original definitions of NARI and the Sahel. Table 4.2 contains the complete definitions of the SST indices used in this chapter (black), and Figure 4.3 gives a visualization of the basins.

Table 4.2: SST (black) and other (gray) indices used in this chapter.

Coordinates	Season	Name	Abbrev.
140°E-80°W and 20°S-20°N	January-March	El Niño	EN
	JAS	Tropical Pacific	Pc
28°W-20°E, 35°S-15°N and the lines 6*lat-5*lon=164° and 6*lat+5*lon=116°	May-June	Gulf of Guinea	GG
30°W-20°E and 20°S-10°N	JAS	South Atlantic	SA
60°-15°W and 10°-20°N	March-May	Atlantic Meridional Mode	AMM
	JAS	North Tropical Atlantic	TA
20°S-20°N	JAS	Global Tropics	GT
75°-15°W and 20°-40°N	JAS	North Atlantic	NA
6°W-36°E and 30°-40°N	JAS	Mediterranean Sea	Md
40°-100°E and 20°S-20°N	JAS	Indian Ocean	IN
12°-18°N and 20°W-40°E, 150 hPa	JAS	Upper-Tropospheric Temperature	TT
12°-18°N and 20°W-40°E, surface	JAS	Sahel precipitation	Pr

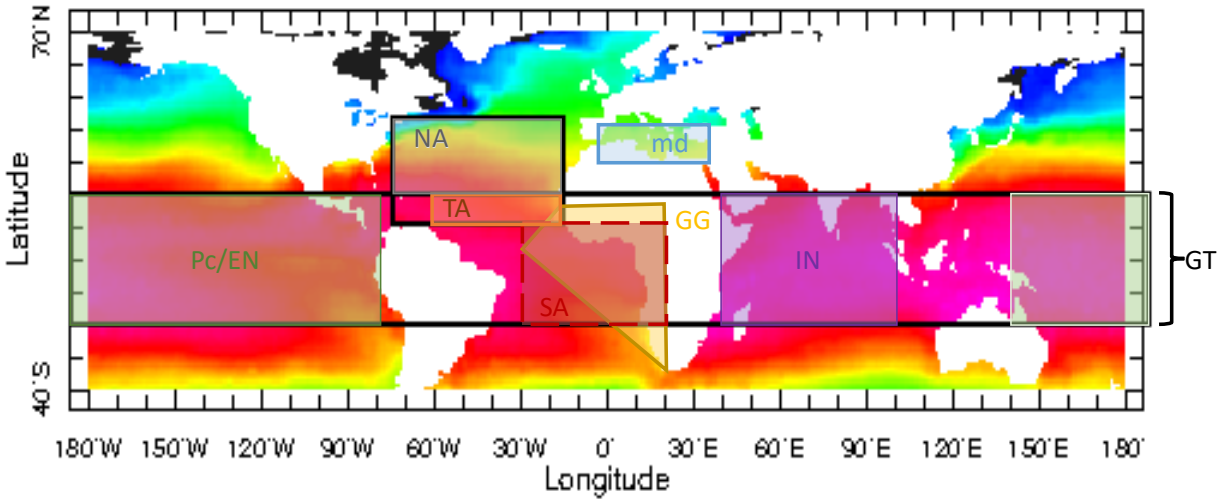

Figure 4.3: Visualization of the SST basins (defined in Table 4.2) used in this chapter.

To avoid overlapping indices, we split G13's North Atlantic index into two: the part that overlaps with the Global Tropics from 10°-20°N (additionally restricted to 60°-15°W, orange), and the rest from 20°-40°N (gray). The former is often called the North Tropical Atlantic, and is the location of decadal SST variability known as the Atlantic Meridional Mode (Nobre and Shukla 1996), which peaks in the spring between March and May. We examine SST in this box

both in the spring (March-May, "AMM") and in the summer (JAS, "TA"), when its temperature might affect the humidity and intensity of the West African Westerly Jet (Pu and Cook 2012), which brings moisture from the North Tropical Atlantic to the Sahel during the rainy season (unfortunately, the jet lies between 8º and 11ºN, just on the southern boundary of TA). The other part of G13's North Atlantic index will be designated the North Atlantic (NA) for the purposes of this study. It is meant to capture low-frequency variability in the North Atlantic Ocean basin associated with the Atlantic Multidecadal Variability (AMV or AMO; Knight et al. 2005). Though it excludes variability in the subpolar North Atlantic north of 40N that is also associated with the AMV, it captures the variability from 20-30N shown to be most strongly associated with the Sahel in simulations (Martin et al. 2014).

To capture independent modes of variability, we keep the original definition of the Global Tropics (GT), but additionally split it up into component basins designed to capture each distinct mode of internal variability. We intentionally do not include an index for Indonesian SST, and also never include SA and IN at the same time, to make sure that GT has variability not explained by its constituent basins, as required by our definition of an SCM (Section 4.2.2).

The EN and Pc indices cover the tropical Pacific (green), and are designed to capture El Niño – the most dramatic and influential mode of internal variability in the global oceans (Wang et al. 2017). It is accompanied by a pressure oscillation in the southern Pacific and Indian oceans which is called the Southern Oscillation (Bjerknes 1969), and the coupled phenomenon is called the El Niño Southern Oscillation (ENSO). It develops in the summer, which is also when it is believed to suppress Sahel rainfall in observations (Joly and Voldoire 2009). This development phase is captured in the Pc index, defined from July to September. El Niño events peak in winter, and this is captured by our EN index defined from January to March.

The Gulf of Guinea index (GG, yellow) is designed to roughly capture the Atlantic Ocean portion of the "precipitationshed" for Sahel rainfall identified by Keys et al. (2014), where the precipitationshed is the land and ocean surface where at least 5mm of evaporated moisture eventually precipitates in the Sahel during the monsoon. It is defined in spring from May to June to capture the atmospheric river that transports moisture from the Gulf of Guinea to the Sahel (Lélé et al. 2015). This index also captures the first mode of interannual variability in the tropical Atlantic, known as the Atlantic Niño (Ruiz-Barradas et al. 2000), which peaks in May and June (see Figure 1.5 in Mechoso et al. 2023) and may affect the latitudinal position of the African rainband (Rodríguez-Fonseca et al. 2011, and references therein). It additionally captures the development of the seasonal tropical Atlantic cold-tongue, which begins in April and is strongly associated with the onset and development of the WAM (Okumura and Xie 2004). The cold tongue extends into July, affecting the South Atlantic index (SA, red), which is defined from July to September to represent, in combination with NA, the argument that summer Atlantic SST gradients affect the latitudinal location of the African rainband (e.g. Zeng 2003). Since this index overlaps with Sahel rainfall, coupled cold tongue dynamics may mean that convection in the Sahel also causes increased wind speeds over SA and reduces SST there, in which case there would be a causal cycle in the graph, violating the assumptions of causal discovery.

The IN index (purple) captures the basin-wide Indian Ocean SST variability believed to affect the Sahel (e.g. Bader and Latif 2003; Giannini et al. 2003). It is defined from July-September, and captures the second peak of the Indian Ocean Basin Mode (IOBM), which persists into summer (Du et al. 2009). It does not capture the Indian Ocean Dipole[12], in which the west and east portions of the ocean basin have opposite sign anomalies, unless IOD variability is

[12] The existence of this mode is disputed (Zhao and Nigam 2015).

associated with positive IOBM variability. The IOBM displays both high- and low-frequency variability (Han et al. 2014), both of which have been argued to be important for the Sahel (Bader and Latif 2003, 2011).

The definition of the Mediterranean Sea captures the vertical extent determined by Rowell (2003) to have a significant statistical relationship with the Sahel in observations, and extends it to cover the width of the basin to capture correlation with precipitation in simulations. Another choice could be the full height of the East Mediterranean, as identified by (Gaetani et al. 2010). The Mediterranean Sea is thought to impact Sahel rainfall via thermodynamic changes in moisture supply and also by causing a dynamic shift in the location of the rainbelt via changing temperature gradients over Africa.

There are other modes of internal variability that we are neglecting in our analysis, including the North and South Pacific Meridional Modes (Chiang and Vimont 2004; Zhang et al. 2014), the Pacific Decadal Oscillation (Mantua et al. 1997), the Southern Annular Mode (Thompson and Wallace 2000), the North Atlantic Tripole (Fan and Schneider 2012), and an atmospheric phenomenon called the North Atlantic Oscillation (Hurrell 1995). If these modes affect Sahel rainfall (e.g. Villamayor and Mohino 2015) and/or other modes of variability, then we might not achieve causal sufficiency.

In Figure 4.4, we present expected simulated interactions between these basins and Sahel precipitation using the language of time-series summary graphs (Section 4.2.10). While we cannot assume stationarity of the underlying SCM between seasons, we can assume stationarity from year to year for the piC simulations, and so we include a node for every examined season for every basin, and treat all indices within the same calendar year as "simultaneous" in our summary diagrams and the causal discovery algorithms. The indices are grouped by season, with

winter (W, JAS) in the top row, spring (Sp, March-May/May-June) in the second row, and summer (Su, JAS) in the third row; and also by location (from left to right: the Indian Ocean, the tropical Pacific, the tropical Atlantic, the North Atlantic, the South Atlantic and Gulf of Guinea, and the Mediterranean Sea). Beneath the first three rows is the Global Tropics (GT), defined to encompass the Pacific, South Atlantic, Tropical Atlantic, and Indian Oceans. The non-causal "ontological" relationships between GT and its constituent basins are notated with black arrows. Beneath GT are atmospheric variables including Sahel tropospheric temperature (TT) and precipitation (pr). Throughout the diagram, straight arrows represent "simultaneous" dependencies in the same calendar year, while curved arrows represent "lagged" dependencies from one year to the next, and colored circles represent the dependence of a variable on itself from the previous year. While SST in every ocean basin should depend on its previous value because ocean circulation is slow and water has a high heat capacity, only IN, NA, and md have red circles because the other auto-dependencies are unrolled over different seasons, and consist of one simultaneous edge and one lagged edge.[13] In green arrows, we present interactions we expect to see between ocean basins based on the observed climate system, as summarized by Mechoso et al. (2023).

[13] In some ocean basins around the world, seasonal SST can affect SST one year later directly—without mediation by the intermediate seasons—via subsurface temperatures that resurface seasonally (Hanawa and Sugimoto 2004). But in the tropics, we expect the autocorrelation of SST in a given basin to be mediated by the intervening seasons.

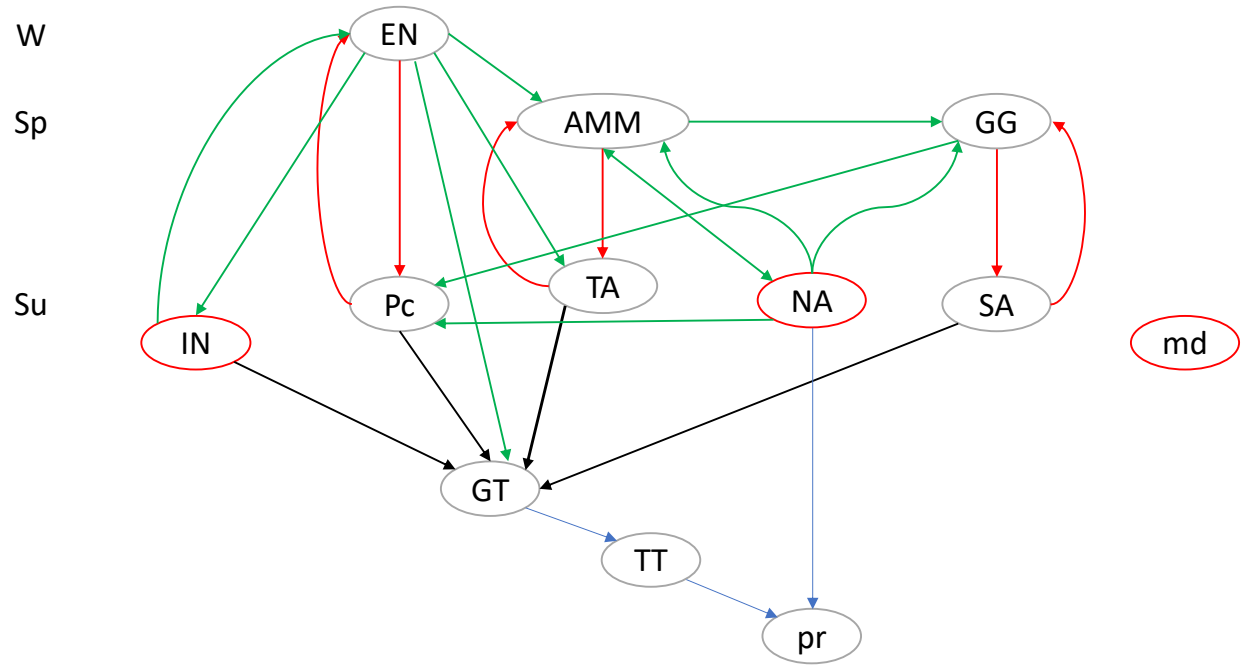

Figure 4.4: Summary graph containing relationships between ocean basins and relevant atmospheric variables identified in the literature. Basins are organized according to season, with winter (W) in the top row, spring (Sp) in the second row, and summer (Su) below. Some basins are similar or identical to each other, and are grouped by column; from left to right: Indian Ocean (IN), tropical Pacific (EN and Pc), tropical Atlantic (AMM and TA), North Atlantic (NA), Gulf of Guinea or South Atlantic (GG and SA), and Mediterranean Sea (md). The fourth row contains the Global Tropics (GT), which encompasses the Pacific, South Atlantic, and Tropical Atlantic, as well as the Indian Ocean, and below that are Sahel tropospheric temperature (TT) and precipitation (pr). Straight arrows represent dependencies within the same year, while curved arrows represent time-lagged dependencies from one year to the next between variables and red circles represent auto-dependence from year to year. The colors differentiate the reasons why we expect to see an edge: red is given (dependence of an ocean basin on itself within the last year), black is by construction, green is based on observations and theory, and blue is the hypothesis of G13.

ENSO is believed to affect global atmospheric and oceanic states three to six months after its peak (EN), including the North Tropical Atlantic and the Indian Ocean in spring (Sp) and summer (Su) (Klein et al. 1999), so we draw green arrows from EN to AMM, TA, and IN, and also GT, in case other areas of the Global Tropics are also effected. These effects are mediated by tropospheric temperature in the tropics: convection transports heat from warm tropical oceans to the troposphere, and gravity waves transmit the local increases in tropospheric

temperature both eastward and westward until tropical tropospheric temperature is nearly uniform (Sobel et al. 2001), obeying the "Weak Temperature Gradient" constraint (see section 1.4). These anomalies then are transmitted to cooler remote ocean basins via downwelling and turbulent surface fluxes (Parhi et al. 2016) in an "atmospheric bridges" mechanism (Alexander et al. 2002). This mechanism transmits temperature anomalies from whichever tropical ocean basin happens to be the warmest to the rest of the Global Tropics, but the Pacific is considered the main driver because it has the most dramatic variability and is often the warmest basin in the Global Tropics during El Niño.

Some have argued that ENSO events can be triggered by a warm phase of the Pacific Meridional Mode (PMM) the preceding spring (Alexander et al. 2010). Others have argued that the PMM affects the location of the warm El Niño anomalies (Yu et al. 2010), which may change the effect of ENSO on other parts of the world (Ashok et al. 2007). We have not included the PMM explicitly, but the SST pattern extends to 10°S and may be somewhat captured in the Pc index. Unfortunately, however, our EN and Pc indices will likely not be able to distinguish between the types of ENSO events because we take an average over the whole basin. PMM may respond to the Atlantic Multidecadal Variability (Yu et al. 2015), so we connect NA to Pc.

AMV may also modulate interannual modes in the tropical Atlantic (Martín-Rey et al. 2018), so we connect NA to the following year's AMM and GG. Some have argued that the Atlantic Niño can reduce the likelihood of El Niño (Losada et al. 2010; Rodríguez-Fonseca et al. 2009); accordingly, we draw an arrow from GG to Pc. There is disagreement about whether (Ham et al. 2013) or not (Zhang et al. 2021) the AMM can also trigger ENSO; for now, we do not connect them. It has also been argued that the AMM dampens the Atlantic Niño between May and July via an atmospheric teleconnection (Foltz and McPhaden 2010); accordingly, we

add an arrow from AMM to GG. The AMM is thought to be driven by the atmospheric North Atlantic Oscillation (NAO; Hurrell 1995) in addition to ENSO (Enfield and Mayer 1997), and the AMV also may be driven by the NAO (Clement et al. 2015), so AMM and NA are also confounded (double-sided green arrow). There may also be an oceanic connection between the Atlantic basins since surface waters flow northward through the Atlantic Oceans. Though the oceanic connection may not produce the spatial pattern of variability associated with the Atlantic Niño, the Atlantic Meridional Mode, or the AMV, it may still affect our area-mean SST indices, giving the connection SA → TA → NA (not pictured in our diagram). Though AMM and GG overlap in May, we hope that the fact that GG is defined later than AMM will mostly prevent any flow of information from GG to AMM which would conflict with the 'simultaneous' atmospheric teleconnection under the assumption of acyclicity.

The IOBM may play a role in determining the persistence of ENSO events via an atmospheric pathway (Okumura and Deser 2010), so we draw an arrow from IN to EN in the following year. The Indian Ocean is connected to the Pacific Ocean via the Indonesian Throughflow, and so there could also be oceanic communication between these basins. For instance, it has been proposed that the Pacific Decadal Oscillation affects the IOD mode with a delay of about 10 years (Ummenhofer et al. 2017), but this time lag is too long to be captured in our analysis and is not pictured in our hypothesis. The Indian Ocean is also oceanically connected to the Gulf of Guinea via the Agulhas current, but the time delay is also likely too long to be observed in our analysis. If the Indian Ocean affects the NAO (Bader and Latif 2003), then there may be a dependence of NA and AMM on IN as well; these links are not included in our diagram.

The Mediterranean Sea (md) is correlated with NA in observations. There is outflow of Mediterranean water to the Atlantic, but it is usually identified by salinity and not with a temperature profile (Zhao and Nigam 2015). Variability in md has generally been poorly understood, but some studies links it to the AMV via atmospheric (Mariotti and Dell'Aquila 2012) and oceanic (Skliris et al. 2012) pathways, while others link it to the NAO (Yan and Tang 2021). Thus, md may affect NA with a lag, be affected by NA, and/or be confounded with NA and AMM. For now, we leave it disconnected.

In addition to these SST indices, we examine precipitation (pr) and upper-tropospheric temperature (TT) over the Sahel (gray in Table 4.2), defined as in Chapter 2 and Chapter 3 to be the area within 12°-18°N and 20°W-40°E. If G13's view of Sahel rainfall is correct, then we would see a connection from NA to Sahel rainfall, and from GT to the upper tropospheric temperature over the Sahel (TT) and then to Sahel precipitation (blue arrows). According to other hypotheses we discussed in the Introduction, we may see direct connections to other basins as well or instead. TT is defined over the Sahel and in summer to minimize its impact on other ocean basins so it can act solely as a mediator for the effect of remote SST on Sahel rainfall. However, since warm tropospheric temperature anomalies spread both eastward and westward throughout the tropics, TT may still be detected as a mediator for the effect of winter and spring oscillatory modes on summer tropical SST. If this is the case, the fact that TT also responds to tropical SST on shorter than seasonal timescales (Zhang and Fueglistaler 2020) may pose a problem for causal discovery because cyclic adjacencies are not allowed. We also hope to define TT high enough in the atmosphere that it will affect Sahel precipitation without responding to it, but if we are unsuccessful, TT and pr may also be coupled.

4.4. Methods

It is likely that simulations from different climate models differ causally from observations and from each other (see Section 4.1). Furthermore, it is not clear a priori how to choose the best parameters for statistical conditional independence testing (see Section 4.2.9) that forms the backbone of causal inference. In light of these considerations, rather than attempting to validate the conditional independencies implied by our hypothesized causal relationships (see Section 4.2.7 and Figure 4.4) in climate simulations, we use causal discovery (see Section 4.2.8) for time series (see Section 4.2.10) to learn the causal structure of individual climate models. As mentioned in Section 4.3, we expect causal relationships to vary seasonally, so in order to satisfy stationarity (see Section 4.2.10), we treat indices defined at all seasons within the same calendar year as "simultaneous." This hides some background knowledge from the causal discovery algorithm that could have improved its performance, but it also gives us some independent ground-truth knowledge about a subset of the underlying causal relationships with which to evaluate the performance of the method (see Section 4.2.12). Because we likely do not include all causally-relevant factors (see Section 4.3), we choose an algorithm that allows latent confounding (see Section 4.2.5). As mentioned in the Introduction, we focus on coupled simulations with constant external radiative forcing (see Section 4.4.1) to reduce latent confounding and stationarity violations. Future work can use the discovered causal structures specific to each climate model to properly estimate the causal effects (see Section 4.2.6) of SST in various ocean basins on Sahel precipitation in each simulation, and those results can then be validated against atmospheric simulations with prescribed historical SST from those same climate models (with the caveat that if the system is not linear and the coupled simulations do

not cover the distribution of observed historical SST, then the learned relationships may not transfer well from the coupled simulations to the atmospheric simulations).

4.4.1. Data

We use pre-Industrial control (piC) simulations from the Coupled Model Intercomparison Project phase 6 (Eyring et al. 2016), in which all anthropogenic emissions and other external radiative forcings—factors that affect the climate system but are not caused by the climate system—are held constant at pre-Industrial levels. Table 4.3 lists the piC simulations used in this chapter. We focus on climate models that also provide amip-piForcing simulations (Webb et al. 2017)—which are long atmospheric simulations with prescribed observed historical SST and no external radiative forcing—that can be used to validate causal effects estimated using the results of this chapter, and we include a few additional models for tuning (see Section 4.4.4).

Table 4.3: Simulations used in this chapter, along with the sample size (N) and the largest autocorrelation at a lag of one year for any included variable in a given season (r_{max}). The first 5 models were chosen for tuning algorithm parameters. The last 5 models were chosen to explain the performance of the amip-piF simulations analyzed in Chapter 3.

	GCM	Run	N	r_{max}
For tuning	AWI_AWI-CM-1-1-MR	r1i1p1f1	500	0.18
	CMCC_CMCC-ESM2	r1i1p1f1	500	0.66
	AS-RCEC_TaiESM1	r1i1p1f1	500	0.29
	CAMS_CAMS-CSM1-0	r1i1p1f1	250	0.29
	CSIRO_ACCESS-ESM1-5	r1i1p1f1	1000	0.30
Provide amip-piF simulations	CCCma_CanESM5	r1i1p2f1	1051	0.34
	CNRM-CERFACS_CNRM-CM6-1	r1i1p1f2	500	0.35
	IPSL_IPSL-CM6A-LR	r1i1p1f1	1200	0.30
	MRI_MRI-ESM2-0	r1i1p1f1	701	0.30
	NCAR_CESM2	r1i1p1f1	1200	0.37

We examine area-weighted means of SST in the regions listed in Table 4.2. Initial results exclude IN, and later results include it. We chose 150 hPa (which is 50 hPa higher than the approximate tropical tropospause pressure, Figure 4.5a) as the level for TT because it maximized

the absolute value of the correlation of Sahel precipitation with the difference in moist static energy between the upper troposphere and near the surface (b, red) and of GT with TT across CMIP6 historical simulations (c). (If we were to repeat the analysis, we would use piC simulations instead.) Though correlation may not maximize at the height most associated with the causal mechanism, we hope that this choice will provide a reasonable signal-to-noise ratio.

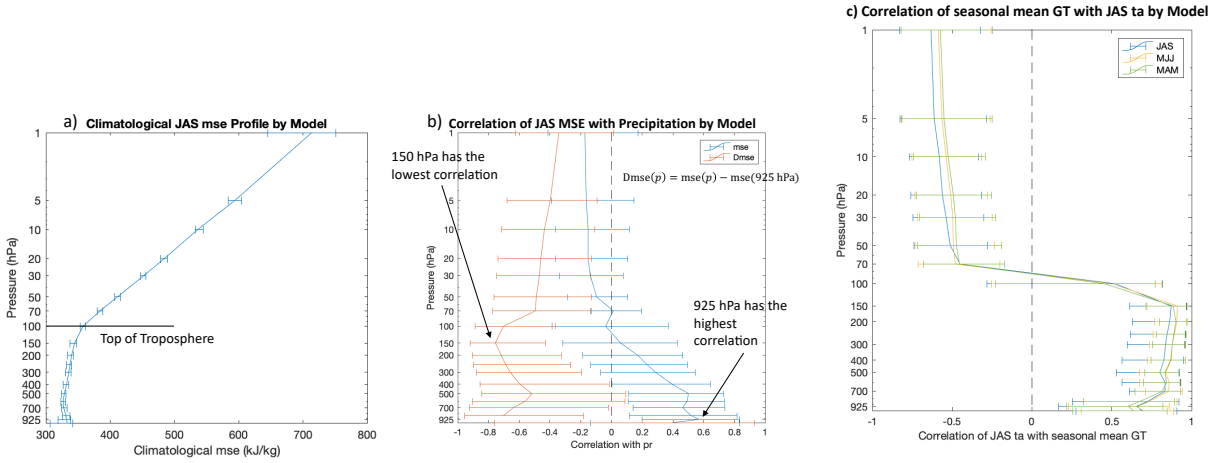

Figure 4.5: Ensemble-mean and variance of Sahelian profiles of time-mean thermodynamic quantities across CMIP6 historical simulations. (a) JAS moist static energy (MSE, estimated with c_{pd}) profile. (b) Correlation of Sahel precipitation with MSE (blue) and with the difference in MSE at a given pressure and at 925 hPa (red, Dmse). (c) Correlation of GT in spring (green, March-May), late spring (yellow, May-July), and summer (blue, JAS) with air temperature.

4.4.2. Code

We study piC simulations using the "Latent PCMCI" (LPCMCI; Gerhardus and Runge 2021) causal discovery algorithm and also employ PCMCI+ (Runge 2020), both of which are implemented in a python package called "tigramite". This package supports multiple kinds of statistical conditional independence tests (see Section 4.2.9). For continuous dependencies, there are three of note. First, there is a partial correlation test (ParCorr) for linear dependencies with Gaussian noise, and two variants: RobustParCorr for different marginal distributions, and ParCorrWLC for heteroskedastic data. Second, there is a distance correlation test (GPDC). Since

distance correlation is currently only implemented for unconditional dependencies, it first removes the effect of the conditioning set on the variables of interest using a Bayesian procedure called Gaussian process regression (Williams and Rasmussen 2006), which assumes that all marginal distributions are Gaussian. Finally, there is a k-nearest-neighbor estimator of conditional mutual information (CMIknn; Runge 2018b) that assumes no parametric form and can be used for continuous data with general dependencies. We use the High-Performance Computing resources at Columbia University to run parallel iterations of this algorithm.

4.4.3. Parameter Choices

Adjustable parameters in PCMCI+ include the significance level for the conditional independence test (α, see Section 4.2.8) and the maximum time lag (τ_{max}, see Section 4.2.10). LPCMCI includes these parameters and one more – the number of preliminary iterations (p, see Section 4.2.11). Furthermore, all causal discovery algorithms implemented in tigramite support multiple choices of conditional independence test (see Section 4.4.2) which may come with additional parameter choices (see Section 4.2.9).

To determine the best choice for τ_{max}, we perform a bivariate conditional independence testing procedure (described in Section S3.2 of Runge et al. 2019a) which evaluates $X_i(t-\tau) \perp\!\!\!\perp X_j(t)|\{X_j(t-\tau')|\tau' \in 0, \dots, \tilde{\tau}\}$ for all pairs of variables X_i and X_j at a range of lags $0 \leq \tau \leq \tilde{\tau}$ ($\tau = 0$ is excluded when $i = j$). This test conditions on the past of X_t as would be done in Granger causality, so it helps reduce the effects of autocorrelation to give a rough idea of the important lags. However, it may miss some true dependencies due to low detection power that derives from the high dimensionality of the conditional independence tests, and it may contain spurious non-causal dependencies that result from autocorrelation or that cannot be removed or properly oriented without a proper causal discovery algorithm. For this procedure, we use

CMIknn (Runge 2018b) – the most general test available in tigramite (see Sections 4.2.9 and 4.4.2) – with its default parameters, and examine lags up to $\tilde{\tau} = 5$. For almost all climate models examined in this chapter that provide amip-piF simulations, the bivariate CMI is consistently near 0 for time lags $\tau > 1$, so we focus on $\tau_{max} = 1$ in this chapter. For CMCC and AS-RCEC (two of the tuning simulations), the bivariate CMI between EN and itself at a lag of $\tau = 2$ years is greater than or roughly equivalent to the bivariate CMI at $\tau = 1$, respectively. We thus tested $\tau_{max} = 2$ as well for AS-RCEC.

To motivate our choice of conditional independence test for LPCMCI, we examine scatter plots for each pair of variables at the time lag that produces the maximum CMI. While most variables appear to be linearly related or completely uncorrelated (not shown), most of the simulations we examine have some relationships that appear to be nonlinear and non-Gaussian. For instance, Figure 4.6 displays a non-linear relationship between EN and itself in the following year in NCAR. Consistent with the oscillatory nature of the El Niño Southern Oscillation, some years show a positive relationship with their predecessors and others show a negative relationship with their predecessors, giving the resulting joint density distribution a complex shape. Not only is this process non-linear, but it is also non-Gaussian, because drawing a vertical or horizontal line through the distribution (equivalent to conditioning on $\mathrm{EN}(t+1)$ or $\mathrm{EN}(t)$, respectively) can yield a distribution with multiple peaks. This suggests that partial correlation and GPDC would be poor choices of conditional independence test for this data. In fact, a univariate linear or Gaussian process regression will not be able to capture a second-order relationship because the complete description must include the time-derivate of the variable in addition to the value of the variable. Thus, we prefer the CMIknn conditional independence test,

even though it has relatively low detection power, meaning it is more likely to find independence when variables are actually dependent (Runge et al. 2019a).

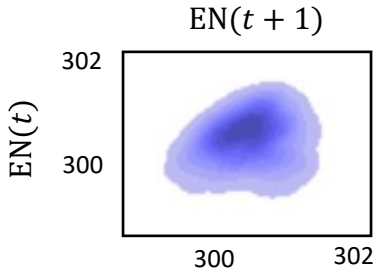

Figure 4.6: Scatter plot of EN(t) with EN($t + 1$) in kelvins for NCAR. The relationship of EN to itself appears to be non-linear and non-Gaussian.

CMIknn has a tunable parameter that affects its null hypothesis (SN, see Section 4.2.9), so for LPCMCI we stick to the default significance level $\alpha = 0.05$. We tune the number of preliminary iterations in LPCMCI (p) along with CMIknn's parameters (see the next section).

4.4.4. Tuning

For tuning, we chose simulations that will not be used in our analysis and have varying time series length (N, second to last column, bottom part of Table 4.3). Time series length is roughly equivalent to sample size, but autocorrelation may reduce the degrees of freedom and the effective sample size, so we also choose simulations with varying autocorrelation, estimated by the maximum autocorrelation at lag 1 over all variables (r_{max}; last column of Table 4.3). We use 11 variables, including all the indices listed in Table 4.2 but the Indian Ocean, which was initially excluded. We test CMIknn parameters knn $= 0.1, 0.15, \ldots 0.6$ and SN $= 5, 6, \ldots 25$, and LPCMCI parameter p$= 1, 2 \ldots 4$ and search for a neighborhood of parameter space that consistently performs well. We hope to find a consistent set of preferable parameter choices for climate simulations with different characteristics, or a trustable dependence of preferable parameter choices on the time series length (sample size) and maximum autocorrelation.

We evaluate the performance of LPCMCI by comparing the resulting causal diagram to our prior knowledge and beliefs as summarized in Figure 4.4. We have high confidence that all simulations will demonstrate auto-dependence of (similar) SST indices in time (red circles and arrows) and the ontological relationships that result from our definitions of the basins (black arrows), but the green adjacencies in our hypothesized causal diagram are likely to differ from the true causal structure in any given climate model. We could use our partial knowledge of the true adjacencies to estimate recall without bias, but we cannot estimate precision without complete knowledge of the graph (see Section 4.2.12). Luckily, as explained in Section 4.2.8, edges cannot be oriented in causal discovery in the absence of triples $A - B - C$ where A is not adjacent to C, so we can use our knowledge of the orientation of our known adjacencies to reward a graph for appropriately removing edges.

We define a new score for evaluating the output of our causal discovery algorithm relative to *partial* background knowledge and call it the *oriented-recall* score. A graph's oriented-recall score is the fraction of known adjacencies it contains, weighted by their orientation relative to the ground-truth. For illustration, say the ground truth is a right adjacency (→). A matching directed edge (→) receives one point, a consistent partially-oriented edge (o→) receives 2/3 of a point, and an unoriented adjacency (o−o) receives 1/3 of a point, while incorrectly-oriented edges (←, ←∘) receive no points. While a fully-connected undirected graph would receive a classical recall score of 1, it will only receive an oriented-recall score of 1/3. The fully-connected graph will be outperformed by graphs that remove other edges when this results in correct orientation of known adjacencies, and it will outperform graphs that remove other adjacencies when this leads to incorrect orientation of known adjacencies. Graphs with x's

anywhere are assigned a score of 0 because they violated some assumption of LPCMCI (see Section 4.2.11).

By construction, we also know that any "simultaneous" edge that points backwards in time within the calendar year must be a false edge or a false orientation. It is difficult to use this knowledge to evaluate performance in a continuous and fair way when we don't know if the edge should be present or not, so we do not use this knowledge for tuning, but we will take advantage of this knowledge in the next section.

4.4.5. Graph Selection Criteria

We define a *likely-accurate* score, which tests a PAG's accuracy (the ratio of true positive and negative detected adjacencies to total adjacencies) relative to a distribution of "likely" ground-truth adjacencies constructed from our full background knowledge and from the ensemble of discovered PAGs within the chosen neighborhood of parameter space. This score will identify the PAG most likely to represent the underlying causal structure of the simulations.

To begin, we estimate the likelihood of each type of adjacency (edge or lack thereof) for each pair of variables at each time lag, using the language of maximal ancestral graphs (MAGs), which admit exactly four types of adjacencies: right (→), left (←), confounded (↔), and no edge, defined as they are for PAGs (see Section 4.2.7). The likelihood assigned to each possible adjacency is the fraction PAGs discovered with parameters from within the ideal parameter neighborhood (identified according to Section 4.4.4) that contain an adjacency consistent with it. For the "no edge" MAG adjacency, this only includes PAGs without an edge connecting the variables in question. For the other three MAG adjacencies, this includes discovered PAGs that connect the variables in question with a fully- or partially-oriented edge that matches the PAG adjacency wherever it has no circle (the reader may consult the first two rows and the last three

columns of Table 4.1). The resulting likelihood distribution for a given pair of nodes at a given time lag will exceed 1 when PAG edges are partially-oriented; this was a conscious choice because we want to learn from, rather than disadvantage, any PAGs with a slightly different skeleton that were able to orient that adjacency. This edge-oriented approach is simplified causally in that pairs of edges with neighboring circles cannot always be oriented independently (see Section 4.2.7 and Zhang (2008)), and so our score may sometimes reward one graph for appearing to locally match another graph that is actually locally inconsistent with it; nevertheless it seems a simple and reasonable representation of our interpretation of the causal mechanisms at each adjacency. When performing this calculation, we entirely exclude any PAG that anywhere contains an x (reflecting a violation of LPCMCI's assumptions) or an edge pointing backwards in time within the calendar year (which violates our background knowledge): we do not want these PAGs to affect the likelihood of any adjacency because the conflict or false orientation may have resulted from incorrect adjacencies and orientations in other parts of the graph, and a false orientation may also lead to further false orientations elsewhere. To explicitly incorporate our background knowledge in the edge likelihoods, we override the likelihood of MAG adjacencies that correspond to "expected" red and black edges from Figure 4.4, setting the likelihood of a fully-oriented adjacency → to 1 and the others to 0, and also override the likelihood of all edges pointing backwards in time, setting them to 0.

The likely-accurate score for each discovered PAG is an aggregate over all pairs of variables and time lags of the likelihood of each potential MAG adjacency consistent with the discovered PAG adjacency (including no edge). The likelihood distributions for the MAG adjacencies sum to more than 1, but the likelihood of any given adjacency does not exceed 1, so each PAG adjacency earns a weighted average (rather than a sum) of the likelihood for each type

of MAG adjacency consistent with it: a PAG edge with two circles ∘ – ∘ receives an unweighted average over the likelihood of the forwards →, backwards ←, and confounded ↔ MAG adjacencies; and a partially-oriented PAG edge with one circle on its left and an arrowhead on the other side ∘→ receives a weight of 2/3 on the likelihood of the forwards MAG adjacency → and of 1/3 on that of the confounded MAG adjacency ↔. The forwards MAG adjacency is weighted more highly than the confounded adjacency because the partially-oriented PAG edge is consistent with the existence of an unobserved confounder, a forwards causal relationship, or both, and the last two would both be expressed with a forwards adjacency in a MAG (see Table 4.1). Finally, the score is offset by the average number of non-adjacencies (the sum of the likelihood of no edge for each pair of nodes) and normalized by the average number of edges in the PAGs selected for that climate simulation, so that scores are comparable across graphs of different dimensionality and connectivity and never exceed 1. We assign likely-accurate scores to LPCMCI results from the entire parameter space – not just the ideal parameter space chosen to have high oriented-recall (see Section 4.4.4). The reason is that though oriented-recall attempts to reward graphs for appropriately removing edges, it is still a recall score, and may be biased toward having too many edges (see Section 4.2.12). If each graph in the ideal parameter space has a different false positive edge, then a graph from outside the 'ideal' parameter space with slightly lower detection power may be a better representative of the robust qualities of the graphs used to create the edge likelihood scores. The graphs with the highest likely-accurate score for each climate model will constitute our results from a scientific point of view.

4.4.6. Performance and Robustness

Oriented-recall measures the performance of LPCMCI according to adjacencies required by our partial background knowledge (see Section 4.4.4) To evaluate our trust in other parts of

the graph, we re-calculate the likely-accurate scores for PAGs from within the ideal parameter space completely without background knowledge: no graphs are excluded from the ideal parameter space, no adjacency likelihoods are overwritten, and graphs are not assigned a score of zero when they contain a backwards edge. We call this the *probably-accurate* score, and compare it to the original likely-accurate scores to evaluate the likelihood of having at least one incorrectly-oriented edge outside our background knowledge in high-scoring graphs. We focus on the likelihood of having at least one edge rather than the number of incorrectly-oriented edges because one false orientation is likely to lead to more, depending on the connectivity of the graph. To the extent that false orientations are associated with detectable time-backwards orientations and are disassociated with high oriented-recall, we may alternately estimate the likelihood of an incorrect orientation by analyzing the prevalence of backwards edges in graphs with high likely-accurate scores that are not set to 0 when the graph contains a backwards edge. We will call this the "full" likely-accurate score.

We can evaluate the robustness or our chosen graph by re-calculating the likely-accurate scores while refraining only from overwriting adjacency likelihoods according to our background knowledge. The resulting score represents the similarity of our graph to the actual graphs in the ideal parameter neighborhood rather than to the set of most likely adjacencies, and we call this score *robustness.*

4.5. Results

4.5.1. Tuning

Figure 4.7 displays the oriented-recall scores when $\tau_{max} = 1$ over all indices listed in Table 4.2 aside from IN for the climate models chosen for tuning, organized from left to right by increasing time series length (N) and from bottom to top by increasing maximum autocorrelation

(r_{max}), so that effective sample size increases moving down and to the right. Within each subplot, each of the three axes corresponds to one of the tunable parameters, and each dot on the grid corresponds to a PAG learned by LPCMCI under those parameters. Where a dot is absent, the algorithm failed to converge in a reasonable amount of time (knn=.2 is an exception; it was sometimes excluded from our analysis by mistake). Where the dots are present, the color represents the oriented-recall score. Blue dots received a score of 0 because the graph had at least one edge with an x, meaning orientation rules gave conflicting results for that edge and there must have been a violation of assumptions. The best-performing graphs are presented in yellow.

For all simulations, we make the following observations. First, as expected, increasing the number of preliminary iterations in LPCMCI (p, y axis) never hurts performance. Some simulations show dramatic improvement for $p = 2$ relative to $p = 1$, and none show much improvement after that, so $p = 2$ is desired and larger p may require more resources than necessary. Second, SN (z-axis) does not have a large impact on the performance of LPCMCI as long as it is large enough, where values between 10 and 15 are sufficient for the different simulations. The fact that increasing SN further doesn't eventually reduce performance suggests that SN does not control the null distribution completely enough to encompass the effect of changing α (the significance parameter of LPCMCI), and future work should focus on tuning α as well. Finally, knn has a large impact on performance.

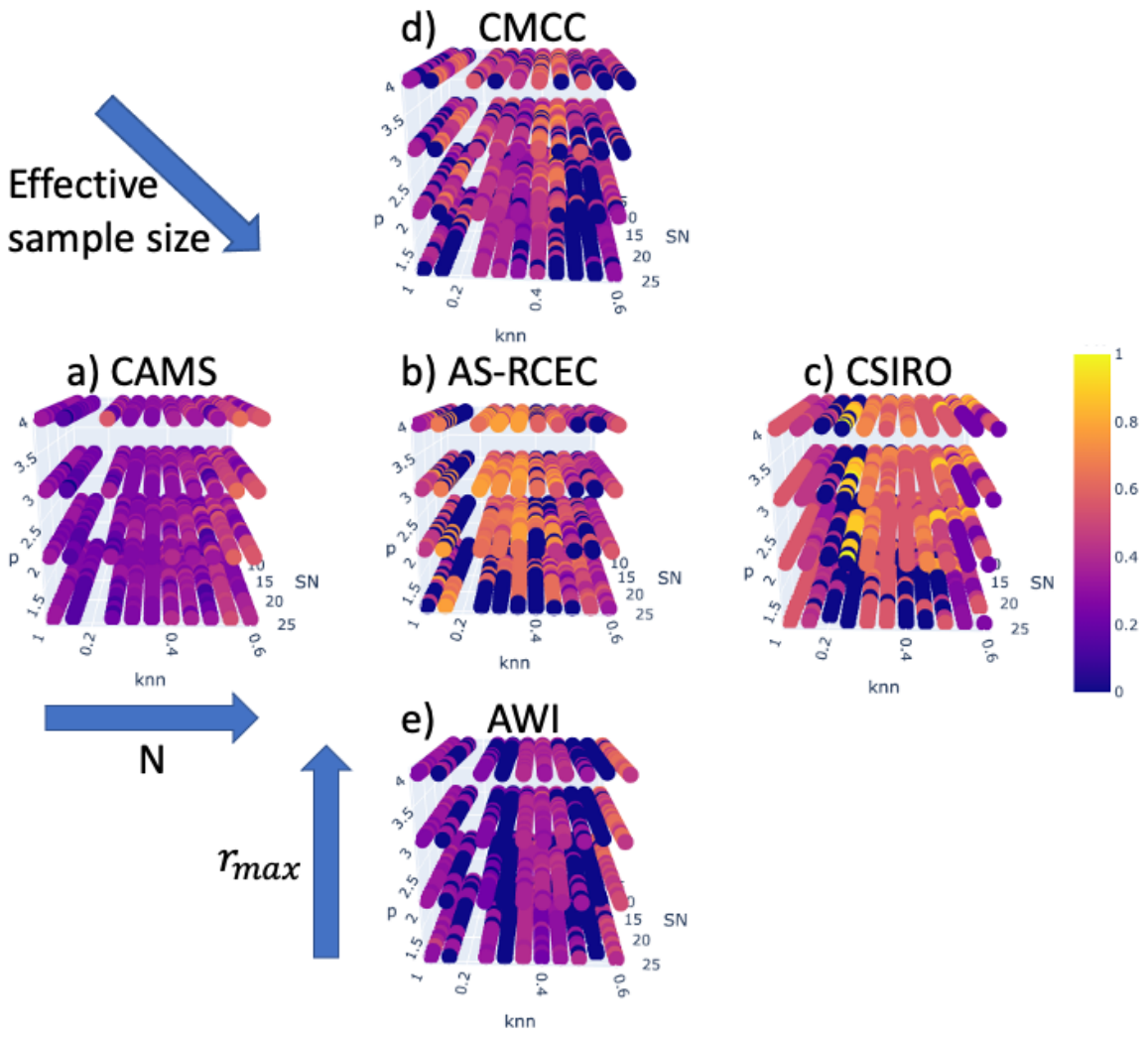

Figure 4.7: Oriented-recall scores (colors) for the "tuning" simulations. The scores are displayed on a 3D grid with axes corresponding to different parameters, including k nearest neighbors for CMIknn (knn, x ordinates), the number of preliminary LPCMCI iterations (p, y ordinates), and shuffle neighbors for CMIknn (SN, z ordinates). Knn = 0.2 was excluded by mistake for all simulations but CSIRO; missing dots where knn ≠ 0.2 indicate that LPCMCI failed to converge because of time constraints. The simulations are organized by increasing time series length (N) from left to right: N=250 for CAMS (a), N=500 for AS-RCEC (b) and panels (d) and (e), and N=1000 for CSIRO (c); and by increasing maximum autocorrelation (r_{max}) from bottom to top: $r_{max} = 0.18$ for AWI (e), r_{max}~0.3 for panels (a-c), and $r_{max} = 0.66$ for CMCC (d).

Across the middle row, we can see that longer time-series length leads to larger oriented-recall given the appropriate choice of knn: CAMS (N=250) is mostly purple, AS-RCEC (N=500)

achieves some orange, and CSIRO (N=1000) reaches yellow. Besides this expected improvement in the optimal performance of LPCMCI with increasing sample size, we see that the best choice for knn—the fraction of the data to use in each data-adaptive 'bin' for estimation (see Section 4.2.9)—depends on time series length. For N=250 (a), the best value for knn appears to be 0.55, much higher than the range recommended for CMIknn in isolation (0.1-0.2). For N=500 (b), the parameter space with good performance is relatively wide, but the best choice of knn appears to be 0.35. Finally, for N=1000, the best choice is 0.25. This lead us to hypothesize that the most effective number of nearest neighbors is indeed related to the sample size linearly by a factor of 0.15—as suggested by Runge (2018b)—but with an offset of 100, where the offset may depend on the dimensionality: $\text{knn} = (0.15N + 100)/N = 0.15 + 100/N$. This is logical, because as the size of the conditioning sets used during causal discovery increase with the number of variables, we may need some minimum number of points to evaluate conditional independence even as the sample size approaches zero.

Looking vertically, AWI and CCMC—the climate models chosen to test varying autocorrelation—do not follow the pattern of improved performance at lower knn for larger effective sample size. Either our measure of autocorrelation is a poor representative of effective sample size (perhaps we should have taken the average instead of the maximum over each variable's autocorrelation), or the relationship between autocorrelation and LPCMCI performance is non-linear, or there is another explanation for why AWI and CMCC perform poorly that is unrelated to the parameters of CMIknn. Perhaps CMCC performs poorly because it is not statistically stationary, with low-frequency variability emerging halfway through the run; and perhaps AWI performed poorly because cross-correlation (in addition to autocorrelation) is low, perhaps reflecting a causally-noisy dataset. They both had such low performance over the

entire tested parameter space that it's not clear that they should be used for optimizing parameter choices. We will discuss other potential reasons for low performance in Section 4.5.4.d. In the meantime, we are not able to effectively investigate the effect of autocorrelation on preferred parameter choices. This is acceptable because the simulations chosen for our analysis (bottom half of Table 4.3) all have a maximum autocorrelation (r_{max}) similar to that of the simulations used for tuning to the time-series length.

We repeated the tuning experiment for AS-RCEC with $\tau_{max} = 2$ (not pictured). Oriented recall scores are more varied with increasing τ_{max}, with more PAGs receiving 0s but a higher maximum score. The maximum occurred at knn=0.4, consistent with the hypothesis that a larger τ_{max} increases the dimensionality of the problem, requiring a larger offset. On average, only one dependency with a lag of two years was found, either reflecting short causal lags in the underlying causal structure or the low detection power of CMIknn. Since the climate models selected for analysis give no indication of having detectable causal lags longer than 1 year, we focus on $\tau_{max} = 1$. As explained in Section 4.2.11, because LPCMCI does not require causal sufficiency, a small τ_{max} may limit the algorithm's ability to remove autocorrelation effects, but it will not violate the assumptions of LPCMCI.

To test our hypothesized relationship between time-series length and preferred choice for knn, we repeat the tuning experiment on the piC simulations chosen for our analysis (Figure 4.8). We confirm that performance varies with knn while SN and the number of preliminary iterations (p) have little effect on performance (though sometimes, like for IPSL and NCAR, $p > 1$ can be beneficial). We summarize the apparent best choice of knn for each model in the first two columns of the top part of Table 4.4. CNRM (panel a) appears to have a high-performing neighborhood of parameter space centered at knn = 0.35 (which has the highest individual

oriented-recall score) and 0.4 (which is the most consistently high-performing region), in line with our prediction for a time series length of 500. MRI (panel b), which has a time series length of 700, also is consistent with our hypothesis, and appears to maximize its performance at knn= $.3 \approx .15 + 100/700$. Unfortunately, performance for MRI and the other climate models is lower than expected for longer time-series lengths, and for some of them it's so low or inconsistent that there doesn't appear to be any high-performing neighborhood of parameter values (these say "all" in the table). Thus, we are neither able to confirm nor reject our hypothesized formula for ideal parameter choices.

For the purposes of this chapter, we consider the ranges of knn noted in Table 4.4 together with $p > 1$ and $SN > 9$ to be the 'ideal' parameter space for each climate model.

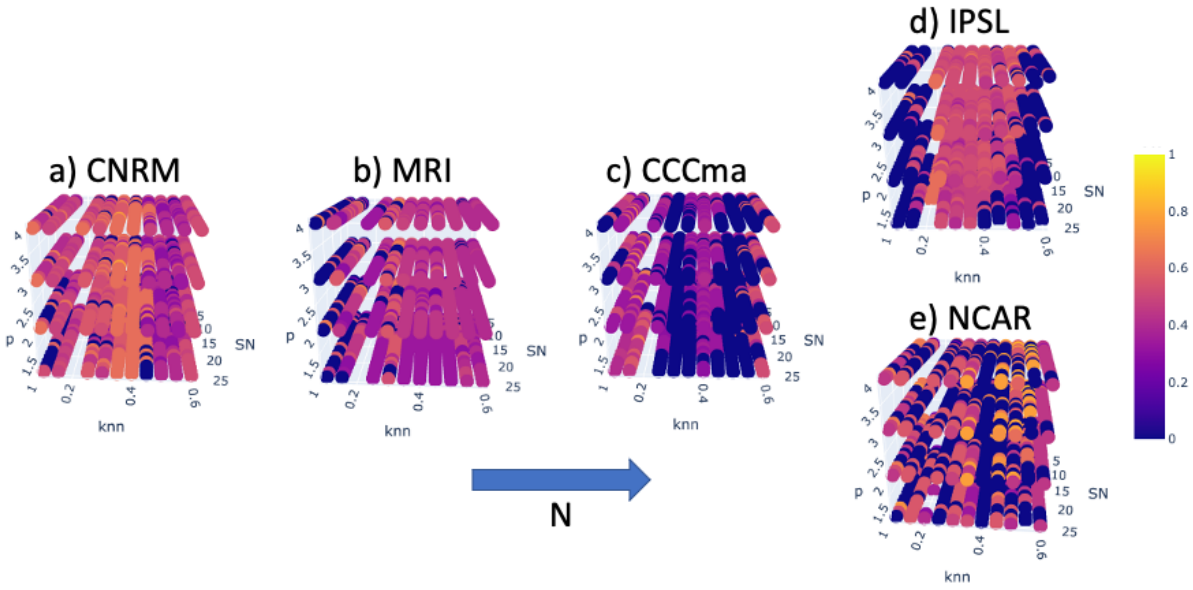

Figure 4.8: As in Figure 4.7, but for piC simulations from the climate models chosen for our analysis. Time series lengths are, from left to right: 500 (a), 700 (b), 1051 (c), and 1200 (d, e).

Table 4.4: The first major column lists the climate models used for analysis with a suffix that denotes whether SA and TT were included ("TT", top portion) or whether IN was included instead ("IN", bottom portion; this is motivated in Section 4.5.4). The second major column shows the knn values that resulted in high oriented-recall scores in Figure 4.7. The ideal parameter space used as probabilistic background knowledge when calculating the likely-accurate scores is defined by these knn values, SN>9, and p>1. The second major column also displays the number of ideal parameter combinations (and the fraction of the ideal parameter space) that was used for calculating the likely-accurate score (the rest had edges pointing backwards in time or ending with an x). When the number of admissible parameter combinations was low, we also used a larger 'ideal' parameter space for comparison, and noted this in a second minor row. The third major column shows the maximum likely-accurate score achieved by PAGs from the entire parameter space (see Section 4.4.5). The fourth major column shows parameter combinations (see Sections 4.4.3 and 4.4.4) that achieved the highest likely-accurate score, and the fifth major column shows the robustness (see Section 4.4.6) and oriented-recall score (see Section 4.4.4) of the associated graph.

Climate Model	Best knn	Admissible		Likely-accurate	knn	SN	p	Robustness	Oriented-Recall
		Num	Frac						
CNRM-TT	.35,.4	90	.94	.62	.4	19-25[14]	1-4	.78	0.58
MRI-TT	.3	0	0	0					
	all	0	0	0					
CCCma-TT	all	0	0	0					
IPSL-TT	.2-.35	7	.05	.63	.30	11	2	.82	.42
	all	10	.02	.58[15]	.35	12	2	.76	.42
NCAR-TT	all	58	.13	.57	.3	9-10[16]	3-4	.68	.58
MRI-IN	.15,.2	3	.03	.74	.2	19	4	.90	.61
	all	57	.11	.44	.5	12	2-4	.65	.36
NCAR-IN	.35,.4	67	.74	.68	.35	21	2	.81	.64
CCCma-IN	.3	0	0						
	.1-.35	26	.10	.42	.15	22	2	.73	.45

4.5.2. Graph Selection

From the causal graphs discovered using every tested parameter combination, we select the one with the highest likely-accurate score for each climate model. The likely-accurate score (Section 4.4.5) is informed by our background knowledge (red circles and arrows and black

[14] The exact combinations of SN and p that produced the highest-scoring graph are SN=19 with p=2,3,4; SN=20 with p=4; SN=21,23,25 with p=1; SN=22 with p=2; and SN=24 with p=1,3

[15] This PAG is the third-highest performing PAG according to the ideal parameter space in the first minor row, with a likely-accurate score of .61 and a robustness score of .8

[16] All combinations of the SN and p values listed here produced the highest-scoring graph except SN=10 with p=3

arrows in Figure 4.4) and graphs from within each climate model's 'ideal' parameter space that don't contain obviously non-physical edges (edges with x's, or "simultaneous" edges within the same calendar year that point backward in time across the seasons). The first minor column of the second major column in Table 4.4 notes the knn values that help define the ideal parameter space, and the other minor columns show the number of parameter combinations from the ideal parameter space for each climate model that produced usable or "admissible" PAGs and the fraction of the ideal parameter space they comprise. When this number or fraction is very small, we re-calculated the likely-accurate scores using a larger 'ideal' parameter space for comparison, and note this in another minor row.

IPSL—which had only seven usable graphs within the ideal parameters space (knn=0.2-0.35)—gained only three additional usable PAGs elsewhere, and this did not improve its maximum likely-accurate score (third major column of Table 4.4). We thus prefer the graph that maximizes the likely-accurate score when optimized over the ideal parameter space, but we examine this alternate choice as well.

MRI and CCCma, on the other hand, did not produce any usable PAGs in anywhere in the entire parameter space. In Section 4.3, we discussed the possibility that SA and TT may be coupled with Sahel precipitation (pr) and that TT may be coupled with SST throughout the tropics. Thus, we hypothesized that the non-physical false orientations might be due to a violation of LPCMCI's assumption of acyclicity relating to SA and TT. Indeed, all of the non-physical graphs for MRI and 80% of the non-physical graphs for CCCma contained backwards edges from the South Atlantic in the summer to the Gulf of Guinea in the previous spring, and almost all of the non-physical graphs for IPSL contained an effect of summer upper tropospheric

temperature over the Sahel (TT) on the Atlantic Meridional Mode from the previous spring (AMM).

We attempted to improve the performance of LPCMCI on this dataset by removing SA and TT. Since we removed two variables, we added the Indian Ocean, which may also have unique effects on Sahel rainfall. Still, we have one fewer variable than used in our tuning experiments (Section 4.5.1), so we cannot implicitly trust the tuning results. We repeated the tuning experiments for all climate simulations chosen for analysis with this new set of variables (not shown). Under these conditions, MRI and CCCma still struggled, but did produce some usable graphs. However, two of the three models that produced usable PAGs when the Indian Ocean was excluded produced backwards edges or edges with x's with every tested parameter combination. In the bottom half of Table 4.4, we list the ideal parameter space and the number of usable PAGs for each climate simulation that produced any usable results with the new set of parameters.

MRI produces only three usable graphs in the ideal parameter space, but the maximum likely-accurate score is much lower when calculated using all values of knn, so we again prefer the graph that maximizes the likely-accurate score when optimized over the smaller ideal parameter space. For CCCma, we do find some usable graphs when we expand the 'ideal' parameter space, but the maximum likely-accurate score is below 0.5, meaning that, in our estimation, a given edge is on average more likely to be incorrect than correct. Expanding the ideal parameter space to the entire tested parameter space only reduces the likely-accurate score in this case because the performance is incredibly low for graphs with large knn. Thus, we do not trust the results for CCCma, and we do not present or discuss them.

4.5.3. Discovered Causal Relationships

The fourth major column of Table 4.4 identifies the parameter combinations that produced the highest-scoring PAG, and these PAGs are displayed in Figure 4.9a-e (all the highest-scoring PAGs for each simulation are the same). The first row (panels a-c) shows results that omit the Indian Ocean, and panels (d) and (e) show results that omit SA and TT; these will be used to shed light on causal relationships including the Indian Ocean.

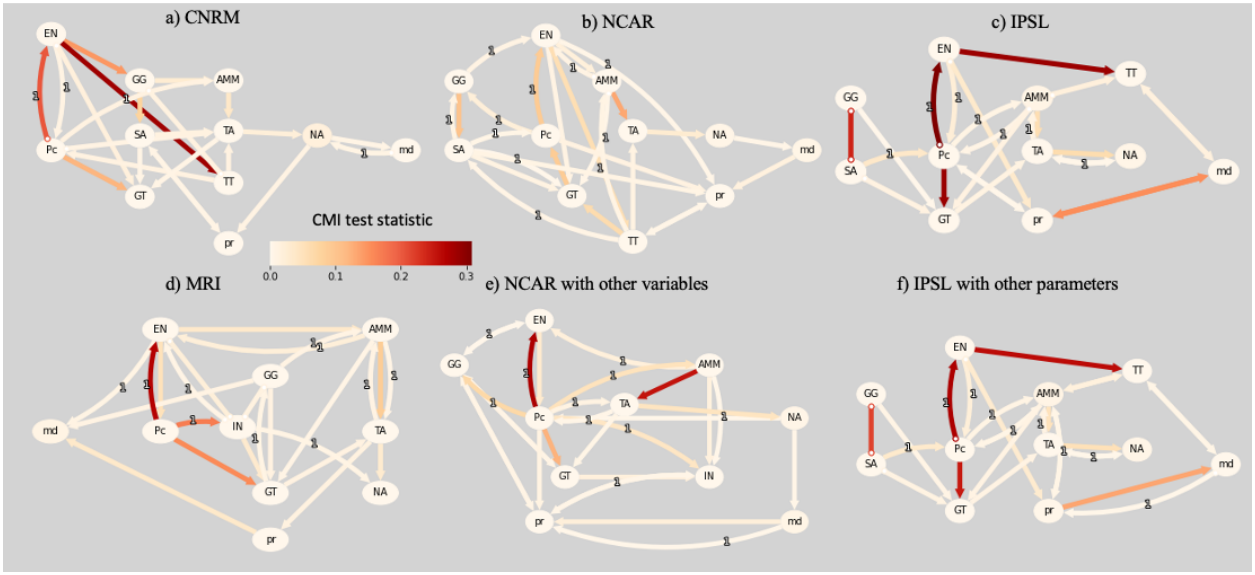

Figure 4.9: Discovered causal PAGs with maximum likely-accurate scores for (a) CNRM, (b) NCAR, and (c) IPSL when SA and TT were included and IN was excluded, and for (d) MRI and (e) NCAR when SA and TT were excluded but IN was included (see the fourth major column of Table 4.4 for the parameter values employed to discover each PAG). Panel (f) shows an alternate choice of PAG for IPSL: it receives the maximum likely-accurate score when optimized over the entire tested parameter space (second minor row for IPSL in Table 4.4) and receives the third-highest score when optimized over the ideal parameter space (second major column, first minor row for IPSL in Table 4.4). The strength of the conditional mutual information test statistic between a variable and its past value is notated in shade of red of the circle behind the name of the variable, and between otherwise adjacent variables is marked in the color of the edge connecting them.

The graphs can be read in a manner similar to our hypothesized causal diagram (Figure 4.4): Mostly we try to group SST indices in rows by season and in columns by location, though this format must be compromised often for clarity of the graph. The colors represent the strength

of the test statistic for dependence between each pair of adjacent variables and for each auto-dependence, which can be roughly interpreted as a normalized detected causal importance of that direct mechanism relative to total variability (though this should not be taken as an approximation of the causal effect). We've used CMIknn as the conditional independence test, which is a distance metric applied to the probability distributions used to define conditional independence (see Sections 4.2.2 and 4.2.9), not an estimate of the functional parameters relating two variables; thus, the values are strictly positive, and are presented from white to red. Recall that in partial ancestral graphs (see Section 4.2.7) a directed adjacency (→ or ←) asserts ancestorship rather than parenthood, and edges may contain circles that permit multiple orientations within the discovered equivalence class. Straight arrows are dependencies within the same calendar year, while curved arrows are lagged dependencies labelled with the length of the time lag, which in this case is always one calendar year. Auto-dependence of a single variable is represented with the color of the circle behind the variable name, but note that auto-dependence is often so weak that it can be difficult to identify the presence or absence of true auto-dependence in the summary graphs. Furthermore, it is never possible to see the orientation of auto-dependencies—which could be causal or confounded—in the summary diagrams. Refer to the text for an account of the active auto-dependencies and their orientations.

In Sections 4.5.3.a and 4.5.3.b, we analyze the discovered adjacencies between ocean basins and between ocean basins and the Sahel, respectively, and compare them to the green and blue arrows from our causal hypothesis (Figure 4.4), which are of greatest scientific interest. We discuss the recovery of ground-truth red and black (circles and) arrows when we analyze the performance of LPCMCI in Section 4.5.4.b.

4.5.3.a SST

We begin by examining CNRM (Figure 4.9a). The Pacific (EN and Pc) is not influenced by any of the other ocean basins, contrary to our expectations. Furthermore, we had expected the effects of the Pacific on the North Atlantic (NA) and the Gulf of Guinea (GG and SA) to be mediated by the North Tropical Atlantic (AMM and TA) via atmospheric pathways. Instead, we find a direct effect of peak El Niño conditions (EN) on the Atlantic Niño (GG) that mediates part of the effect of ENSO on the North Tropical Atlantic. Both of the Tropical Atlantic indices (GG and SA) both affect the North Tropical Atlantic indices in their respective seasons (AMM and TA), likely representing an oceanic pathway via northward surface water fluxes due to the Atlantic Meridional Overturning Circulation, rather than the expected atmospheric teleconnection from the Atlantic Meridional Mode to the Atlantic Niño. (As pointed out in Section 4.3, if both these mechanisms are active simultaneously, this would cause a simultaneous cycle and violate the assumptions of LPCMCI.)

The North Tropical Atlantic is also directly affected by ENSO—but not quite in the way we expected. AMM supposedly responds to the preceding development of El Niño (Pc) rather than the peak (EN). This is not inconsistent statistically with our hypothesis that relates peak El Niño conditions in winter to AMM later that same calendar year, but it is causally distinct—suggesting that intervening on the peak El Niño temperatures in winter would not prevent the response of AMM to El Niño's development the previous summer. The 'simultaneous' effect of the peak of ENSO (EN) on TA is supposedly mediated by upper tropospheric temperature over the Sahel (TT). TT might be standing in for upper tropospheric temperature over the tropical Atlantic, which must be quite similar to TT by the weak temperature gradient constraint. There is also potentially an effect of the development of ENSO (Pc) on TA that same summer, which

would be much faster than our expected effect, though this edge isn't fully oriented, so it may also be confounding.

Rather than unmeasured confounding between the Atlantic Meridional Mode (AMM) and the North Atlantic (NA), TA is seen to cause NA, which is also likely an oceanic connection via surface water fluxes. The North Atlantic does not affect the other oscillatory modes as we had predicted, and instead interacts with the Mediterranean (md) in two ways we discussed in Section 4.3: NA affects md that same summer, as concluded by Mariotti and Dell'Aquila (2012) and Skliris et al. (2012), while md affects NA one year later, perhaps via outflow of its water into the Atlantic basin.

Finally, in addition to the basins that fall within the bounds of the Global Tropics (GT; see Figure 4.3), there is a direct influence of EN on GT, as we had predicted if the other areas in GT also respond to ENSO.

In NCAR (panel b), El Niño is not as prominent a driver of global SST variability, as evidenced by the lighter colors of the arrows. Instead of this strong relationship, many additional weaker causal relationships are discovered. Peak El Niño (EN) responds to the Atlantic Niño (GG) from the year before, similar statistically to our hypothesis; but it also possibly responds to the Atlantic Meridional Mode (AMM) from the year before directly instead of via GG, similar statistically to the argument of Ham et al. (2013), which we did not depict in Figure 4.4 (this edge is not fully-oriented, so it may be confounding). For this simulation, the development of ENSO (Pc) affects the following year's Atlantic Niño (GG), similar statistically to CNRM, but with a different causal lag. Unlike CNRM, the Atlantic Meridional Mode (AMM) responds to peak El Niño temperature (EN) directly, as we had originally predicted; and, like CNRM, the response of summer temperatures (TA) to EN is mediated by AMM and summer tropospheric

temperature over the Sahel (TT). Again as in CNRM, summer Tropical Atlantic temperatures (TA) are seen to affect the North Atlantic (NA) which in turn affects the Mediterranean sea (md) and no other ocean basin. Unlike our hypothesis and the other simulations, we don't see any confounding between NA and other ocean basins. LPCMCI finds that GT may be affected by the Atlantic Meridional Mode from the previous year (this edge is not fully oriented), and is also affected by summer SST in the South Atlantic (SA) from the year before. Perhaps these basins affect SST elsewhere in the Global Tropics via an atmospheric bridges pathway that operates slower than predicted.

In IPSL (panel c), there is no direct effect of ENSO (EN and Pc) or the Atlantic Meridional Mode (AMM and TA) on the Atlantic Niño (GG and SA). The Atlantic Meridional Mode (AMM) responds to the development of El Niño in summer the year before, as in CNRM. The development of El Niño (Pc) is affected by other ocean basins, but not in the way we had expected. LPCMCI found potential effects from the previous year's Atlantic Meridional Mode (AMM; this edge is not fully oriented), as argued by Ham et al. (2013); and from the previous year's SA, which is not only causally distinct from our hypothesis, but may even result in a statistical association one year different from that seen in observations.

The North Atlantic (NA) has no effect on other basins, contrary to expectations and other simulations, and is again affected by TA. Summer SST in the North Atlantic is also confounded with SST in the North Tropical Atlantic, but during the following summer ($\mathrm{TA}(t+1)$) rather than the previous spring ($\mathrm{AMM}(t)$); the source of this confounding is unclear. The Mediterranean (md) is only confounded with atmospheric variables, and is not connected to other ocean basins at all, as in our hypothesis. The Global Tropics are found to respond additionally to spring Atlantic Niño (GG) and Atlantic Meridional Mode (AMM) variability, which may represent

atmospheric teleconnections to other parts of the Global Tropics similar to the response to El Niño.

The first 2 PAGs in the next row of Figure 4.9 allow us to learn about the Indian Ocean. For MRI (panel d), almost all SST relationships excluding the Indian Ocean (IN) are ones we've seen in the other models, aside from a dependence of the Mediterranean Sea on the Atlantic Niño (GG) and on the previous year's peak El Niño (EN), and the lagged response of AMM to GG. The Indian Ocean (IN) responds to the previous year's developing El Niño (Pc), which may be an atmospheric response or an oceanic communication via the Indonesian Throughflow. It also affects the following year's NA (perhaps via an effect on the North Atlantic Oscillation) and potentially the following winter's peak El Niño (EN), as predicted. IN is related in an unspecified way to the preceding peak El Niño (EN) in addition to the preceding development of El Niño (Pc), and the Atlantic Niño (GG), the last of which could be an oceanic connection via the Agulhas current. When we include IN in the NCAR analysis (panel e), we find again that IN responds to the previous year's developing El Niño (Pc), but it responds to the Atlantic Meridional Mode rather than the Atlantic Niño, and it also oddly responds to the Global Tropics rather than causing changes in GT. In NCAR, IN appears to have no effects on other ocean basins.

These results make it clear that there may be teleconnections that are important in some climate models despite not appearing statistically in other simulations or in observations, like the effect of peak El Niño on the Atlantic Niño in CNRM. Additionally, it would appear that the causally-relevant lags—and even the lags of statistical associations—may differ significantly between observations and climate simulations and between different climate simulations. Not only do the climate models demonstrate differing connectivity and differing relative importance

of different causal effects, but they also have causal effects that are inconsistent with one another. For instance: in NCAR (b), the Tropical Atlantic (TA) affects the Global Tropics (GT) in the way we had predicted, but in IPSL (c), the Global Tropics affects the Tropical Atlantic. Both cannot be possible without producing a simultaneous cycle, which is a violation of our assumptions. This suggests that simulations may need to be analyzed separately.

Nevertheless, the causal discovery process also identified some causal relationships that appear in multiple climate simulations, and thus can inform our expectations for observed relationships. LPCMCI discovered an effect of TA on NA in every single climate simulation, suggesting that in order to measure truly-causal effects of basin-wide SST—whether in simulations or observations—it is robustly important to account for meridional advection of (near-)surface ocean water in the Atlantic basin from the North Tropical Atlantic to the North Atlantic, and perhaps also from the Gulf of Guinea to the North Tropical Atlantic – not just atmospheric teleconnections between the modes of internal variability. This means we must be careful during causal effect estimation and causal discovery that expected atmospheric teleconnections from northern to southern basins (Martín-Rey et al. 2018) don't cause simultaneous cycles in our data. Additionally, an effect of the North Atlantic on the Mediterranean Sea appeared in half of our climate simulations, potentially consistent with the conclusions of Mariotti and Dell'Aquila (2012), Skliris et al. (2012), or both. The results also suggest that it may be important to account for the effects of the Atlantic Meridional Mode and the Atlantic Niño, in addition to El Niño, on SST elsewhere Global Tropics via atmospheric bridges. To that end, three of the four climate models support the hypothesis of Ham et al. (2013) that AMM can trigger El Niño, though the exact causal lag is unclear. Finally, causal lags interpreted from observed statistical relationships should be questioned; for instance, the Indian

Ocean responds to El Niño's development during the previous year, rather than its peak, in both of the climate simulations that produced usable results using IN.

4.5.3.b Drivers of Sahel Precipitation and Tropospheric Temperature

Precipitation in the Sahel (pr) is driven by different ocean basins in each simulation (see first column of Table 4.5). In CNRM (a), the Sahel responds only to the North Atlantic; in NCAR (b), Sahel rainfall responds to the developing phase of El Niño, to South Atlantic SST, and to the Mediterranean; in IPSL (c), the Sahel responds only to peak El Niño conditions from the previous winter (which is consistent with analysis of coupled simulations even though it differs from observations; see Joly and Voldoire 2009); and in MRI (d), Sahel precipitation is affected only by summer SST in the North Tropical Atlantic (TA). All the dependencies are consistent with the literature except the dependence of Sahel precipitation on peak El Niño temperatures from two winters ago in NCAR, though it's possible this represents a slow teleconnection in this climate model. The fact that basins other than GT and NA drive Sahel precipitation in many of these simulations would suggest that GT and NA—though they correlate well with Sahel precipitation in historical simulations and observations—are not a sufficient summary of the causal drivers of Sahel rainfall, and so the observed statistical relationship to NARI may not hold in a changing climate. It may be that the Global Tropics only appeared to be a good indicator for TT and Sahel rainfall in historical simulations because SST across the global tropics was confounded by anthropogenic emissions and by atmospheric bridges to individual important ocean basins. (TT also responds to ocean basins other than GT – in all simulations that included TT, it responds to peak El Niño temperatures (EN), and it additionally responds to the Atlantic Meridional Mode (AMM) in IPSL, and potentially to the Atlantic Niño in CESM.)

Table 4.5: Discovered (from Figure 4.9) and expected (from Figure 4.4) parents (see Section 4.2.2) and spouses (see Section 4.2.5) of Sahel precipitation. Variables are italicized if they were found to be independent of Sahel precipitation in Section 4.5.5.

Climate Model	Parents of pr	Spouses of pr
CNRM	NA	SA
NCAR-TT	md, Pc, SA, *EN-1*	TT
IPSL	*EN*	md, *Pc*
MRI	TA	∅
Hypothesis	NA, TT	∅

Not only did NA and GT fail to mediate all causal effects between Sahel precipitation and other ocean basins, but LPCMCI did not even discover a causal dependence of Sahel precipitation (pr) and tropospheric temperature over the Sahel (TT) on the Global Tropics (GT) in any of the climate simulations, contrary to the claims of G13. TT was also not identified as a driver of Sahel precipitation variability in any of the simulations. In most of the climate models, TT can be separated from pr by conditioning on EN. Perhaps the altitude chosen for TT is not the one that is causally-relevant for Sahel rainfall in these simulations, making variations in TT another effect of ENSO that doesn't actually mediate the effect on Sahel rainfall.

The orientation is different from our expectations, but, in NCAR, Sahel precipitation is directly connected to tropospheric temperature, which in turn mediates a causal relationship with GT. If the orientations of these edges are incorrect, NCAR may experience the expected relationship between GT, TT, and pr, in addition to other direct causal relationships with EN, Pc, and SA. As discussed in Section 4.3, it is possible that convection associated with Sahel precipitation affects TT, and the confounding relationship is a misrepresentation of a disallowed cyclic dependency. Furthermore, causality may flow both ways between global tropical upper tropospheric temperature and SST in any given ocean basin, depending on which ocean basins are warm and convecting, and TT over the Sahel may be statistically indistinguishable from global tropical TT, or may be an actual mediator via gravity waves propagating both directions

around the tropics (see Section 1.4). Our discovered PAGs appear to support this possibility: in two of the three PAGs that include Sahel TT, it affects the temperature of some tropical ocean basin or other: both CESM and NCAR show an effect of TT on summertime temperatures in the North Tropical Atlantic (TA); in CNRM, TT additionally affects the development of El Niño (Pc) rather than simply responding to it; and in NCAR, tropospheric temperature over the Sahel is found to affect SA the following year and also GT, rather than responding to it. These potentially-cyclic relationships may complicate the interpretation of the presence and orientation or absence of causal adjacencies to TT.

Sahel precipitation is unfortunately also found to be confounded with other climate indices in almost every simulation (see second column of Table 4.5): in CNRM (a), it is confounded with summer SST in the Gulf of Guinea (SA); in NCAR (b), it's confounded with upper tropospheric temperature over the Sahel (TT); and in IPSL (c), it is confounded with the developing phase of El Niño (Pc) and with the Mediterranean (md). The apparent confounded relationship between pr and SA in CNRM (a) could be a misrepresentation of a coupled relationship between of the formation of the South Atlantic cold tongue to the monsoon circulation, in which case we cannot interpret the relationship as causal, and it may be impossible to appropriately estimate the causal effect of SA on Sahel precipitation using any method at this timescale. If these confounding relationships are physical, they would directly prevent us from measuring the simulated teleconnection between the implicated basin and Sahel rainfall in that climate model in a causally-responsible way, unless another variable can be found to de-confound them. Since these confounded relationships are not discovered in all climate simulations, causal effect estimation for each teleconnection might be possible in a subset of the climate models, suggesting that simulations are better analyzed separately.

4.5.4. LPCMCI Performance

The likely-accurate scores (third major column, top part of Table 4.4) are all above 0.5, but are all below 0.65, limiting the trust we can place in our results. This score is reduced when the results are not robust to small parameter perturbations, but is also reduced when LPCMCI robustly fails to recall known adjacencies or robustly returns edge orientations that are inconsistent with our background seasonal knowledge (see Section 4.4.5). In the following sections, we examine the performance of LPCMCI according to robustness (Section 4.5.4.a), (oriented-)recall (Section 4.5.4.b), and orientation (Section 4.5.4.c). In Section 4.5.4.d, we investigate the reasons for the non-physical orientations seen in Section 0 and discussed in Section 4.5.4.c.

4.5.4.a Robustness

The robustness score (Section 4.4.6) quantifies how well our chosen graph represents the discovered graphs from the ideal parameter space for each climate model that don't have backwards edges or x's, and thus it quantifies the sensitivity of our results to small parameter perturbations. The robustness scores (second-to-last column of Table 4.4) are all between 0.7 and 0.85, suggesting that the chosen graph is a good representation of the ensemble of results, but that there are still some substantial differences between them.

In Figure 4.9f, we picture the third-highest performing PAG for IPSL; it would have been the chosen PAG if we optimized the likely-accurate score over the entire parameter space (see Table 4.4), and it achieves a likely-accurate score of 0.61 and a robustness score at 0.80 according to the original ideal parameter space. Most of the skeleton is the same as the highest-scoring PAG for IPSL (panel c), but it has a slightly different connectivity: it does not find confounding between Pc and pr, and instead has some additional adjacencies, including

confounding between AMM and pr the following year, and confounding or a causal relationship between md and pr the following year. It also has some different orientations that are inconsistent with our chosen graph in panel c). For one, LPCMCI detects a collider at AMM in the triple $\mathrm{TT}(t) \leftrightarrow \mathrm{AMM}(t) \leftrightarrow \mathrm{Pc}(t+1)$,[17] which is an unoriented chain in the highest-performing PAG. It must be that the change in parameter choices changed the result of the conditional independence test there, and, unfortunately, our robustness score is not able to strongly detect this type of inconsistency (see Section 4.4.6). But this PAG also has an inconsistent orientation that can likely be traced back to changes in the skeleton that caused the algorithm to perform a different conditional independence test entirely to orient the edge: the unlikely conclusion that Sahel precipitation affects simultaneous SST in the Mediterranean in panel (f) may be a result of replacing the adjacency between pr and Pc in panel (c) with adjacencies to AMM and md from the previous year in panel (f).

We can learn about the robustness of our results to perturbations in the data by comparing the discovered causal diagram for NCAR when SA and TT are included (b) to that when IN is included instead (e). When a variable is removed, all its adjacencies must be removed and all non-collider paths that passed through that variable must be replaced with direct adjacencies, but the rest of the underlying causal structure should theoretically remain unchanged. For SA in NCAR, we would expect the path $\mathrm{GG} \rightarrow \mathrm{SA} \rightarrow \mathrm{GT}$ from panel (b) to be replaced with the adjacency $\mathrm{GG} \rightarrow \mathrm{GT}$ in panel (e), the path $\mathrm{GG} \rightarrow \mathrm{SA} \rightarrow \mathrm{pr}$ to be replaced with the adjacency $\mathrm{GG} \rightarrow \mathrm{pr}$, and the path $\mathrm{GG}(t) \rightarrow \mathrm{SA}(t) \rightarrow \mathrm{Pc}(t+1)$ to be replaced with the adjacency $\mathrm{GG}(t) \rightarrow \mathrm{Pc}(t+1)$. In panel (e), we can see two of the three expected adjacencies, connecting GG to GT

[17] This corresponds to the lower of the two curved edges connecting Pc and AMM. Unfortunately, the lag notation in the summary graphs is not currently fully specified when the relationship is confounded.

and to pr; but the orientations are both confounded rather than causal, which is inconsistent with our first diagram.

We can also see an instability in the exact timing of causal connections, even when the causal connections seem to be unrelated to SA, TT, and IN: in panel (b), there is a simultaneous edge from winter Pacific SST (EN) to AMM, and in panel (e) there is a lagged edge from the previous summer (Pc) instead. Furthermore, there are unexpected changes in the causal connectivity of the graph even when lag and orientation are ignored. We see a lagged coupled dependence of Pc on TA in panel (e) that was not present in panel (b) as well as a direct effect of NA on the development of El Niño, as we originally hypothesized. In panel (b), the direct effect of winter Pacific SST (EN) on summer Pacific SST (Pc) is oddly mediated by the Global Tropics (GT): $\mathrm{EN} \rightarrow \mathrm{GT} \rightarrow \mathrm{Pc}$, while in panel (f) we see the expected relationship $\mathrm{EN} \rightarrow \mathrm{Pc} \rightarrow \mathrm{GT}$. These changes are difficult to explain even statistically, because the connectivity of the Pacific to itself at different seasons and the effect of NA on the development of El Niño supposedly have nothing to do with SA, TT, or IN in the final discovered causal diagrams, suggesting room for improvement in the algorithm itself.

4.5.4.b Recall

The low likely-accurate scores are strongly affected by low oriented-recall (last major column of Table 4.4) of expected auto-dependencies and ontological edges (red and black, respectively, in Figure 4.4). This score is affected by the skeleton and the orientation of the discovered PAGs; to understand the low oriented-recall in more depth, we examine the expected auto- and ontological dependencies in the PAGs in the first row of Figure 4.9. The details of the auto-dependencies that are not unrolled over multiple seasons are difficult or impossible to see in the summary graphs, but will be identified in the text.

In CNRM (panel a), LPCMCI successfully recovers the ontological response of GT to its constituent basins (Pc, SA, and TA) and no auto-dependence of GT on itself. However, the other auto-dependencies are not successfully recovered. The effects of summer temperatures in the North Tropical Atlantic (TA) and Gulf of Guinea (SA) on the following spring (AMM and GG, respectively) are not captured, and while there is a 'simultaneous' effect of winter Pacific SST (EN) on the immediately-following summer (Pc), it is indirect, mediated by tropospheric temperature over the Sahel (TT); this cannot represent the expected auto-dependence due to the high heat capacity of the ocean. Of the basins that only appear at one season, only the North Atlantic (NA) has a direct link with itself in the previous year, and LPCMCI failed to fully orient the edge: $\mathrm{NA}(t) \circ\!\!\rightarrow \mathrm{NA}(t+1)$.

In the other two climate models, LPCMCI also struggles to reproduce the expected auto-dependencies. LPCMCI is able to detect the unrolled auto-dependence of the Gulf of Guinea $\mathrm{GG}(t) \rightarrow \mathrm{SA}(t) \rightarrow \mathrm{GG}(t+1)$ in NCAR (panel b) and the unrolled auto-dependence of the North Tropical Atlantic $\mathrm{AMM}(t) \rightarrow \mathrm{TA}(t) \rightarrow \mathrm{AMM}(\mathrm{t}+1)$ in IPSL (panel c), but doesn't find both in the same climate model. Auto-links for NA and md are found in NCAR (though neither is fully-oriented), but in IPSL, LPCMCI finds that NA is confounded with—rather than caused by—its previous value, and fails to find any auto-dependence for md (instead finding an unexpected auto-dependence of springtime AMM on itself in the previous year that is not mediated by TA in addition to the expected unrolled auto-dependence). LPCMCI isn't able to detect a direct dependence of Pc on EN in the same calendar year for either climate model (though NCAR does exhibit an indirect simultaneous relationship that is mediated by GT, or by TT and then GT).

LPCMCI also has trouble recovering the ontological relationships between GT and its constituent basins for these simulations, mostly due to orientations that are inconsistent with our

expectations rather than missing adjacencies in the discovered skeleton. With the exception of TA in NCAR, GT is connected to all its constituent basins in all simulations; but in NCAR, GT causes Pc rather than responding to it; and in IPSL, GT causes TA rather than responding to it. We cannot tell from this analysis whether these orientations are nonphysical and only reflect (strong) faithfulness violations during orientation, or whether the expected causal effect is simply much weaker than a competing causal or confounded relationship. Thus, while these dependencies seem highly unlikely, we do not consider these orientations obviously non-physical. We examine the likelihood of definitively incorrect orientations in the next section.

4.5.4.c Orientation

Recall that, for a variety of reasons, we hid the seasonal time ordering of the variables from LPCMCI (see Section 4.4), making it possible for the algorithm to return an oriented edge that is obviously non-physical because it points backwards in time within a calendar year. This decision certainly reduces the performance of LPCMCI on some simulations in this dataset, but an advantage of this decision is that it allows us to bypass mechanistic uncertainty and objectively evaluate the limits of the performance of LPCMCI – limits we expect to hold even in the presence of complete time information. The accuracy of the orientation of the other edges in the graphs we identify as being the most likely to represent the underlying causal structure (Figure 4.9) cannot be directly tested given our partial knowledge. To the extent that the prevalence of general false orientations is comparable to the prevalence of detectable backwards orientations in the ideal parameter spaces for our climate models, we can estimate the likelihood that a graph has at least one backwards orientation by detecting backwards orientations is graphs that achieve high likely-accurate scores when our seasonal background knowledge is not taken into account (see Section 4.4.6). The probably-accurate score ignores seasonal knowledge

completely, while the "full" likely-accurate score simply does not set the likely-accurate score 0 when a graph contains a backwards edge. The latter would better represent the likelihood of having general incorrect orientations than the former if general incorrect orientations are associated strongly with detectable time-backwards edges.

There is ample evidence that LPCMCI sometimes robustly discovers non-physical edges for some climate models. Recall that MRI and CCCma didn't produce a single usable PAG, and IPSL produced very few (second major column of Table 4.4, see Section 0). We will investigate the reasons for these non-physical edges in Section 4.5.4.d. Here, we approximately quantify our trust in the orientations of discovered simultaneous edges within the same season.

In Figure 4.10, we compare the probably-accurate scores (top row) and the "full" likely-accurate scores (bottom row) to the true likely-accurate scores (y-axis) for the three models that produced physical graphs when including SA and TT: CNRM (left column), NCAR (middle column), and IPSL (right column). All the probability density plots are a superposition of density centered around a diagonal line $y \sim x$ and around a horizontal line $y = 0$. When graphs with high probably-accurate and "full" likely-accurate scores do not contain backwards edges, they will also achieve a similar likely-accurate score and contribute to increased density along the diagonal; when they contain backwards edges, they will contribute density to the horizontal line at likely-accurate $= 0$. The likelihood of an incorrect orientation in a graph with a given likely-accurate score is estimated to be between the marginal likelihoods of a true likely-accurate score of 0 (y-axis) in the vertical cross sections at comparable probably-accurate and "full" likely-accurate scores (x-axis). Comparable scores can be estimated by tracing the desired likely-accurate score from the y-axis to the peak density along the diagonal (grey dashed lines), and the vertical cross section (black solid lines) will also intersect this point.

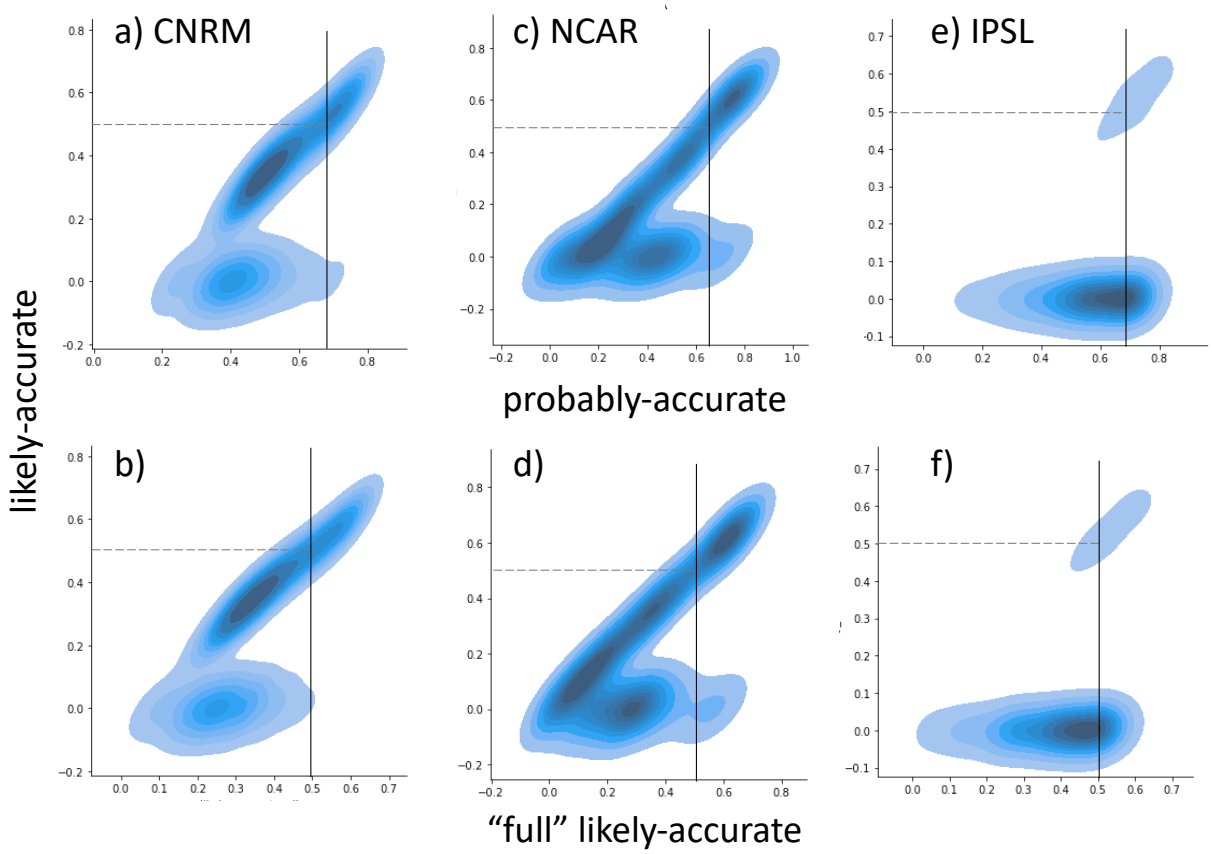

Figure 4.10: Comparison of probably-accurate scores (x-axes on the top row), which do not include background knowledge, and of "full" likely-accurate scores that are not set to 0 when they contain backwards edges (x-axes on the bottom row) to true likely-accurate scores (y-axes) for CNRM (a-b), NCAR (c-d), and IPSL (e-f) when TT and SA are included. Dashed grey lines identify a likely-accurate score of 0.5, and solid black lines highlight the vertical cross-sections at comparable probably-accurate and "full" likely-accurate scores. These cross-sections give the estimated relative likelihood of discovering an edge with a reversed simultaneous orientation.

In all cases, including our background knowledge when calculating edge likelihoods (bottom row) makes little difference to the qualitative comparison. For CNRM (the climate model with the highest performance according to oriented-recall; see Figure 4.7), high-scoring graphs (comparable to likely-accurate scores above 0.5) are unlikely to have incorrect orientations. For the lower-performing climate models, the chances of an incorrect orientation increases. Results for NCAR are still reasonably trustworthy, showing a much higher density

along the diagonal than the horizontal for high-scoring graphs, while results for IPSL are highly suspect. The results in Figure 4.9 are organized from left to right by decreasing trustworthiness.

4.5.4.d Causes for non-Physical Edges

Because multiple climate simulations produced non-physical graphs over the entire parameter space, we cannot attribute these persistent false orientations to the climate model or poor choices of CMIknn parameters. In Section 0, we removed SA and TA – variables we believed might violate LPCMCI's assumption of acyclicity – and replaced them with IN, but LPCMCI still struggled. Could IN also be causing problems for LPCMCI?

In CCCma, one of the most prominent backwards orientations under these conditions is from the Indian Ocean in summer (IN) to the Atlantic Meridional Mode in the previous spring (AMM), but equally prominent are backwards edges from the Global Tropics in summer to AMM the previous spring, and from the Pacific and Global Tropics in summer (Pc and GT) to the Gulf of Guinea the previous spring (GG). In the other climate models, the struggles seem to have nothing to do with the Indian Ocean: in CNRM, the most common backwards edge is from the Pacific Ocean in summer (Pc) to the previous winter (EN); in IPSL, all graphs with backwards edges include this backwards edge and also backwards edges from the Global Tropics in summer (GT) to AMM the previous spring and the Pacific the previous winter (EN); and MRI struggled by producing backwards edges to (GG) from the Global Tropics (GT) in lieu of the South Atlantic (SA). This suggests that all backwards orientations—whether in the graphs including SA and TT or those including IN—might be the result of missing adjacencies and incorrect orientations elsewhere, rather than acyclicity violations due to including potentially-problematic indices.

GT itself could theoretically pose problems for causal discovery because it is ontologically, rather than causally, related to its constituent basins, and causal inference requires that every variable has unique causally-relevant noise (see Section 4.2.2). However, we intentionally left the Indonesian ocean surface out of our individual SST indices (see Section 4.3), and we never included SA and IN at the same time, so GT should have variability that cannot be completely determined by the included constituent basins.

It is also possible that simply having multiple indices that correlate highly with each other could cause strong faithfulness violations that pose a problem for causal discovery (see Section 4.2.8), especially with a conditional independence test with low detection power like CMIknn. In our hypothesized diagram (Figure 4.4), the auto-dependence of the Gulf of Guinea region is unrolled: $\mathrm{SA}(t-1) \to \mathrm{GG}(t) \to \mathrm{SA}(t-1)$. We would have hoped that LPCMCI would separate $\mathrm{SA}(t-1)$ from $\mathrm{SA}(t)$ by finding that $\mathrm{SA}(t) \perp\!\!\!\perp \mathrm{SA}(t-1)|\mathrm{GG}(t)$. But if $\mathrm{GG}(t) \approx SA(t)$, as can be expected since they are almost the same basin (see Figure 4.3) less than one year apart, then LPCMCI may incorrectly find instead that $\mathrm{GG}(t) \perp\!\!\!\perp \mathrm{SA}(t-1)|\mathrm{SA}(t)$, potentially leading to the unoriented adjacencies $\mathrm{SA}(t-1) \circ\!\!\to \mathrm{SA}(t) \circ\!\!-\!\!\circ \mathrm{GG}(t)$, where the partial orientation of the first edge comes from the algorithm's knowledge of the time order. This incorrect skeleton forces the reversed orientation: since $\mathrm{SA}(t)$ is in the separating set, LPCMCI would find that $\mathrm{SA}(t)$ is not a collider, and must orient the second edge away from SA, giving $\mathrm{SA}(t-1) \circ\!\!\to \mathrm{SA}(t) \to \mathrm{GG}(t)$.

For CCCma, all graphs contain backwards edges, but only 80% contain backwards edges from SA to GG. We find that none of the graphs with backwards edges from SA to GG retain the expected adjacency between $\mathrm{SA}(t-1)$ and $\mathrm{GG}(t)$, while 82% of them discover the unlikely auto-dependence of SA on its value from the previous calendar year that forces the reversed orientation. In MRI as well, all of the graphs with backwards edges miss the expected adjacency

and contain the unlikely auto-dependence of SA on itself. Eleven graphs for MRI contained x's elsewhere rather than a backwards edge from SA to GG; more than half still found the false auto-dependence of SA on itself, but all of these also found the true dependence of $\mathrm{GG(t)}$ on $\mathrm{SA}(t-1)$, preventing the application of the orientation rule. All this seemingly supports our hypothesis for the source of reversed orientations. Removing SA might not have been enough to produce usable graphs for some climate models because other SST indices—though we have no reason to suspect they are involved in cyclic dependencies—also correlate highly with each other, either because their auto-dependence is unrolled over time or because they contribute to GT. Indeed, all the graphs for MRI and CCCma when SA and TT were included also contain backwards edges from Pc or GT to EN, respectively, and these three indices are also highly correlated. All this might suggest that it is unwise to include unrolled dependencies in causal discovery, because this will necessarily produce a high cross-correlation while depriving the algorithm of time information that could help it avoid unnecessary conditional independence tests that lead to incorrect conclusions.

But the story is slightly more complicated. Recall that, in Section 4.5.4.a, we found that adding and removing variables sometimes caused changes to the skeleton even in unconnected regions of the discovered graph. Examining the verbose output from LPCMCI for MRI, we see that $\mathrm{GG}(t) \perp\!\!\!\perp \mathrm{SA}(t-1)|\mathrm{SA}(t)$ is sometimes used to justify separating GG from the previous year's SA, but often the algorithm actually justifies the separation by affirming the independence relation $\mathrm{GG}(t) \perp\!\!\!\perp \mathrm{SA}(t-1)|\mathrm{GT}(t-1)$. Though GG and SA are essentially the same ocean basin, GT is theoretically distinct. We had included it in order to test the hypothesis that $\mathrm{GT}(t)$ mediates the effect of $\mathrm{SA}(t)$ and other summer tropical SST on $\mathrm{pr}(t)$, but we do not believe that

$\mathrm{GT}(-1)$ is a driver of $\mathrm{GG}(t)$. Rather, we believe it is a child of the true parent of our target variable $\mathrm{GG}(t)$: $\mathrm{SA}(t-1) \rightarrow \mathrm{GT}(t-1)$ (see Figure 4.4).

Given an 'oracle' conditional independence test (see Section 4.2.12), conditioning on a child of the cause variable that is not a mediator of its effect on our target variable will change the shape of the underlying probability distributions, but will not affect the result of the independence test or the outcome of the causal discovery algorithm. However, Runge et al. (2019a) showed that, in the presence of finite data, conditioning on children of the parent variable dramatically reduces detection power of the conditional independence test and thus the performance of causal discovery algorithms. PCMCI (see Section 4.2.11) was specifically written to avoid conditioning on children of true causes that cannot be parents of the target variable, but it assumes that there are no simultaneous dependencies, and so it is able to leverage time information to identify which conditions to avoid. When simultaneous dependencies are allowed in PCMCI+ and LPCMCI, it is not possible to avoid this as easily.

$\mathrm{GT}(t-1)$ may persist as a potential parent of $\mathrm{GG}(t)$ because it is connected via multiple confounded pathways, and thus requires a large conditioning set to separate it from $\mathrm{GG}(t)$, but small sets are tested first. In our hypothesized diagram (Figure 4.4), there are many paths connecting $\mathrm{GT}(t-1)$ to $\mathrm{GG}(t)$, including $\mathrm{GT}(t-1) \leftarrow \mathrm{SA}(t-1) \rightarrow \mathrm{GG}(t)$, $\mathrm{GT}(t-1) \leftarrow \mathrm{Pc}(t-1) \leftarrow \mathrm{NA}(t-1) \rightarrow \mathrm{GG}(t)$, $\mathrm{GT}(t-1) \leftarrow \mathrm{TA}(t-1) \rightarrow \mathrm{AMM}(t) \rightarrow \mathrm{GG}(t)$, and multiple paths $\mathrm{GT}(t-1) \leftarrow \cdots \circ - \circ \mathrm{GG}(t-1) \rightarrow \mathrm{SA}(t-1) \rightarrow \mathrm{GG}(t)$. In LPCMCI's output, $\mathrm{Pc}(t-1)$ is required in addition to $\mathrm{SA(t)}$ (which introduces a selection bias into the calculation) to separate $\mathrm{GT}(t-1)$ from $\mathrm{GG}(t)$. Furthermore, to promote independence relative to the order of the variables in the dataset, all these causal discovery algorithms prioritize variables that correlate well (or have high conditional mutual information) with the target variable in the separating sets,

and due to autocorrelation effects, $\mathrm{GG}(t)$ has a higher unconditional mutual information with $\mathrm{GT}(t-1)$ than with its true parent $\mathrm{SA}(t-1)$ in all of the climate simulations. It would seem that the only way to successfully use LPCMCI with CMIknn on this dataset would be to remove all strong covariates including GT – the very index we wanted to examine.

To investigate whether this problem persists with a conditional independence test with higher detection power, we ran PCMCI+ on all tuning and analysis simulations using RobustParCorr (see Section 4.4.2) as the conditional independence test. Though RobustParCorr has higher detection power than CMIknn, the problem of producing backwards edges was even more severe than before: every climate simulation produced a backwards edge from SA to GG, and the non-linearity of EN (see section 4.4.4) does not appear to be the source of the problem. In PCMCI+, which does not allow latent confounding, only one potential separating set of every size is tested when learning the skeleton of the graph, and so the effect of prioritizing indices that correlate well with the target variable for separating sets is even more severe: an analysis of the verbose output in one case showed that $\mathrm{GT}(t-1)$ is retained in every potential separating set until the conditioning set gets large enough and the detection power of RobustParCorr small enough that $\mathrm{GG}(t)$ is separated from $\mathrm{SA}(t-1)$, and at the same time, $\mathrm{SA}(t-1)$ is used to help separate $\mathrm{GT}(t-1)$ from $\mathrm{GG}(t)$. Causal logic and basic probability theory dictate that these decisions are inconsistent with each other, but the algorithm does not check for such inconsistencies because they would not occur in the absence of faithfulness violations (see Section 4.2.8). The combination of these conflicting decisions means that PCMCI+ produces a graph that may not even be Markov to the data according to the very conditional independence test used to discover the graph.

4.5.5. Synthesis

Our orientation analysis (Section 4.5.4.c) showed that LPCMCI struggled more with some datasets than others, so while the discovered PAGs for high-performing datasets such as NCAR and CESM may be reasonably trusted (Figure 4.9a and b), the orientations of causal relationships in the discovered PAGs for low-performing datasets such as IPSL (Figure 4.9c and f) should not be taken with high confidence to represent observed or simulated causal relationships. Our investigation into the source of incorrect orientations (Section 4.5.4.d) showed that they can often be traced to strong faithfulness violations in the skeleton phase, rather than the orientation phase, of causal discovery, meaning that even the skeleton for low-performing datasets such as IPSL cannot be trusted.

Even for high-performing datasets, our robustness analysis (Section 4.5.4.a) shows that while a the discovered PAG is reasonably robust to small parameter changes, we may see a few causally-significant changes when perturbing the parameters or the dataset in the orientation of adjacencies and their time lag, and even in the basic skeleton of the discovered graph, that are not easily explained by the perturbations we made. While we would expect to see changes in the algorithm's decisions based on the parameters chosen for the conditional independence test, the changes that result from adding and removing supposedly-irrelevant variables are especially concerning, suggesting that the algorithm is not as robust to the presence of causally-unnecessary variables in real data as it appeared to be on partially-simulated data (Runge et al. 2019a). While we may place some scientific weight on the results for higher-performing datasets including NCAR and CESM, discovered causal relationships and lags still must be using targeted simulations and theoretical arguments.

According to our recall analysis (Section 4.5.4.b), it would appear that LPCMCI is especially prone to omitting true causal adjacencies (that are presumably statistically weak). If this does not result in false orientations, it might not cause a large practical problem for estimation of individual causal effects using the same dataset, because the omitted relationships may only weakly bias the results. But if the statistical strengths of our known edges are comparable to the strengths of other causal relationships we wished to test, then we cannot place strong physical significance on the absence of adjacencies in our discovered PAGs when forming a conceptual model of climate variability – omitted relationships may play a larger role in the future. We therefore cannot strongly interpret that fact that LPCMCI did not discover causal effects from GT to TT and pr.

However, if NARI truly captures the causal mechanism relating global SST to the Sahel, we would expect that NA together with TT or GT would be able to separate Sahel precipitation from all other ocean basins. Instead, in most simulations, LPCMCI finds that Sahel precipitation is directly driven by at least one ocean basin other than NA and GT (see Table 4.5). NCAR is the only simulation that does not display such a causal relationship, but Sahel precipitation is found to be confounded with SA, which is part of GT. If the orientation is correct, then NARI is not a clean causal driver of Sahel precipitation because the two are also confounded. If the skeleton is correct but the orientation is incorrect, then NCAR also demonstrates a causal effect of global SST on Sahel precipitation that is not mediated by NARI.

Unfortunately, we are not able to directly test the prevalence of false positive adjacencies in our performance analysis given our partial knowledge (see Sections 4.2.12 and 4.4.4), but we can directly validate individual dependence relationships. As explained in Section 4.2.10, false positives may arise during causal discovery or correlation analyses due to autocorrelation effects,

but conditioning on the parents of both the cause and the effect variables should address this problem. Causal discovery is not specifically necessary to test this hypothesis in general, but it is helpful for making sure we have not missed important variables to add to our conditioning set. Furthermore, a tuning analysis like we performed in Section 4.5.1 may be necessary to determine the best choice of parameters for conditional independence testing.

To validate our conclusion that NARI is not sufficient to explain the influence of global SST on Sahel precipitation independent of the decisions of our causal discovery algorithm, we directly test whether Sahel precipitation is independent from the ocean basins listed in Table 4.5 when conditioning on NA, GT, and additional variables chosen to eliminate autocorrelation effects. Reasonable choices for this additional conditioning set include the causal parents from our hypothesis (Figure 4.4), the discovered causal parents (both with and without potential parents, which are only partially-oriented $\circ\!\!\rightarrow$ in Figure 4.9), and combinations of these. Reasonable choices for CMIknn parameters include the CMIknn parameters employed to discover the pictured PAG (fourth major column of Table 4.4), and the 'ideal' CMIknn parameters based on oriented-recall alone (second column of Table 4.4, Section 4.5.1). We repeat each test ten times using combinations of these conditions, and the repeated tests almost uniformly give the same results. NARI can separate Sahel precipitation from lagged and simultaneous EN (peak El Niño) in NCAR and IPSL, and from Pc (the Pacific in summer) in IPSL (noted with italics in Table 4.5). Nevertheless, all other basins remain connected, confirming in every climate simulation that NA and GT are not sufficient to separate Sahel precipitation from global SST. Our analysis specifically suggests that summertime SSTs in the Pacific Ocean (Pc), the Gulf of Guinea (SA), the North Tropical Atlantic (TA), and the Mediterranean Sea (md) may be independently important for Sahel precipitation. Potential roles

for the Indian Ocean should be re-examined when LPCMCI achieves a higher performance on datasets that include it.

4.6. Discussion

One might try to dismiss the poor performance of (L)PCMCI(+) on the dataset examined here by arguing that our problem is contrived; after all, LPCMCI would not discover any detectable false orientations if we had not withheld seasonal time information (see Section 4.4). However, I argue that hiding partial time information from LPCMCI did not cause these difficulties, it simply exposed LPCMCI's weaknesses by preventing us from justifying the incorrectly-oriented edges with some contrived physically-plausible explanation.

In Section 4.5.4.d, we find that the true source of the backwards edges in the Gulf of Guinea is often not the high correlation of GG and SA, but the presence of another causally-irrelevant variable that happens to correlate with the target better than its true parent. This reflects the combination of two common problems in time-series causal discovery that the authors of (L)PCMCI(+) identify and specifically try to address (see Section 4.2.11). The first is that conditioning on related variables that turn out to be not causally-relevant decreases the detection power of conditional independence tests and thus leads to strong faithfulness violations (see Section 4.2.8). The second is that autocorrelated time series data by definition violate the independence assumption, and can artificially inflate the apparent dependence between two variables (see Section 4.2.10).

Surprisingly, the implications of these practical finite-data effects on conditional independence tests and on the performance and consistency of causal discovery algorithms are never addressed rigorously in theory. All causal discovery algorithms have been proven to be consistent, meaning they make correct decisions and converge on the correct graph; but, as noted

in Section 4.2.12, most of these proofs rely on faithfulness (and all other assumptions of the algorithm) and assume an 'oracle' conditional independence test that is not constrained by the statistics of finite samples and always tells the algorithm truthfully whether or not the probability distributions of two variables are conditionally independent.

Reliance on the oracle conditional independence test means that decreased detection power and autocorrelation effects do not factor into the proof of consistency, and Runge et al. (2019a) offer no theoretical results on the implications of iid violations on conditional mutual information or distance correlation metrics, let alone the effects that changes in these metrics would have on PCMCI. In the end, autocorrelation effects are only addressed in the last stage of the algorithm without consideration for how they might affect the rest of the discovered graph in the earlier stages of the algorithm. A theoretical analysis of these finite-data effects might show, for instance, that—because correlation (and potentially other dependence metrics) can be artificially inflated by autocorrelation effects, and because increasing the size of the conditioning set reduces detection power—it is unwise in PCMCI to check only one potential conditioning set of each size consisting of the variables that are most strongly correlated, even though this approach has been proven to be sound in the oracle case.

Furthermore, the reliance on the assumption of (strong) faithfulness means that most causal discovery algorithms (not just time-series algorithms and not just those available through tigramite) are designed to run potentially-conflicting conditional independence tests—sometimes in order to promote independence of the discovered graph from the order in which the variables appear in the dataset (Colombo and Maathuis 2014), and sometimes simply of poor design[18]—

[18] The classic PC algorithm (Spirtes and Glymour 1991) checks potential separating sets for X and Y using potential parents of Y whether or not they are d-separated from X; in the oracle case, variables not d-connected to both X and

without ever considering the possibility that the results might conflict, either due to low detection power or true faithfulness violations.

There is certainly room for improvement in the conditional independence tests themselves to reduce the prevalence of strong faithfulness violations and approach the oracle case: perhaps conditional mutual information can be estimated using Empirical Orthogonal Functions like distance correlation, and perhaps distance correlation can be generalized to allow for conditional independence testing so that it no longer applies only to Gaussian processes. But it may also be that true faithfulness violations are more prevalent than previously considered in the time series case, where causal lags must be rounded up or down to match discrete timesteps in a way that may not scale over multiple causal effects.

If the performance of time-series causal discovery is to be improved and the results are to be trusted, there is an urgent need for theoretical results regarding the implications of iid violations on conditional independence metrics, and regarding the implications of autocorrelation effects and (strong) faithfulness violations on the consistency of causal discovery algorithms at every stage.

In lieu of theoretical results, Runge et al. (2019a) attempt to prove computationally that they have succeeded in addressing low detection power and autocorrelation effects by showing that PCMCI performs better than other causal discovery methods on a specific contrived dataset with two real variables and additional generated causally-irrelevant variables, and on a large ensemble of entirely-generated datasets for causal discovery where the ground-truth is known. Since the true causal structure of observed data is usually *not* known, it is standard to test causal

---

Y should not be able to help separate X and Y. I am currently exploring ways to address the potential sources of inconsistency in the PC algorithm.

discovery algorithms by generating Structural Causal Models (SCMs; see Section 4.2.2), and the stated goal of such SCM generation (which is often not achieved; see Reisach et al. 2021) is usually to generate truly-random causal graphs with truly-random causal effects to prevent the algorithm from taking unfair advantage of patterns in the simulated data.

But in real causal discovery problems, the scientist does not randomly choose indices that may or may not be related; she chooses indices because they are correlated with each other and believed to be related. The scientist is furthermore only likely to turn to causal discovery if she does not know which variables are causally relevant, and if she believes a simple correlation analysis will fail to identify the most prominent driver of her target variable. I argue that simulated SCMs should not be generated randomly, but should instead be selected for high correlation between variables—and perhaps even for SCMs where the lagged variable with the highest correlation is not a true causal parent—to fairly represent causal discovery tasks the algorithm will likely have to face. If causal discovery is to be preferred over correlation analysis for discovering the one strongest causal driver, we must show specifically that causal discovery can identify the correct causal parents when other variables correlate better with the target – not only that it can recover randomly-generated SCMs. If causal discovery algorithms cannot perform well in data environments like these, it undermines the main selling points of causal discovery: that it can separate direct effects from indirect effects (mediated by another highly-correlated variable), and that it is able to distinguish between variables that are correlated and variables that are truly causally-relevant. In our dataset, it would appear that (L)PCMCI(+) cannot handle the inclusion of the potential mediator we wished to test.

Additionally, algorithm performance is always tied to recovery of the underlying graph, but it is not standard to check for inconsistencies that might easily result from conflicting

decisions that violate strong faithfulness. This is imperative because, to the extent that causal discovery is used to generate background assumptions for causal effect estimation, the discovered diagram cannot be used in good faith if its testable implications do not hold in the data. While an inappropriate choice of conditional independence test might affect the algorithm's ability to recover the underlying graph, a practically-sound algorithm should always return a graph that is Markov to the data according to the chosen conditional independence test. Thus, I argue that Markov-consistency of the discovered PAG should actually be valued as a performance metric above recovery of the generating causal diagram, and call first for the improvement of causal discovery algorithms according to this metric.

4.7. Conclusions

In this chapter, we employ a novel technique known as *causal discovery* to analyze the simulated impacts of sea surface temperature (SST) variability in various ocean basins on Sahel precipitation and on each other in the Coupled Model Intercomparison Project phase 6.

Our results suggest that the North Atlantic Relative Index (NARI) does not mediate the full simulated effects of all ocean basins on Sahel precipitation. Instead, we find that simulated Sahel precipitation in most of the examined climate models responds to summertime SST in the Pacific Ocean, the Gulf of Guinea, the North Tropical Atlantic, or the Mediterranean Sea via a pathway that is not mediated by NARI. In fact, our results fail to detect any causal dependence of Sahel rainfall on the Global Tropics – a key component of NARI.

We cannot initially place too much physical significance on either of these findings because, according to our performance analysis, discovered causal graphs are prone to errors in the discovered causal lags, the orientations of causal relationships, and even in the skeleton of the graph, likely resulting from violations of algorithmic assumptions. However, we verify the

first claim by confirming that NARI still does not separate simulated Sahel rainfall from summertime SST in the identified basins even when we rely on our causal hypothesis, rather than the error-prone discovered graph, to determine how to remove autocorrelation effects. To evaluate this, we rely on the results of our tuning experiment to determine conditional independence test parameters. The second claim—that NARI does *not* affect Sahel precipitation—cannot easily be verified. But even if we still believe that NARI has some direct effect on Sahel precipitation, the fact that individual ocean basins also have their own direct effects complicates estimation of the causal effect of NARI. Summertime SST in the Pacific, the North Tropical Atlantic, and the Gulf of Guinea all contribute to the Global Tropics index by construction, and the North Tropical Atlantic additionally impacts the North Atlantic Ocean in every single climate simulation, perhaps through surface water fluxes. Whenever any of these basins also has a direct effect on the Sahel, it confounds the relationship between NARI and Sahel precipitation, meaning that the causal effect cannot be responsibly estimated using bivariate regression as done in Chapter 3.

Even though NARI, as it is defined here, does not capture the full effect of tropical—let alone global—SST variability on Sahel rainfall, it may still be possible that a single index could suffice to capture the effects of SST variability throughout the tropics on the Sahel. Perhaps the "upped ante" mechanism discussed in G13 would be best captured by tropical-maximum SST or precipitation-weighted mean SST (see Section 1.4) rather than the area-weighted mean SST employed by NARI, which potentially over-simplifies and obscures the causal mechanism. The practical need for this distinction may not have been clearly evident in the coupled historical simulations and simulated projections employed by G13, both of which are dominated by anthropogenic forcing that causes SST in the tropical ocean basins largely to vary together.

However, in freely-varying pre-Industrial control simulations such as those used in this chapter, tropical-mean and -maximum SST are more likely to diverge, exposing the limitations of the NARI index.

So far, we have commented only on causal relationships in simulations, but when multiple climate models from the ensemble all demonstrate the same relationships, we gain confidence that certain claims may hold in observations as well. Our discovered causal graphs demonstrate a number of similarities. NARI does not mediate the effects of global SST on Sahel rainfall in *any* of the simulations we examine, so our claim that individual ocean basins directly impact Sahel rainfall is likely to hold in observations. Our results can also comment on some causal relationships between ocean basins that are currently debated in the literature. For instance, our analysis supports the claim that the Atlantic Meridional Mode can trigger El Niño. Such analysis has scientific merit in its own right, and furthermore could be important background information that affects how the NARI teleconnection and a wide range of other important causal relationships in the observed climate system should be estimated from the observational record.

This study also detects significant differences in the causal structures of the climate models which could have important implications for causal effect estimation. Notably, our discovered causal graphs for different simulations often differ in edge orientation such that they are incompatible with each other, meaning that a different mathematical expression may be required for each climate model when attempting to estimate a causal effect from non-interventional simulations. When ensemble-mean methods are applied to non-interventional data, as we did when examining the effect of SST on Sahel precipitation in coupled simulations in Chapter 3, they implicitly assume that the causal structures underlying different simulations are

the same, and so they cannot responsibly be used over the entire ensemble when climate models differ in this way. Since our performance analysis shows that small parameter and dataset perturbations can lead to some incompatible changes in our discovered diagrams, we cannot place too much confidence on the conclusion that teleconnections in different climate models are causally incompatible with each other. However, even if the causal graphs for different climate models and observations were oriented such that they did not directly conflict, the graphs have some notable differences in connectivity. For example, in CNRM, the Atlantic Niño responds strongly to peak El Niño conditions and affects springtime SST in the North Tropical Atlantic, while none of the other climate models demonstrate such an effect. These differences also suggest that attempting to extract causal information from ensemble-means is not optimal because different physical processes are active in different simulations.

Sometimes the discovered graphs also conflict with our expectations based on the observed climate system. For instance, the previously-mentioned causal effect of spring Gulf of Guinea SST (GG) on spring North Tropical Atlantic SST (AMM) in CNRM conflicts with the expected effect of the Atlantic Meridional Mode on the Atlantic Niño in observations. When causal links in our hypothesis were inspired by observed associations, we would hope that the results of causal discovery that are robust to climate model parameterization could rest on equal ground with prior beliefs and help us improve scientific theory. Furthermore, if we are able to obtain a trustworthy causal representation of observations using some combination of theory and causal discovery applied to observations and simulations, and if we can trust that the differences in discovered graphs represent true differences in the underlying dynamics, we could constrain projections and help ensure that even ensemble means will be meaningful and representative of observed dynamics by selecting only climate models whose causal graph is compatible with our

understanding of observations (Nowack et al. 2020). However, it is not immediately clear whether the differences we see between climate models and observations are due to differences in the underlying climate dynamics, statistical errors in the causal discovery algorithm, or confounding and autocorrelation effects that biased our prior scientific beliefs.

Unfortunately, our performance analysis suggests that time series causal discovery algorithms such as LPCMCI and its earlier variants, while promising, are not yet robust methods for learning causal relationships between climate variables. According to our analysis, the algorithms appear to make mistakes for a number of reasons that could be addressed. First, the algorithms rely so strongly on the computational assumption of strong (adjacency) faithfulness that they do not include implementable provisions to prevent the algorithm from making contradictory decisions when the assumption is violated. Second, because the theory of autocorrelation effects is not fully developed, it is overlooked in some stages of the algorithm. In practice, this second concern appears to make these causal discovery algorithms somewhat susceptible to the same pitfalls as correlation analysis, leading them to sometimes treat the variables that correlate best, rather than the true causal drivers, as parents of the target variable during many stages of the causal discovery algorithms.

To improve the performance of causal discovery algorithms for time series, there is an urgent need for theoretical (and computational) results regarding the implications of iid violations on (statistical) conditional independence testing, and also the implications of strong faithfulness violations and autocorrelation effects on the consistency of causal discovery algorithms. Furthermore, automated evaluation of causal discovery algorithms should be tailored to the type of problems researchers are likely to face. Until these goals are accomplished, we

hope that our analysis of the reasons for LPCMCI's failures on this dataset will help usher in needed improvements to the algorithms.

In the meantime, it is not possible to determine whether differences between discovered graphs, or between discovered graphs and prior beliefs about observations, are due to mistakes in the prior beliefs, mistakes in the causal discovery algorithm, or true differences in the underlying dynamics. Thus, while current causal discovery algorithms may still be useful for preliminary characterization of poorly-understood causal relationships or large ensembles of models with differing behavior, the results of causal discovery should be combined with domain-specific theory as much as possible, and should then be validated by checking testable implications in the data. Furthermore, if the results of causal discovery will be used as the underlying causal assumptions for causal effect estimation, then there is a need to state uncertainty in the discovered graph and to propagate that uncertainty through effect estimation (such an automated procedure is currently being developed).

Even if causal discovery algorithms do not perform well, going through the exercise of formalizing implicit beliefs and assumptions in the form of a causal diagram is paramount for directing appropriate causal analysis of non-interventional data. We hope that our tuning experiments will provide a partial basis for selecting parameter values for targeted testing of analysis assumptions, specific causal relationships, and entire causal diagrams.

Conclusion

Identifying the true causal drivers of climate disasters, such as the drought in the Sahel, is essential for prediction, mitigation, and prevention efforts in a changing climate. Unfortunately, it is not possible to perform randomized controlled experiments on large-scale climate phenomena in the observed climate system, so results in climate science are often derived from simulated experiments on individual climate models that may not represent the true climate system, from observations or large ensembles of climate models using correlation analyses or other statistical association techniques that may suffer from autocorrelation effects and confounding, or from theoretically-motivated narratives and *storylines*, or physically self-consistent, plausible pathways (Shepherd 2019), that are contrived to explain observed or simulated associations.

Initial explanations for the Sahel drought focused on the impacts of local land-use change. But since Giannini et al. (2003) showed that observed global sea surface temperature (SST) changes could cause simulated Sahel precipitation variability that correlates with observations, most prominent storylines explain Sahel rainfall using SST. In Chapter 1, we introduce in depth the theory necessary to understand one of the prominent storylines for multidecadal Sahel rainfall change. The narrative is that Sahel rainfall decreased in the 1970's and 80's because moisture supply from the North Atlantic decreased due to anthropogenic aerosol-induced cooling, while the moist static energy threshold for convection was simultaneously raised due to greenhouse gas-induced warming of the global tropics. According to this storyline, the rains (partially) recovered after the 80's because moisture supply from the North Atlantic was able to meet the "ante" for convection after aerosol emissions reduced in response to clean air legislation, warming the North Atlantic. Other storylines attribute Sahel

rainfall variability to changes in the location of the rainband caused by the response of SST gradients to anthropogenic and volcanic aerosols (without a role for greenhouse gases) or to unforced SST variability internal to the climate system. Because all of these competing narratives were inspired in part by the observed statistical association between Sahel rainfall and global SST, it is difficult to differentiate between them in observational studies.

In Chapter 2, we enter the debate on whether or not Sahel precipitation variability was caused by anthropogenic emissions, taking advantage of the release of the 5th phase of the Coupled Model Intercomparison Project (CMIP5), which is a large ensemble of climate models that for the first time includes "detection and attribution" simulations in which the coupled atmosphere-ocean system responds to prescribed combinations of historical anthropogenic emissions and natural (non-anthropogenic) radiative forcings. The simulations are *interventional*, allowing us to make causal claims about the total effects of different forcing agents on Sahel precipitation in climate models. Because all climate models make somewhat arbitrary simplifying "parameterizations" of climate physics that may induce biases into the resulting climatology, it is not possible to confidently generalize causal results from an individual climate model to the observed environment, but the mean over a large ensemble of simulations with different parameterization choices, like CMIP5, has a much better chance of representing the observed climate system. Of course, ensemble-wide biases are also prevalent, and the performance of the ensemble still must be validated against observations.

Simulated ensemble-mean Sahel precipitation variability in CMIP5 *correlates* significantly with observations. To the extent that correlation and magnitude represent the performance of the simulations, we confirm that the ensemble-mean outperforms individual models. This is far from a complete validation of the model ensemble, but might optimistically

be interpreted to mean that true climate dynamics lie somewhere in between the dynamics of individual simulations in the ensemble.

To the extent that simulations from CMIP5 can explain observed variability, we find that the ensemble would attribute observed rainfall changes to changing atmospheric aerosol concentrations alone – a finding that is consistent with other studies using CMIP5 (e.g. Polson et al. 2014). Specifically, CMIP5 does not support any coherent role for historical greenhouse gas emissions in driving Sahel precipitation variability, and suggests that changing historical anthropogenic and volcanic aerosol emissions contributed to both the observed drought and the observed recovery, with anthropogenic emissions mainly responsible for low-frequency variability. These claims are supported by the fact that simulated unforced internal variability does not explain the observations as skillfully as the ensemble mean response to anthropogenic aerosols. The results appear to be consistent with one of the alternate storylines of Sahel precipitation variability – that increasing and then decreasing anthropogenic and volcanic aerosols affect the interhemispheric temperature gradient which in turn shifts the rainband meridionally, affecting how much of the monsoon rains fall in the Sahel. However, we also find that the simulated processes in CMIP5 cannot account for observed variability because the magnitude of simulated low-frequency variability is much smaller than observed, even after bias correction for total variability. This is evidence that there are differences between observed and simulated dynamics, thus preventing responsible use of these simulations for *attribution*, or claims about the fraction of observed variability due to specific causal drivers.

Nevertheless, because the correlation of CMIP5 ensemble-mean precipitation with observations is relatively high and significant relative to other simulated patterns, some studies (e.g. Hua et al. 2019) have been tempted to attribute observed historical rainfall change to

anthropogenic aerosol emissions, even concluding that anthropogenic aerosols likely were the primary driver of observed changes. Such a claim relies on the implicit assumption that model biases are more likely to affect the magnitude of simulated variability than the pattern, and explicitly on the assumption that there is no poorly-simulated real-world process that produces the same pattern of variability as the simulations. Clearly, all studies that standardize or focus solely on correlations (including Giannini and Kaplan 2019; Held et al. 2005) also rely on these assumptions.

Additionally, even studies that acknowledge attribution is impossible when simulated variability differs in magnitude from observations (e.g. Undorf et al. 2018) still rely on both of these assumptions in the same way if they instead make claims about *detection*, or verifying that the simulated response truly reflects an observed process. The premise is that if the pattern of variability associated with a simulated process is statistically distinguishable from other processes, then finding that same pattern in observations—even at a different magnitude—is evidence that the simulated process occurred in the real world. However, when observed variability is larger than simulated variability, if for any reason the simulated signal *cannot* be simply scaled to match the magnitude of the pattern in observations, then the difference in magnitude inadvertently provides evidence of poorly-simulated distinct observed processes that could in theory just as easily explain the entire magnitude of the simulated pattern in observations. Thus, a violation of the first, implicit assumption implies a violation of the second, explicit assumption, and destroys confidence in the conclusion that the simulated process is detectable in observations.

If simulated variability cannot explain the full magnitude of observed variability, one could assume (rather than attempting to verify) that the simulated experiments represent *a subset*

of observed processes. Under this assumption, all simulated processes must be taken at their actual magnitudes regardless of whether they correlate well with observations unless there is some physical, rather than statistical, reason to correct biases in the simulations. In this context, we may re-interpret the significance of the high correlation between observed and aerosol-driven ensemble-mean precipitation, seen in Chapter 2, as a quantification of our confidence that the response to anthropogenic aerosols, though small, *enhances* the intensity of the observed drought rather than inhibiting it or having a negligible effect. However, we cannot claim that the simulated mechanism is stronger in the observed record without finding evidence that some specific physical process should be amplified, and we cannot even take this as a confirmation that the storyline relating anthropogenic aerosols to Sahel precipitation is the dominant explanation for simulated rainfall variability without examining potential mediating variables to confirm the storyline.

In Chapter 3, we begin to address these goals by examining simulated SST, which is an important causal driver or mediator in all of the prominent storylines for Sahel precipitation variability. This will not help us distinguish between the narratives for Sahel precipitation, but it will help us understand what physical processes would need to be strengthened to bring the magnitude of simulated variability into agreement with observations, and to begin to evaluate whether this would make sense physically.

Atmospheric simulations with prescribed global SST matching observed historical variability from the next generation of this climate model ensemble (CMIP6) capture the full magnitude of Sahel precipitation variability for the first time, confirming the importance of global SST in driving observed Sahel rainfall variability, and allowing us to make attribution claims about observed historical Sahel precipitation variability based on these atmospheric

simulations. In coupled simulations with prescribed radiative forcings from both CMIP5 and CMIP6, on the other hand, we find that anthropogenic aerosols cause simulated NARI variability that correlates positively with observations but is smaller in magnitude. At face value, this appears to be consistent with our narrative and to suggest that the simulated SST response to anthropogenic aerosols is amplified in observations. However, we also surprisingly find that simulated variability in NARI's component basins – the North Atlantic and the Global Tropics – is inconsistent with observations and all narratives relating anthropogenic aerosols to Sahel precipitation variability. It would appear that the positive correlation of observed NARI with the simulated NARI response to anthropogenic aerosols is to be due to compensating errors in the two basins. Given these results, we cannot claim that the mechanism of Sahel rainfall change from coupled simulations is amplified in observations because this would result in exacerbated unrealistic differences between simulated and observed SST in the Global Tropics. Instead, we must conclude that CMIP5 ensemble-mean Sahel precipitation correlates with historical observations despite an inability to reproduce the physical phenomena that were most important for the Sahel in the 20th century.

Though the response of NARI to radiative forcing is the same in CMIP5 and CMIP6 and the atmospheric components of the CMIP6 climate models perform incredibly well when provided with observed SST, the ensemble-mean of Sahel precipitation in CMIP6 no longer correlates with observations. Using the atmospheric simulations, we show that the fast response to radiative forcing (not mediated by SST) is not a dominant contributor to observed 20th century Sahel precipitation variability because it is overwhelmed by the influence of observed global SST variability. But in the coupled simulations, relevant SST variability (according to NARI) is

much smaller than observed, so changes in the simulation of the fast responses to anthropogenic aerosols and greenhouse gases are likely to explain the differences between CMIP5 and CMIP6.

Because the atmospheric simulations intervene on global SST, we may directly discuss the simulated causal effects of the observed SST record. Since these simulations appear to explain the pattern and full magnitude of observed low-frequency precipitation variability, we can claim with reasonable confidence that we will not be able to explain the root causes of observed Sahel precipitation variability until we can account for relevant observed SST variability around the globe. We use NARI throughout the chapter to represent the relevant causal information in the global SST field, and we can see that not only are SSTs from the coupled simulations with prescribed radiative forcing inconsistent with observed variability in both the North Atlantic and the Global Tropics, but also that no linear combination of simulated internal variability and the SST responses to anthropogenic aerosols and greenhouse gases could possibly bring simulated SST into alignment with observations in the North Atlantic, which is primarily responsible for low-frequency variability in NARI. As with precipitation, this means that we cannot determine whether observed multidecadal variability in the North Atlantic is anthropogenic or natural, and further work is required to explain the observed record.

Accounting for observed NARI variability would be a start to explaining the root causes of Sahel precipitation, but may not be sufficient. We show using the atmospheric simulations that a linear relationship with NARI is not a complete substitute for the effects of global SST: while NARI matches the pacing of the simulated precipitation response to global SST in the atmospheric simulations, it only explains half of the simulated variability and cannot account for the magnitude of the drought. Thus, though scientific consideration of global SST and NARI in particular as drivers of Sahel precipitation has improved scientific understanding of Sahel rainfall

variability, a more complete characterization of teleconnections affecting Sahel rainfall will now be necessary to explain observed variability. Unfortunately, we cannot identify other important and poorly-simulated ocean basins using the methodology used in this chapter because there are almost no ensemble simulations that prescribe the temperature of different ocean basins independently.

In Chapter 4, we attempt to characterize the simulated impacts of SST variability in individual ocean basins on Sahel precipitation using available non-interventional CMIP6 simulations by turning to a novel technique known as *causal discovery*, which is designed to extract causal information from non-interventional data when possible. Part of the process involves characterizing causal relationships between prominent modes of internal climate variability that affect SST in the ocean basins we analyze, allowing this approach ideally to identify and address confounding and reduce autocorrelation effects which could otherwise lead standard statistical techniques to mistake confounded covariates for causal drivers.

To the extent that prominent physical narratives (and storylines) are contrived to explain observed correlations without explicit regard for confounding or autocorrelation effects, we would expect the output of causal discovery to meaningfully conflict with them to some degree. Such conflicts could help refine our scientific understanding of causal relationships, or could dramatically redefine the set of prominent narratives by identifying (parts of) storylines that are actually inconsistent with the data that inspired them, and instead proposing alternate causal connections that could inspire physical storylines and narratives more likely to represent the observed climate system. Unfortunately, we find that current causal discovery algorithms for non-linear time series are prone to make mistakes that limit the advantages of causal discovery, tempering our confidence in some of the results and preventing us from making many strong

conclusions about observations. Nevertheless, there are some things we can learn from the results.

In each examined climate simulation, the Global Tropics – a key component of NARI – does not have any direct effect on Sahel precipitation, while individual ocean basins within the Global Tropics directly impact Sahel rainfall without mediation by NARI. It is possible that taking a spatial-mean might not be ideal for capturing the proposed causal mechanism that relates the Global Tropics to the Sahel. Due to the uncertainty of our results, more work would be needed to verify that the Global Tropics, even as it is currently defined, actually has no direct impact on Sahel precipitation variability. Nevertheless, we are able to verify that NARI does not mediate the effects of individual ocean basins on simulated Sahel rainfall even when we replace or combine the uncertain discovered causal structure relating different ocean basins to each other with current theory and storylines. If there is a direct effect of NARI on Sahel precipitation, the fact that the tropical ocean basins that directly impact the Sahel also contribute to simulated NARI variability implies that our estimation of the NARI teleconnection in Chapter 3 is likely confounded, with additional implications for estimation of P_{nonNARI}. Regardless, we conclude that we may not be able to explain Sahel precipitation variability without explaining variability in the Mediterranean Sea and in *individual* tropical ocean basins, even when variability in different basins cancels in the global tropical mean.

Regarding internal variability that might confound relationships between individual ocean basins and Sahel rainfall, we find that the discovered causal structure for a given climate model is likely to differ from other climate simulations. Often the differences can be such that the climate simulations are not compatible with each other under the assumption of acyclicity, which is required for most current methods of causal discovery and causal effect estimation. It is

not always clear to what degree these discrepancies reflect differences in the dynamics between the climate models rather than mistakes in the causal discovery output resulting from the statistical limitations of finite data. Nevertheless, to the extent that differences in the structure of simulated causal relationships exist and affect the appropriate mathematical expression for estimating the chosen causal effect, we conclude that it may be advantageous or even necessary to examine individual simulations rather than the ensemble mean when trying to estimate a causal effect from non-interventional simulations.[19]

At the conclusion of this dissertation, the correct narrative and storyline for historical Sahel precipitation variability is still an open question. We hope that future work will continue to attempt to distinguish between different storylines by examining mediating variables related to the proposed mechanism of variability. We have particular interest in distinguishing between the importance of local instability and large-scale dynamics, and in determining the role of moisture supply. We look forward to the continuing steady efforts of the climate science community to determine the degree to which observed North Atlantic multidecadal variability was driven by anthropogenic and volcanic radiative forcing, which would be a first step toward attribution of Sahel precipitation. There is an urgent need to improve simulations of SST so that the combination of the response to radiative forcing and internal variability is consistent with observations. Of course, if past variability was mostly internal, then while truly capturing the physical processes important for past variability will be paramount for near-term predictions of Sahel rainfall even in a warmer world, century-long future projections might still diverge under

[19] When we are interested in teleconnections from individual ocean basins to Sahel rainfall, available atmospheric simulations with prescribed SST must be considered non-interventional because they prescribe global observed historical SST variability, and so variability in different ocean basins is confounded by observed radiative forcing and interactions between ocean basins.

the dominant control of anthropogenic emissions. To gain trust in future projections, we must specifically address the simulation of the SST and fast precipitation responses to radiative forcing, since our results suggest that these simulated processes may be inconsistent with the observed environment even if they are not important for explaining the historical record.

We take some lessons from this work going forward. The evolution throughout the course of this dissertation of our perception of the success of CMIP in capturing observed West African Monsoon dynamics tells a cautionary tale, both for the evaluation of climate models and for the validation of explanatory narratives and storylines. Though having a physically-plausible explanation for the way in which two correlated variables might be related (as we did for anthropogenic aerosols and Sahel precipitation in Chapter 2) does increase the likelihood that a simulated or observed association is causal, it is essential to confirm the true mechanisms of simulated and observed variability. Storylines inspired by observed correlations cannot be evaluated against those same observations, and ensemble-means of simulations that prescribe the hypothesized cause variable are not much better, especially when magnitude is ignored. Compensating errors between mediating variables (as we saw for the components of NARI in Chapter 3) or between simulations (as suggested might be likely in Chapter 4) might mask important differences between the data and the theory, and autocorrelation effects and confounding complicate analysis of non-interventional simulations and observations (as shown for NARI in Chapter 4). Testing our storylines robustly either requires a huge number of CMIP simulations that intervene on all intermediate variables of interest, or the use of causal discovery and causal effect estimation to extract causal information from existing CMIP simulations and observations.

There is a lot of work that still needs to be done to improve the performance of causal discovery algorithms for non-linear time series. Specifically, the algorithms would benefit from theoretical characterization of the implications of iid and faithfulness violations that could motivate provisions in the algorithms to attempt to detect such violations and prevent them from having cascading effects on the resulting causal graph. Nevertheless, despite the struggles of current causal discovery algorithms, thinking informed by causal inference is necessary to glean needed causal information about the climate in the absence of controlled experiments. Correlation, regression, empirical components, fingerprinting, and all methods for detecting and quantifying causal effects from non-interventional data implicitly rely on assumptions that are just as strong as the assumptions of causal effect estimation, and that may not always hold. Because differences in these assumptions affects the appropriate way to estimate a causal effect from observational data, all observational studies would benefit from first formalizing the hypothesized causal structure of the data.

References

Ackerley, D., B. B. B. Booth, S. H. E. Knight, E. J. Highwood, D. J. Frame, M. R. Allen, and D. P. Rowell, 2011: Sensitivity of twentieth-century Sahel rainfall to sulfate aerosol and CO2 forcing. *J. Climate*, **24,** 4999-5014, https://doi.org/10.1175/JCLI-D-11-00019.1.

Alexander, M. A., D. J. Vimont, P. Chang, and J. D. Scott, 2010: The impact of extratropical atmospheric variability on ENSO: Testing the seasonal footprinting mechanism using coupled model experiments. *J. Climate*, **23,** 2885-2901.

Alexander, M. A., I. Bladé, M. Newman, J. R. Lanzante, N.-C. Lau, and J. D. Scott, 2002: The atmospheric bridge: The influence of ENSO teleconnections on air–sea interaction over the global oceans. *J. Climate*, **15,** 2205-2231, https://doi.org/10.1175/1520-0442(2002)015<2205:TABTIO>2.0.CO;2.

Allen, M. R., and W. J. Ingram, 2002: Constraints on future changes in climate and the hydrologic cycle. *Nature*, **419,** 228-232, https://doi.org/10.1038/nature01092.

Andersen, H., 2013: When to expect violations of causal faithfulness and why it matters. *Philosophy of Science*, **80,** 672-683.

Arakawa, A., and W. H. Schubert, 1974: Interaction of a cumulus cloud ensemble with the large-scale environment, Part I. *J. Atmos. Sci.*, **31,** 674-701.

Ashok, K., S. K. Behera, S. A. Rao, H. Weng, and T. Yamagata, 2007: El Niño Modoki and its possible teleconnection. *Journal of Geophysical Research: Oceans*, **112**.

Bader, J., and M. Latif, 2003: The impact of decadal-scale Indian Ocean sea surface temperature anomalies on Sahelian rainfall and the North Atlantic Oscillation. *Geophys. Res. Lett.*, **30**, https://doi.org/10.1029/2003GL018426.

——, 2011: The 1983 drought in the West Sahel: a case study. *Climate Dynam.*, **36,** 463-472.

Baek, S. H., Y. Kushnir, M. Ting, J. E. Smerdon, and J. M. Lora, 2022: Regional Signatures of Forced North Atlantic SST Variability: A Limited Role for Aerosols and Greenhouse Gases. *Geophys. Res. Lett.*, **49,** e2022GL097794.

Becker, A., P. Finger, A. Meyer-Christoffer, B. Rudolf, K. Schamm, U. Schneider, and M. Ziese, 2013: A description of the global land-surface precipitation data products of the Global Precipitation Climatology Centre with sample applications including centennial (trend) analysis from 1901–present. *Earth Syst. Sci. Data*, **5,** 71-99, https://doi.org/10.5194/essd-5-71-2013.

Bellomo, K., L. N. Murphy, M. A. Cane, A. C. Clement, and L. M. Polvani, 2018: Historical forcings as main drivers of the Atlantic multidecadal variability in the CESM large ensemble. *Climate Dynam.*, **50,** 3687-3698, https://doi.org/10.1007/s00382-017-3834-3.

Biasutti, M., 2013: Forced Sahel rainfall trends in the CMIP5 archive. *J. Geophys. Res.-Atmos.*, **118,** 1613-1623, https://doi.org/10.1002/jgrd.50206.

——, 2019: Rainfall trends in the African Sahel: Characteristics, processes, and causes. *Wiley Interdisciplinary Reviews: Climate Change*, **10**, https://doi.org/10.1002/wcc.591.

Biasutti, M., and A. Giannini, 2006: Robust Sahel drying in response to late 20th century forcings. *Geophys. Res. Lett.*, **33**, https://doi.org/10.1029/2006GL026067.

Biasutti, M., I. M. Held, A. H. Sobel, and A. Giannini, 2008: SST forcings and Sahel rainfall variability in simulations of the twentieth and twenty-first centuries. *J. Climate*, **21,** 3471-3486, https://doi.org/10.1175/2007JCLI1896.1.

Bird, G., and S. Medina, 2002: *Africa environment outlook: past, present and future perspectives.* UNEP.

Birkel, S. D., P. A. Mayewski, K. A. Maasch, A. V. Kurbatov, and B. Lyon, 2018: Evidence for a volcanic underpinning of the Atlantic multidecadal oscillation. *NPJ Climate and Atmospheric Science*, **1,** 1-7.

Bjerknes, J., 1969: Atmospheric teleconnections from the equatorial Pacific. *Mon. Weather Rev.*, **97,** 163-172.

Bohren, C. F., and B. A. Albrecht, 2000: Atmospheric thermodynamics. American Association of Physics Teachers.

Bongers, S., P. Forré, J. Peters, and J. M. Mooij, 2021: Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, **49,** 2885-2915.

Booth, B. B., N. J. Dunstone, P. R. Halloran, T. Andrews, and N. Bellouin, 2012: Aerosols implicated as a prime driver of twentieth-century North Atlantic climate variability. *Nature*, **484,** 228-232, https://doi.org/10.1038/nature10946.

Bretherton, C. S., and D. S. Battisti, 2000: An interpretation of the results from atmospheric general circulation models forced by the time history of the observed sea surface temperature distribution. *Geophys. Res. Lett.*, **27,** 767-770.

Bruins, H. J., 2019: Drought Mitigation Policies: Waste Water Use, Energy and Food Provision in Urban and Peri-Urban Africa. *Urban and Peri-Urban Agriculture in Africa*, Routledge, 257-266.

Chan, D., and P. Huybers, 2021: Correcting observational biases in sea surface temperature observations removes anomalous warmth during World War II. *J. Climate*, **34,** 4585-4602.

Chan, D., E. C. Kent, D. I. Berry, and P. Huybers, 2019: Correcting datasets leads to more homogeneous early-twentieth-century sea surface warming. *Nature*, **571,** 393-397.

Chang, C.-Y., J. Chiang, M. Wehner, A. Friedman, and R. Ruedy, 2011: Sulfate aerosol control of tropical Atlantic climate over the twentieth century. *J. Climate*, **24,** 2540-2555, https://doi.org/10.1175/2010JCLI4065.1.

Charney, J. G., 1975: Dynamics of deserts and drought in the Sahel. *Q. J. Roy. Meteorol. Soc.*, **101,** 193-202, https://doi.org/10.1002/qj.49710142802.

Chiang, J. C., and A. H. Sobel, 2002: Tropical tropospheric temperature variations caused by ENSO and their influence on the remote tropical climate. *J. Climate*, **15,** 2616-2631, https://doi.org/10.1175/1520-0442(2002)015<2616:TTTVCB>2.0.CO;2.

Chiang, J. C., and D. J. Vimont, 2004: Analogous Pacific and Atlantic meridional modes of tropical atmosphere–ocean variability. *J. Climate*, **17,** 4143-4158.

Chou, C., and J. D. Neelin, 2004: Mechanisms of global warming impacts on regional tropical precipitation. *J. Climate*, **17,** 2688-2701, https://doi.org/10.1175/1520-0442(2004)017<2688:MOGWIO>2.0.CO;2.

Chou, C., J. D. Neelin, and H. Su, 2001: Ocean-atmosphere-land feedbacks in an idealized monsoon. *Q. J. Roy. Meteorol. Soc.*, **127,** 1869-1891.

Chou, C., J. D. Neelin, C.-A. Chen, and J.-Y. Tu, 2009: Evaluating the "rich-get-richer" mechanism in tropical precipitation change under global warming. *J. Climate*, **22,** 1982-2005.

Clement, A., K. Bellomo, L. N. Murphy, M. A. Cane, T. Mauritsen, G. Rädel, and B. Stevens, 2015: The Atlantic Multidecadal Oscillation without a role for ocean circulation. *Science*, **350,** 320-324.

Colombo, D., and M. H. Maathuis, 2014: Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, **15,** 3741-3782.

Cover, T. M., and J. A. Thomas, 1991: Entropy, relative entropy and mutual information. *Elements of information theory*, **2,** 12-13.
Dong, B., and R. Sutton, 2015: Dominant role of greenhouse-gas forcing in the recovery of Sahel rainfall. *Nat. Clim. Change*, **5,** 757-761, https://doi.org/10.1038/nclimate2664.
Donohoe, A., J. Marshall, D. Ferreira, and D. Mcgee, 2013: The relationship between ITCZ location and cross-equatorial atmospheric heat transport: From the seasonal cycle to the Last Glacial Maximum. *J. Climate*, **26,** 3597-3618, https://doi.org/10.1175/JCLI-D-12-00467.1.
Druyan, L. M., and R. D. Koster, 1989: Sources of Sahel precipitation for simulated drought and rainy seasons. *J. Climate*, **2,** 1438-1446.
Du, Y., S.-P. Xie, G. Huang, and K. Hu, 2009: Role of air–sea interaction in the long persistence of El Niño–induced north Indian Ocean warming. *J. Climate*, **22,** 2023-2038.
Eade, R., D. Stephenson, A. Scaife, and D. Smith, 2021: Quantifying the rarity of extreme multi-decadal trends: how unusual was the late twentieth century trend in the North Atlantic Oscillation? *Climate Dynam.*, 1-14, https://doi.org/10.1007/s00382-021-05978-4.
Ebert-Uphoff, I., and Y. Deng, 2012: Causal discovery for climate research using graphical models. *J. Climate*, **25,** 5648-5665.
Edelmann, D., K. Fokianos, and M. Pitsillou, 2019: An updated literature review of distance correlation and its applications to time series. *International Statistical Review*, **87,** 237-262.
Emanuel, K. A., J. David Neelin, and C. S. Bretherton, 1994: On large-scale circulations in convecting atmospheres. *Q. J. Roy. Meteorol. Soc.*, **120,** 1111-1143.
Enfield, D. B., and D. A. Mayer, 1997: Tropical Atlantic sea surface temperature variability and its relation to El Niño-Southern Oscillation. *Journal of Geophysical Research: Oceans*, **102,** 929-945.
Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, **9,** 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016.
Fan, M., and E. K. Schneider, 2012: Observed decadal North Atlantic tripole SST variability. Part I: Weather noise forcing and coupled response. *J. Atmos. Sci.*, **69,** 35-50.
Fan, Y., and H. Van den Dool, 2008: A global monthly land surface air temperature analysis for 1948–present. *J. Geophys. Res.-Atmos.*, **113**.
Fiedler, S., and Coauthors, 2020: Simulated tropical precipitation assessed across three major phases of the coupled model intercomparison project (CMIP). *Mon. Weather Rev.*, **148,** 3653-3680.
Flannaghan, T. J., S. Fueglistaler, I. M. Held, S. Po-Chedley, B. Wyman, and M. Zhao, 2014: Tropical temperature trends in atmospheric general circulation model simulations and the impact of uncertainties in observed SSTs. *J. Geophys. Res.-Atmos.*, **119,** 13,327-313,337.
Folland, C. K., T. N. Palmer, and D. E. Parker, 1986: Sahel rainfall and worldwide sea temperatures, 1901–85. *Nature*, **320,** 602-607, https://doi.org/10.1038/320602a0.
Foltz, G. R., and M. J. McPhaden, 2010: Interaction between the Atlantic meridional and Niño modes. *Geophys. Res. Lett.*, **37**.
Forré, P., and J. M. Mooij, 2018: Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. *arXiv preprint arXiv:1807.03024*.

——, 2020: Causal calculus in the presence of cycles, latent confounders and selection bias. *Uncertainty in Artificial Intelligence*, PMLR, 71-80.
Forster, P. M., A. C. Maycock, C. M. McKenna, and C. J. Smith, 2020: Latest climate models confirm need for urgent mitigation. *Nat. Clim. Change*, **10,** 7-10.
Gaetani, M., B. Fontaine, P. Roucou, and M. Baldi, 2010: Influence of the Mediterranean Sea on the West African monsoon: Intraseasonal variability in numerical simulations. *J. Geophys. Res.-Atmos.*, **115**.
Gaetani, M., and Coauthors, 2017: West African monsoon dynamics and precipitation: the competition between global SST warming and CO 2 increase in CMIP5 idealized simulations. *Climate Dynam.*, **48,** 1353-1373, https://doi.org/10.1007/s00382-016-3146-z.
Gerhardus, A., and J. Runge, 2021: LPCMCI: Causal Discovery in Time Series with Latent Confounders.
Giannini, A., 2010: Mechanisms of climate change in the semiarid African Sahel: The local view. *J. Climate*, **23,** 743-756, https://doi.org/10.1175/2009JCLI3123.1.
Giannini, A., and A. Kaplan, 2019: The role of aerosols and greenhouse gases in Sahel drought and recovery. *Climatic Change*, **152,** 449-466, https://doi.org/10.1007/s10584-018-2341-9.
Giannini, A., R. Saravanan, and P. Chang, 2003: Oceanic forcing of Sahel rainfall on interannual to interdecadal time scales. *Science*, **302,** 1027-1030, https://doi.org/10.1126/science.1089357.
Giannini, A., M. Biasutti, I. M. Held, and A. H. Sobel, 2008: A global perspective on African climate. *Climatic Change*, **90,** 359-383, https://doi.org/10.1007/s10584-008-9396-y.
Giannini, A., S. Salack, T. Lodoun, A. Ali, A. Gaye, and O. Ndiaye, 2013: A unifying view of climate change in the Sahel linking intra-seasonal, interannual and longer time scales. *Environ. Res. Lett.*, **8,** 024010, https://doi.org/10.1088/1748-9326/8/2/024010.
Gillett, N., F. Zwiers, A. Weaver, G. Hegerl, M. Allen, and P. Stott, 2002: Detecting anthropogenic influence with a multi-model ensemble. *Geophys. Res. Lett.*, **29,** 31-31-31-34, https://doi.org/10.1029/2002GL015836.
Gillett, N. P., and Coauthors, 2016: The detection and attribution model intercomparison project (DAMIP v1. 0) contribution to CMIP6. *Geosci. Model Dev.*, **9,** 3685-3697.
Glossary, M., 2009: American Meteorological Society. *ed*.
Granger, C. W., 1969: Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society***,** 424-438.
Haarsma, R. J., F. M. Selten, S. L. Weber, and M. Kliphuis, 2005: Sahel rainfall variability and response to greenhouse warming. *Geophys. Res. Lett.*, **32**, https://doi.org/10.1029/2005GL023232.
Ham, Y.-G., J.-S. Kug, J.-Y. Park, and F.-F. Jin, 2013: Sea surface temperature in the north tropical Atlantic as a trigger for El Niño/Southern Oscillation events. *Nature Geoscience*, **6,** 112-116.
Ham, Y. G., 2017: A reduction in the asymmetry of ENSO amplitude due to global warming: The role of atmospheric feedback. *Geophys. Res. Lett.*, **44,** 8576-8584.
Han, W., J. Vialard, M. J. McPhaden, T. Lee, Y. Masumoto, M. Feng, and W. P. De Ruijter, 2014: Indian Ocean decadal variability: A review. *Bull. Am. Meteorol. Soc.*, **95,** 1679-1703.

Han, Z., F. Luo, S. Li, Y. Gao, T. Furevik, and L. Svendsen, 2016: Simulation by CMIP5 models of the Atlantic multidecadal oscillation and its climate impacts. *Advances in Atmospheric Sciences*, **33,** 1329-1342, https://doi.org/10.1007/s00376-016-5270-4.

Hanawa, K., and S. Sugimoto, 2004: 'Reemergence'areas of winter sea surface temperature anomalies in the world's oceans. *Geophys. Res. Lett.*, **31**.

Harris, I., P. D. Jones, T. J. Osborn, and D. H. Lister, 2014: Updated high-resolution grids of monthly climatic observations–the CRU TS3. 10 Dataset. *Int. J. Climatol.*, **34,** 623-642, https://doi.org/10.1002/joc.3711.

Haywood, J. M., A. Jones, N. Bellouin, and D. Stephenson, 2013: Asymmetric forcing from stratospheric aerosols impacts Sahelian rainfall. *Nat. Clim. Change*, **3,** 660-665, https://doi.org/10.1038/nclimate1857.

Hegerl, G., and F. Zwiers, 2011: Use of models in detection and attribution of climate change. *Wiley interdisciplinary reviews: climate change*, **2,** 570-591, https://doi.org/10.1002/wcc.121.

Held, I. M., and B. J. Soden, 2006: Robust responses of the hydrological cycle to global warming. *J. Climate*, **19,** 5686-5699, https://doi.org/10.1175/JCLI3990.1.

Held, I. M., T. L. Delworth, J. Lu, K. u. Findell, and T. Knutson, 2005: Simulation of Sahel drought in the 20th and 21st centuries. *Proc. Nat. Acad. Sci.*, **102,** 17891-17896, https://doi.org/10.1073/pnas.0509057102.

Hill, S. A., 2016: Energetic and hydrological responses of Hadley circulations and the African Sahel to sea surface temperature perturbations, Princeton University.

Hill, S. A., 2019: Theories for past and future monsoon rainfall changes. *Current Climate Change Reports*, **5,** 160-171, https://doi.org/10.1007/s40641-019-00137-8.

Hill, S. A., Y. Ming, and M. Zhao, 2018: Robust responses of the Sahelian hydrological cycle to global warming. *J. Climate*, **31,** 9793-9814.

Hill, S. A., Y. Ming, I. M. Held, and M. Zhao, 2017: A moist static energy budget–based analysis of the Sahel rainfall response to uniform oceanic warming. *J. Climate*, **30,** 5637-5660, https://doi.org/10.1175/JCLI-D-16-0785.1.

Hirasawa, H., P. J. Kushner, M. Sigmond, J. Fyfe, and C. Deser, 2020: Anthropogenic Aerosols Dominate Forced Multidecadal Sahel Precipitation Change through Distinct Atmospheric and Oceanic Drivers. *J. Climate*, **33,** 10187-10204, https://doi.org/10.1175/JCLI-D-19-0829.1.

——, 2022: Evolving Sahel rainfall response to anthropogenic aerosols driven by shifting regional oceanic and emission influences. *J. Climate*, **35,** 3181-3193.

Hoerling, M., J. Hurrell, J. Eischeid, and A. Phillips, 2006: Detection and attribution of twentieth-century northern and southern African rainfall change. *J. Climate*, **19,** 3989-4008, https://doi.org/10.1175/JCLI3842.1.

Hua, W., A. Dai, L. Zhou, M. Qin, and H. Chen, 2019: An Externally Forced Decadal Rainfall Seesaw Pattern Over the Sahel and Southeast Amazon. *Geophys. Res. Lett.*, **46,** 923-932, https://doi.org/10.1029/2018GL081406.

Huang, B., and Coauthors, 2017: Extended reconstructed sea surface temperature, version 5 (ERSSTv5): upgrades, validations, and intercomparisons. *J. Climate*, **30,** 8179-8205, https://doi.org/10.1175/JCLI-D-16-0836.1.

Huffman, G. J., and Coauthors, 1997: The global precipitation climatology project (GPCP) combined precipitation dataset. *Bull. Am. Meteorol. Soc.*, **78,** 5-20.

Hurrell, J. W., 1995: Decadal trends in the North Atlantic Oscillation: Regional temperatures and precipitation. *Science*, **269,** 676-679.
Hwang, Y.-T., D. M. W. Frierson, and S. M. Kang, 2013a: Anthropogenic sulfate aerosol and the southward shift of tropical precipitation in the late 20th century. *Geophys. Res. Lett.*, **40,** 2845-2850, https://doi.org/10.1002/grl.50502.
Hwang, Y. T., D. M. Frierson, and S. M. Kang, 2013b: Anthropogenic sulfate aerosol and the southward shift of tropical precipitation in the late 20th century. *Geophys. Res. Lett.*, **40,** 2845-2850.
Iles, C. E., and G. C. Hegerl, 2014: The global precipitation response to volcanic eruptions in the CMIP5 models. *Environ. Res. Lett.*, **9**, https://doi.org/10.1088/1748-9326/9/10/104012.
Izrael, Y. A., E. Volodin, S. Kostrykin, A. Revokatova, and A. Ryaboshapko, 2014: The ability of stratospheric climate engineering in stabilizing global mean temperatures and an assessment of possible side effects. *Atmospheric Science Letters*, **15,** 140-148.
Johnson, N. C., and S.-P. Xie, 2010: Changes in the sea surface temperature threshold for tropical convection. *Nature Geoscience*, **3,** 842-845.
Joly, M., and A. Voldoire, 2009: Influence of ENSO on the West African monsoon: temporal aspects and atmospheric processes. *J. Climate*, **22,** 3193-3210.
Joly, M., A. Voldoire, H. Douville, P. Terray, and J.-F. Royer, 2007: African monsoon teleconnections with tropical SSTs: validation and evolution in a set of IPCC4 simulations. *Climate Dynam.*, **29,** 1-20.
Kamae, Y., M. Watanabe, M. Kimoto, and H. Shiogama, 2014: Summertime land–sea thermal contrast and atmospheric circulation over East Asia in a warming climate—Part I: Past changes and future projections. *Climate Dynam.*, **43,** 2553-2568.
Kang, S. M., D. M. Frierson, and I. M. Held, 2009: The tropical response to extratropical thermal forcing in an idealized GCM: The importance of radiative feedbacks and convective parameterization. *J. Atmos. Sci.*, **66,** 2812-2827, https://doi.org/10.1175/2009JAS2924.1.
Kang, S. M., I. M. Held, D. M. Frierson, and M. Zhao, 2008: The response of the ITCZ to extratropical thermal forcing: Idealized slab-ocean experiments with a GCM. *J. Climate*, **21,** 3521-3532, https://doi.org/10.1175/2007JCLI2146.1.
Kawase, H., M. Abe, Y. Yamada, T. Takemura, T. Yokohata, and T. Nozawa, 2010: Physical mechanism of long-term drying trend over tropical North Africa. *Geophys. Res. Lett.*, **37**, https://doi.org/10.1029/2010GL043038.
Keys, P. W., E. Barnes, R. Van Der Ent, and L. J. Gordon, 2014: Variability of moisture recycling using a precipitationshed framework. *Hydrology and Earth System Sciences*, **18,** 3937-3950, https://doi.org/10.5194/hess-18-3937-2014.
Klein, S. A., B. J. Soden, and N.-C. Lau, 1999: Remote sea surface temperature variations during ENSO: Evidence for a tropical atmospheric bridge. *J. Climate*, **12,** 917-932.
Klimont, Z., S. J. Smith, and J. Cofala, 2013: The last decade of global anthropogenic sulfur dioxide: 2000–2011 emissions. *Environ. Res. Lett.*, **8**, https://doi.org/10.1088/1748-9326/8/1/014003.
Knight, J. R., C. K. Folland, and A. A. Scaife, 2006: Climate impacts of the Atlantic multidecadal oscillation. *Geophys. Res. Lett.*, **33**, https://doi.org/10.1029/2006GL026242.
Knight, J. R., R. J. Allan, C. K. Folland, M. Vellinga, and M. E. Mann, 2005: A signature of persistent natural thermohaline circulation cycles in observed climate. *Geophys. Res. Lett.*, **32**, https://doi.org/10.1029/2005GL024233.

Kretschmer, M., J. Runge, and D. Coumou, 2017: Early prediction of extreme stratospheric polar vortex states based on causal precursors. *Geophys. Res. Lett.*, **44,** 8592-8600.
Kretschmer, M., D. Coumou, J. F. Donges, and J. Runge, 2016: Using causal effect networks to analyze different Arctic drivers of midlatitude winter circulation. *J. Climate*, **29,** 4069-4081.
Kretschmer, M., S. V. Adams, A. Arribas, R. Prudden, N. Robinson, E. Saggioro, and T. G. Shepherd, 2021: Quantifying causal pathways of teleconnections. *Bull. Am. Meteorol. Soc.*, **102,** E2247-E2263.
Kucharski, F., N. Zeng, and E. Kalnay, 2013: A further assessment of vegetation feedback on decadal Sahel rainfall variability. *Climate Dynam.*, **40,** 1453-1466, DOI 10.1007/s00382-012-1397-x.
Lélé, M. I., L. M. Leslie, and P. J. Lamb, 2015: Analysis of low-level atmospheric moisture transport associated with the West African monsoon. *J. Climate*, **28,** 4414-4430, https://doi.org/10.1175/JCLI-D-14-00746.1.
Lin, L., Y. Xu, Z. Wang, C. Diao, W. Dong, and S. P. Xie, 2018: Changes in extreme rainfall over India and China attributed to regional aerosol-cloud interaction during the late 20th century rapid industrialization. *Geophys. Res. Lett.*, **45,** 7857-7865, https://doi.org/10.1029/2018GL078308.
Liu, L., and Coauthors, 2018: A PDRMIP multimodel study on the impacts of regional aerosol forcings on global and regional precipitation. *J. Climate*, **31,** 4429-4447.
Liu, Y., J. C. Chiang, C. Chou, and C. M. Patricola, 2014: Atmospheric teleconnection mechanisms of extratropical North Atlantic SST influence on Sahel rainfall. *Climate Dynam.*, **43,** 2797-2811.
Lohmann, U., and J. Feichter, 2005: Global indirect aerosol effects: a review. *Atmos. Chem. Phys.*, **5,** 715-737, https://doi.org/10.5194/acp-5-715-2005.
Losada, T., B. Rodriguez-Fonseca, E. Mohino, J. Bader, S. Janicot, and C. R. Mechoso, 2012: Tropical SST and Sahel rainfall: A non-stationary relationship. *Geophys. Res. Lett.*, **39**, https://doi.org/10.1029/2012GL052423.
Losada, T., B. Rodríguez-Fonseca, I. Polo, S. Janicot, S. Gervois, F. Chauvin, and P. Ruti, 2010: Tropical response to the Atlantic Equatorial mode: AGCM multimodel approach. *Climate Dynam.*, **35,** 45-52.
Ma, J., and S.-P. Xie, 2013: Regional patterns of sea surface temperature change: A source of uncertainty in future projections of precipitation and atmospheric circulation. *J. Climate*, **26,** 2482-2501.
Mantua, N. J., S. R. Hare, Y. Zhang, J. M. Wallace, and R. C. Francis, 1997: A Pacific interdecadal climate oscillation with impacts on salmon production. *Bull. Am. Meteorol. Soc.*, **78,** 1069-1080.
Mariotti, A., and A. Dell'Aquila, 2012: Decadal climate variability in the Mediterranean region: roles of large-scale forcings and regional processes. *Climate Dynam.*, **38,** 1129-1145.
Martin, E. R., C. Thorncroft, and B. B. Booth, 2014: The multidecadal Atlantic SST—Sahel rainfall teleconnection in CMIP5 simulations. *J. Climate*, **27,** 784-806, https://doi.org/10.1175/JCLI-D-13-00242.1.
Martín-Rey, M., I. Polo, B. Rodríguez-Fonseca, T. Losada, and A. Lazar, 2018: Is there evidence of changes in tropical Atlantic variability modes under AMO phases in the observational record? *J. Climate*, **31,** 515-536.

Marvel, K., M. Biasutti, and C. Bonfils, 2020: Fingerprints of external forcings on Sahel rainfall: aerosols, greenhouse gases, and model-observation discrepancies. *Environ. Res. Lett.*, **15**, https://doi.org/10.1088/1748-9326/ab858e.

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

McCoy, D. T., F. A.-M. Bender, J. K. C. Mohrmann, D. L. Hartmann, R. Wood, and D. P. Grosvenor, 2017: The global aerosol-cloud first indirect effect estimated using MODIS, MERRA, and AeroCom *J. Geophys. Res.-Atmos.*, **122,** 1779-1796, https://doi.org/10.1002/2016JD026141.

Mechoso, C. R., C. Wang, J. F. Lübbecke, B. Rodríguez-Fonseca, and M. Diakhate, 2023: Interactions among climates of ocean basins. *Frontiers in Climate*, **5,** Art. Nr. 1138642.

Meehl, G., and Coauthors, 2007: THE WCRP CMIP3 Multimodel Dataset: A New Era in Climate Change Research. *Bull. Am. Meteorol. Soc.*, **88,** 1383-1394, https://doi.org/10.1175/BAMS-88-9-1383.

Meehl, G. A., W. M. Washington, T. Wigley, J. M. Arblaster, and A. Dai, 2003: Solar and greenhouse gas forcing and climate response in the twentieth century. *J. Climate*, **16,** 426-444, https://doi.org/10.1175/1520-0442(2003)016<0426:SAGGFA>2.0.CO;2.

Menary, M. B., and Coauthors, 2020: Aerosol-forced AMOC changes in CMIP6 historical simulations. *Geophys. Res. Lett.*, **47**, https://doi.org/10.1029/2020GL088166.

Mote, P., L. Brekke, P. B. Duffy, and E. Maurer, 2011: Guidelines for constructing climate scenarios. *Eos, Transactions American Geophysical Union*, **92,** 257-258.

Murphy, L. N., K. Bellomo, M. Cane, and A. Clement, 2017: The role of historical forcings in simulating the observed Atlantic multidecadal oscillation. *Geophys. Res. Lett.*, **44,** 2472-2480, https://doi.org/10.1002/2016GL071337.

Mutton, H., R. Chadwick, M. Collins, F. H. Lambert, R. Geen, A. Todd, and C. M. Taylor, 2022: The Impact of the Direct Radiative Effect of Increased CO 2 on the West African Monsoon. *J. Climate*, **35,** 2441-2458, https://doi.org/10.1175/JCLI-D-21-0340.1.

Naylor, A. W., and J. Ford, 2023: Vulnerability and loss and damage following the COP27 of the UN framework convention on climate change. *Regional Environmental Change*, **23**.

Neelin, J., C. Chou, and H. Su, 2003: Tropical drought regions in global warming and El Nino teleconnections. *Geophys. Res. Lett.*, **30**, https://doi.org/10.1029/2003GL018625.

Neupane, N., and K. H. Cook, 2013: A nonlinear response of Sahel rainfall to Atlantic warming. *J. Climate*, **26,** 7080-7096, https://doi.org/10.1175/JCLI-D-12-00475.1.

Nicholson, S. E., 2009: A revised picture of the structure of the "monsoon" and land ITCZ over West Africa. *Climate Dynam.*, **32,** 1155-1171, https://doi.org/10.1007/s00382-008-0514-3.

——, 2013: The West African Sahel: A review of recent studies on the rainfall regime and its interannual variability. *ISRN Meteorology*, **2013**, https://doi.org/10.1155/2013/453521.

Nobre, P., and J. Shukla, 1996: Variations of sea surface temperature, wind stress, and rainfall over the tropical Atlantic and South America. *J. Climate*, **9,** 2464-2479.

Nowack, P., J. Runge, V. Eyring, and J. D. Haigh, 2020: Causal networks for climate model evaluation and constrained projections. *Nature communications*, **11,** 1415.

Okonkwo, C., and Coauthors, 2015: Combined effect of El Niño southern oscillation and Atlantic multidecadal oscillation on Lake Chad level variability. *Cogent Geoscience*, **1**, https://doi.org/10.1080/23312041.2015.1117829.

Okumura, Y., and S.-P. Xie, 2004: Interaction of the Atlantic equatorial cold tongue and the African monsoon. *J. Climate*, **17,** 3589-3602.
Okumura, Y. M., and C. Deser, 2010: Asymmetry in the duration of El Niño and La Niña. *J. Climate*, **23,** 5826-5843.
Palmer, T., 1986: Influence of the Atlantic, Pacific and Indian oceans on Sahel rainfall. *Nature*, **322,** 251-253, https://doi.org/10.1038/322251a0.
Parhi, P., A. Giannini, P. Gentine, and U. Lall, 2016: Resolving contrasting regional rainfall responses to El Niño over tropical Africa. *J. Climate*, **29,** 1461-1476, https://doi.org/10.1175/JCLI-D-15-0071.1.
Park, J.-Y., J. Bader, and D. Matei, 2015: Northern-hemispheric differential warming is the key to understanding the discrepancies in the projected Sahel rainfall. *Nature communications*, **6,** 1-8, https://doi.org/10.1038/ncomms6985.
——, 2016: Anthropogenic Mediterranean warming essential driver for present and future Sahel rainfall. *Nat. Clim. Change*, **6,** 941-945, https://doi.org/10.1038/nclimate3065.
Pearl, J., 2009: *Causality.* Cambridge university press.
——, 2022: Direct and indirect effects. *Probabilistic and Causal Inference: The Works of Judea Pearl*, 373-392.
Pearl, J., M. Glymour, and N. P. Jewell, 2016: *Causal inference in statistics: A primer.* John Wiley & Sons, 136 pp.
Penner, J. E., and Coauthors, 2006: Model intercomparison of indirect aerosol effects. *Atmos. Chem. Phys.*, **6,** 3391-3405, https://doi.org/10.5194/acp-6-3391-2006.
Peyrillé, P., J.-P. Lafore, and J.-L. Redelsperger, 2007: An idealized two-dimensional framework to study the West African monsoon. Part I: Validation and key controlling factors. *J. Atmos. Sci.*, **64,** 2765-2782.
Polson, D., M. Bollasina, G. Hegerl, and L. Wilcox, 2014: Decreased monsoon precipitation in the Northern Hemisphere due to anthropogenic aerosols. *Geophys. Res. Lett.*, **41,** 6023-6029, https://doi.org/10.1002/2014GL060811.
Pomposi, C., Y. Kushnir, and A. Giannini, 2015: Moisture budget analysis of SST-driven decadal Sahel precipitation variability in the twentieth century. *Climate Dynam.*, **44,** 3303-3321, https://doi.org/10.1007/s00382-014-2382-3.
Pomposi, C., A. Giannini, Y. Kushnir, and D. E. Lee, 2016: Understanding Pacific Ocean influence on interannual precipitation variability in the Sahel. *Geophys. Res. Lett.*, **43,** 9234-9242, https://doi.org/10.1002/2016GL069980.
Pu, B., and K. H. Cook, 2010: Dynamics of the West African westerly jet. *J. Climate*, **23,** 6263-6276.
——, 2012: Role of the West African westerly jet in Sahel rainfall variations. *J. Climate*, **25,** 2880-2896.
Qin, M., A. Dai, and W. Hua, 2020: Quantifying contributions of internal variability and external forcing to Atlantic multidecadal variability since 1870. *Geophys. Res. Lett.*, **47**, https://doi.org/10.1029/2020GL089504.
Rahmstorf, S., J. E. Box, G. Feulner, M. E. Mann, A. Robinson, S. Rutherford, and E. J. Schaffernicht, 2015: Exceptional twentieth-century slowdown in Atlantic Ocean overturning circulation. *Nat. Clim. Change*, **5,** 475-480, https://doi.org/10.1038/nclimate2554.
Ramsey, J., J. Zhang, and P. L. Spirtes, 2012: Adjacency-faithfulness and conservative causal inference. *arXiv preprint arXiv:1206.6843*.

Reisach, A., C. Seiler, and S. Weichwald, 2021: Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, **34,** 27772-27784.

Richter, I., and H. Tokinaga, 2020: An overview of the performance of CMIP6 models in the tropical Atlantic: mean state, variability, and remote impacts. *Climate Dynam.*, **55,** 2579-2601.

Robock, A., and Y. Liu, 1994: The volcanic signal in Goddard Institute for Space Studies three-dimensional model simulations. *J. Climate*, **7,** 44-55, https://doi.org/10.1175/1520-0442(1994)007<0044:TVSIGI>2.0.CO;2.

Rodríguez-Fonseca, B., I. Polo Sanchez, J. Garcia Serrano, T. Losada Doval, E. Mohino, R. Mechoso, and F. Kucharski, 2009: Have Atlantic Niños been leading Pacific ENSO events in recent decades? *EGU General Assembly Conference Abstracts*, 9476.

Rodríguez-Fonseca, B., and Coauthors, 2015: Variability and predictability of West African droughts: A review on the role of sea surface temperature anomalies. *J. Climate*, **28,** 4034-4060, https://doi.org/10.1175/JCLI-D-14-00130.1.

Rodríguez-Fonseca, B., and Coauthors, 2011: Interannual and decadal SST-forced responses of the West African monsoon. *Atmospheric Science Letters*, **12,** 67-74.

Rosenfeld, D., and Coauthors, 2008: Flood or drought: How do aerosols affect precipitation? *Science*, **321,** 1309-1313, https://doi.org/10.1126/science.1160606.

Rotstayn, L. D., and U. Lohmann, 2002: Tropical rainfall trends and the indirect aerosol effect. *J. Climate*, **15,** 2103-2116, https://doi.org/10.1175/1520-0442(2002)015<2103:TRTATI>2.0.CO;2.

Rowell, D. P., 2003: The impact of Mediterranean SSTs on the Sahelian rainfall season. *J. Climate*, **16,** 849-862.

Rowell, D. P., C. K. Folland, K. Maskell, and M. N. Ward, 1995: Variability of summer rainfall over tropical north Africa (1906–92): observations and modelling. *Q. J. Roy. Meteorol. Soc.*, **121,** 669-704, https://doi.org/10.1002/qj.49712152311.

Rubenstein, P. K., S. Weichwald, S. Bongers, J. M. Mooij, D. Janzing, M. Grosse-Wentrup, and B. Schölkopf, 2017: Causal consistency of structural equation models. *arXiv preprint arXiv:1707.00819*.

Ruiz-Barradas, A., J. A. Carton, and S. Nigam, 2000: Structure of interannual-to-decadal climate variability in the tropical Atlantic sector. *J. Climate*, **13,** 3285-3297.

Runge, J., 2018a: Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, **28,** 075310.

——, 2018b: Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. *International Conference on Artificial Intelligence and Statistics*, PMLR, 938-947.

——, 2020: Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. *Conference on Uncertainty in Artificial Intelligence*, PMLR, 1388-1397.

Runge, J., V. Petoukhov, and J. Kurths, 2014: Quantifying the strength and delay of climatic interactions: The ambiguities of cross correlation and a novel measure based on graphical models. *J. Climate*, **27,** 720-739.

Runge, J., P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, 2019a: Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, **5,** eaau4996.

Runge, J., A. Gerhardus, G. Varando, V. Eyring, and G. Camps-Valls, In Review: Causal inference for time series. *Nature*.

Runge, J., and Coauthors, 2019b: Inferring causation from time series in Earth system sciences. *Nature communications*, **10,** 2553.

Scaife, A., and Coauthors, 2009: The CLIVAR C20C project: selected twentieth century climate events. *Climate Dynam.*, **33,** 603-614, https://doi.org/10.1007/s00382-008-0451-1.

Schneider, T., T. Bischoff, and G. H. Haug, 2014: Migrations and dynamics of the intertropical convergence zone. *Nature*, **513,** 45-53, https://doi.org/10.1038/nature13636.

Seth, A., S. A. Rauscher, M. Biasutti, A. Giannini, S. J. Camargo, and M. Rojas, 2013: CMIP5 projected changes in the annual cycle of precipitation in monsoon regions. *J. Climate*, **26,** 7328-7351.

Shekhar, R., and W. R. Boos, 2016: Improving energy-based estimates of monsoon location in the presence of proximal deserts. *J. Climate*, **29,** 4741-4761, https://doi.org/10.1175/JCLI-D-15-0747.1.

Shepherd, T. G., 2019: Storyline approach to the construction of regional climate change information. *Proceedings of the Royal Society A*, **475,** 20190013, https://doi.org/10.1098/rspa.2019.0013.

Skliris, N., S. Sofianos, A. Gkanasos, A. Mantziafou, V. Vervatis, P. Axaopoulos, and A. Lascaratos, 2012: Decadal scale variability of sea surface temperature in the Mediterranean Sea in relation to atmospheric variability. *Ocean Dynamics*, **62,** 13-30.

Smith, S. J., J. v. Aardenne, Z. Klimont, R. J. Andres, A. Volke, and S. Delgado Arias, 2011: Anthropogenic sulfur dioxide emissions: 1850–2005. *Atmos. Chem. Phys.*, **11,** 1101-1116, https://doi.org/10.5194/acp-11-1101-2011.

Sobel, A., 2010: Raised bar for rain. *Nature Geoscience*, **3,** 821-822.

Sobel, A. H., 2007: Simple models of ensemble-averaged precipitation and surface wind, given the sea surface temperature. Princeton University Press, 219-251.

Sobel, A. H., J. Nilsson, and L. M. Polvani, 2001: The weak temperature gradient approximation and balanced tropical moisture waves. *J. Atmos. Sci.*, **58,** 3650-3665.

Sobel, A. H., I. M. Held, and C. S. Bretherton, 2002: The ENSO signal in tropical tropospheric temperature. *J. Climate*, **15,** 2702-2706, https://doi.org/10.1175/1520-0442(2002)015<2702:TESITT>2.0.CO;2.

Spirtes, P., and C. Glymour, 1991: An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, **9,** 62-72.

Spirtes, P., C. N. Glymour, R. Scheines, and D. Heckerman, 2000: *Causation, prediction, and search.* MIT press.

Stevens, B., and G. Feingold, 2009: Untangling aerosol effects on clouds and precipitation in a buffered system. *Nature*, **461,** 607, https://doi.org/10.1038/nature08281.

Storelvmo, T., W. Boos, and N. Herger, 2014: Cirrus cloud seeding: a climate engineering mechanism with reduced side effects? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **372,** 20140116.

Sutton, R. T., and D. L. Hodson, 2005: Atlantic Ocean forcing of North American and European summer climate. *Science*, **309,** 115-118, https://doi.org/10.1126/science.1109496.

Székely, G. J., M. L. Rizzo, and N. K. Bakirov, 2007: Measuring and testing dependence by correlation of distances.

Taylor, C. M., E. F. Lambin, N. Stephenne, R. J. Harding, and R. L. H. Essery, 2002a: The influence of land use change on climate in the Sahel. *J. Climate*, **15,** 3615-3629, 10.1175/1520-0442(2002)015<3615:TIOLUC>2.0.CO;2.

Taylor, C. M., E. F. Lambin, N. Stephenne, R. J. Harding, and R. L. Essery, 2002b: The influence of land use change on climate in the Sahel. *J. Climate*, **15,** 3615-3629.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.*, **93,** 485-498, https://doi.org/10.1175/BAMS-D-11-00094.1.

Thompson, D. W., and J. M. Wallace, 2000: Annular modes in the extratropical circulation. Part I: Month-to-month variability. *J. Climate*, **13,** 1000-1016.

Thorncroft, C., and K. Hodges, 2001: African easterly wave variability and its relationship to Atlantic tropical cyclone activity. *J. Climate*, **14,** 1166-1179.

Thorncroft, C. D., H. Nguyen, C. Zhang, and P. Peyrillé, 2011: Annual cycle of the West African monsoon: regional circulations and associated water vapour transport. *Q. J. Roy. Meteorol. Soc.*, **137,** 129-147.

Ting, M., Y. Kushnir, R. Seager, and C. Li, 2009: Forced and internal twentieth-century SST trends in the North Atlantic. *J. Climate*, **22,** 1469-1481, https://doi.org/10.1175/2008JCLI2561.1.

Uhler, C., G. Raskutti, P. Bühlmann, and B. Yu, 2013: Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics***,** 436-463.

Ummenhofer, C. C., A. Biastoch, and C. W. Böning, 2017: Multidecadal Indian Ocean variability linked to the Pacific and implications for preconditioning Indian Ocean dipole events. *J. Climate*, **30,** 1739-1751.

Undorf, S., D. Polson, M. Bollasina, Y. Ming, A. Schurer, and G. Hegerl, 2018: Detectable impact of local and remote anthropogenic aerosols on the 20th century changes of West African and South Asian monsoon precipitation. *J. Geophys. Res.-Atmos.*, **123,** 4871-4889, https://doi.org/10.1029/2017JD027711.

Uppala, S. M., and Coauthors, 2005: The ERA-40 re-analysis. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, **131,** 2961-3012.

Vellinga, M., and Coauthors, 2016: Sahel decadal rainfall variability and the role of model horizontal resolution. *Geophys. Res. Lett.*, **43,** 326-333, https://doi.org/10.1002/2015GL066690.

Vignaroli, P., 2017: Building resilience to drought in the sahel by early risk identification and advices. *Renewing Local Planning to Face Climate Change in the Tropics***,** 151-167.

Villamayor, J., and E. Mohino, 2015: Robust Sahel drought due to the Interdecadal Pacific Oscillation in CMIP5 simulations. *Geophys. Res. Lett.*, **42,** 1214-1222.

Wang, C., L. Zhang, S.-K. Lee, L. Wu, and C. R. Mechoso, 2014: A global perspective on CMIP5 climate model biases. *Nat. Clim. Change*, **4,** 201-205.

Wang, C., C. Deser, J.-Y. Yu, P. DiNezio, and A. Clement, 2017: El Niño and southern oscillation (ENSO): a review. *Coral reefs of the eastern tropical Pacific: Persistence and loss in a dynamic environment***,** 85-106.

Wang, Y., J. H. Jiang, and H. Su, 2015: Atmospheric responses to the redistribution of anthropogenic aerosols. *J. Geophys. Res.-Atmos.*, **120,** 9625-9641, https://doi.org/10.1002/2015JD023665.
Webb, M. J., and Coauthors, 2017: The cloud feedback model intercomparison project (CFMIP) contribution to CMIP6. *Geosci. Model Dev.*, **10,** 359-384.
Williams, C. K., and C. E. Rasmussen, 2006: *Gaussian processes for machine learning.* Vol. 2, MIT press Cambridge, MA.
Yan, X., and Y. Tang, 2021: Multidecadal variability in Mediterranean Sea surface temperature and its sources. *Geophys. Res. Lett.*, **48,** e2020GL091814.
Yan, X., R. Zhang, and T. R. Knutson, 2018: Underestimated AMOC Variability and Implications for AMV and Predictability in CMIP Models. *Geophys. Res. Lett.*, **45,** 4319-4328, https://doi.org/10.1029/2018GL077378.
——, 2019: A multivariate AMV index and associated discrepancies between observed and CMIP5 externally forced AMV. *Geophys. Res. Lett.*, **46,** 4421-4431, https://doi.org/10.1029/2019GL082787.
Yano, J. I., and M. H. Ambaum, 2017: Moist static energy: Definition, reference constants, a conservation law and effects on buoyancy. *Q. J. Roy. Meteorol. Soc.*, **143,** 2727-2734.
Yu, J.-Y., H.-Y. Kao, and T. Lee, 2010: Subtropics-related interannual sea surface temperature variability in the central equatorial Pacific. *J. Climate*, **23,** 2869-2884.
Yu, J.-Y., P.-k. Kao, H. Paek, H.-H. Hsu, C.-w. Hung, M.-M. Lu, and S.-I. An, 2015: Linking emergence of the central Pacific El Niño to the Atlantic multidecadal oscillation. *J. Climate*, **28,** 651-662.
Zelinka, M. D., and Coauthors, 2020: Causes of higher climate sensitivity in CMIP6 models. *Geophys. Res. Lett.*, **47,** e2019GL085782.
Zeng, N., 2003: Drought in the Sahel. *Science*, **302,** 999-1000.
Zhang, H., A. Clement, and P. Di Nezio, 2014: The South Pacific meridional mode: A mechanism for ENSO-like variability. *J. Climate*, **27,** 769-783.
Zhang, J., 2008: On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, **172,** 1873-1896.
Zhang, R., 2017: On the persistence and coherence of subpolar sea surface temperature and salinity anomalies associated with the Atlantic multidecadal variability. *Geophys. Res. Lett.*, **44,** 7865-7875, https://doi.org/10.1002/2017GL074342.
Zhang, R., and T. L. Delworth, 2006: Impact of Atlantic multidecadal oscillations on India/Sahel rainfall and Atlantic hurricanes. *Geophys. Res. Lett.*, **33**, https://doi.org/10.1029/2006GL026267.
Zhang, R., R. Sutton, G. Danabasoglu, T. L. Delworth, W. M. Kim, J. Robson, and S. G. Yeager, 2016: Comment on "The Atlantic Multidecadal Oscillation without a role for ocean circulation". *Science*, **352,** 1527-1527, https://doi.org/10.1126/science.aaf1660.
Zhang, R., and Coauthors, 2013: Have aerosols caused the observed Atlantic multidecadal variability? *J. Atmos. Sci.*, **70,** 1135-1144, https://doi.org/10.1175/JAS-D-12-0331.1.
Zhang, W., F. Jiang, M. F. Stuecker, F.-F. Jin, and A. Timmermann, 2021: Spurious north tropical Atlantic precursors to El Niño. *Nature Communications*, **12,** 3096.
Zhang, X., and Coauthors, 2007: Detection of human influence on twentieth-century precipitation trends. *Nature*, **448,** 461, https://doi.org/10.1038/nature06025.
Zhang, Y., and S. Fueglistaler, 2020: How tropical convection couples high moist static energy over land and ocean. *Geophys. Res. Lett.*, **47,** e2019GL086387.

Zhao, Y., and S. Nigam, 2015: The Indian Ocean dipole: A monopole in SST. *J. Climate*, **28,** 3-19.

Zhou, T., and Coauthors, 2016: GMMIP (v1. 0) contribution to CMIP6: global monsoons model inter-comparison project. *Geosci. Model Dev.*, **9,** 3589-3604.

Appendix A. Supplement for Chapter 2: The Effects of Anthropogenic and Volcanic Aerosols and Greenhouse Gases on Twentieth Century Sahel Precipitation in CMIP5

Fig. A.1 shows the scaled power spectra (PS) from Figure 2.5(c) and (d) in the style of panels (a) and (b), where the PS for each model (averaged over the runs for that model, and in the case of the piC, over different sections of the long piC run) are represented separately, colored by the rainfall bias of that model's ALL runs relative to observations. While the correction seems to completely get rid of the stratification by total rainfall bias at medium and low frequency in the ALL simulations, it seems to overcorrect the power in the simulations of the driest models at high frequency in the ALL simulations, and at all frequencies in the piC simulations. This is perhaps not surprising, as when a model is particularly dry, normal variability may make up a larger fraction of the total rainfall. As this correction is imperfect, we do not use it in the calculation of the MMM; rather, only to facilitate comparison of the models in Figure 2.5.

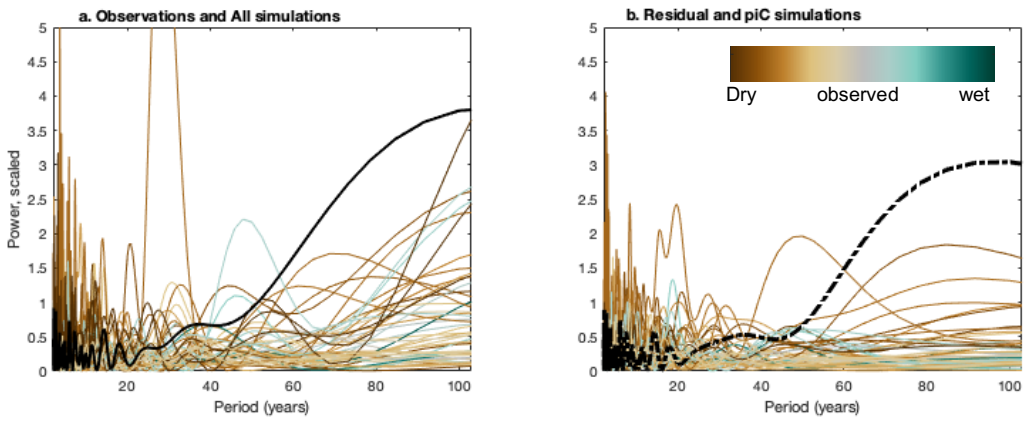

Fig. A.1: Scaled Stratification: Same as Figure 2.5(c) and (d), but displayed as in panels (a) and (b). Power spectra (PS) of observed 20th century Sahel rainfall (solid black, a) and the residual after removing the ALL MMM (black dotted-dashed, b), and mean PS by model of individual ALL (a) and piC (b) runs which were first rescaled by model so their corresponding ALL runs match 20th century observed JAS rainfall, colored by original simulated average JAS rainfall bias of the ALL runs compared to 20th century observations, where observed rainfall is grey, wet models are turquoise, and dry models are brown. piC PS are averaged over multiple segments of the simulations.

Tbl. A.1 displays the models and runs used in this study, as well as their institution classifications.

Tbl. A.1: Models and runs used in this chapter for the different forcing experiments. "p" is the physics number – different physics numbers within the same model are treated as different models. Blank spaces exist in the chart where there were no runs from that model under that forcing experiment. *no accompanying piC run. Doubled lines divide different research institutions.

	ALL				AA				GHG				NAT	
Models	p	Num runs used	runs excluded	reason	p	Num runs used	runs excluded	reason	p	Num runs used	runs excluded	reason	p	Num runs used
ACCESS1-0	1	1												
ACCESS1-3	1	1							1	1			1	3
bcc-csm1-1	1	3							1	1			1	1
bcc-csm1-1-m	1	3												
BNU-ESM	1	1							1	1			1	1
CanCM4*	1		all	no data before 1961										
CanESM2	1	5			4	5			1	5			1	5
CCSM4	1	6			10	3			1	3			1	4
					14									
CESM1-BGC	1	1				2	r6i1p14	access error						
CESM1-CAM5	1	3			10	3			1	1	r1i1p1, r2i1p1	contain NaN	1	3
CESM1-CAM5-1-FV2*	1	4												
CESM1-FASTCHEM	1	3												
CESM1-WACCM	1	1	r4i1p1, r3i1p1, r2i1p1	no data before 1955										
CMCC-CESM	1	1												
CMCC-CM	1	1												
CMCC-CMS	1	1												
CNRM-CM5	1	10							1	6			1	6
CNRM-CM5-2	1	1												
CSIRO-Mk3-6-0	1	10			4	5			1	5			1	5
EC-EARTH	1	1												
FGOALS-g2	1	4	r2i1p1	no data before 1902	1	1			1	1			1	3
FGOALS-s2	1	3												
FIO-ESM	1	3												
GFDL-CM3	1	5			1	3			1	3			1	3
GFDL-ESM2G	1	3												
GFDL-ESM2M	1	1			5	1			1	1			1	1
GISS-E2-H	1	6			107	5			1	5			3	5
	2	5			310	5							1	5
GISS-E2-H-CC	1	1												
GISS-E2-R	1	6			107	5			1	5			3	5
	2	5			310	5							1	5
	3	5												
GISS-E2-R-CC	1	1												
HadCM3*	1	10												
HadGEM2-AO	1	1												
HadGEM2-CC	1	1	r3i1p1, r2i1p1	no data before 1960										
HadGEM2-ES	1	4							1	4			1	4
inmcm4	1	1												
IPSL-CM5A-LR	1	6			3	1			1	3			1	3
IPSL-CM5A-MR	1	3							2	3				
IPSL-CM5B-LR	1	1												
MIROC-ESM	1	3							1	3			1	3
MIROC-ESM-CHEM	1	1							1	1			1	1
MIROC4h	1		all	no data before 1950										
MIROC5	1	5												
MPI-ESM-LR	1	3												
MPI-ESM-MR	1	3												
MPI-ESM-P	1	2												
MRI-CGCM3	1	3							1	1			1	1
	2	2												
MRI-ESM1*	1	1												
NorESM1-M	1	3			1	1			1	1			1	1
NorESM1-ME	1	1												
Total Models used	51				14				21				22	

Appendix B. Supplement for Chapter 3: Deficiencies in Simulated Low-Frequency Sahel Precipitation Variability from CMIP5 and CMIP6

The following tables enumerate the simulations used in this paper.

Tbl. B.1: CMIP6 AMIP (atmosphere-only) simulations used in this chapter.

		Number of Runs Used	
Institutions	Models	amip-hist	amip-piF
CCCma	CanESM5p2	10	3
CNRM-CERFACS	CNRM-CM6-1p1	10	1
	CNRM-CM6-1-HRp1	1	
	CNRM-ESM2-1p1	1	
IPSL	IPSL-CM6A-LRp1	20	1
MRI	MRI-ESM2-0p1	5	1
NCAR	CESM2p1	3	1
	Total Runs Used	50	7
	Total Models Used	7	5
	Total Institutions Used	5	5

Tbl. B.2: Fully coupled CMIP6 simulations used in this chapter. Where different for precipitation and SST, the two are presented in that order separated by a slash. *piC simulations extended past 100 years by repeating the first 14 values. MIROC-ES2H are excluded for only containing one year (1850). NCC_NorESM2-LM r1i1p1f1 excluded for precipitation because it begins in 1950.

		Number of Runs Used			
Institution	Model	ALL	AA	NAT	GHG
BCC	BCC-CSM2-MR p1	3	3	3	3
	BCC-ESM1 p1	3			
CCCma	CanESM5 p1	25	15	25/10	25
	CanESM5 p2	40	15	25/0	25
	CanESM5-CanOE p2	3			
CNRM-CERFACS	CNRM-CM6-1 p1	29	10	10	10
	CNRM-CM6-1-HR p1	1			
	CNRM-ESM2-1 p1	10			
IPSL	IPSL-CM5A2-INCA p1	1/0			
	IPSL-CM6A-LR p1	32	10	10	10
	IPSL-CM6A-LR-INCA p1	1			
MIROC	MIROC-ES2L p1	31			
	MIROC6 p1	50	3	50/0	3
MOHC	HadGEM3-GC31-LL p1	5	4	4	4
	HadGEM3-GC31-MM p1	4			
	UKESM1-0-LL p1	16			
MRI	MRI-ESM2-0 p1	12	3	5/3	3
NASA-GISS	GISS-E2-1-G p1*	28	5	25/5	6
	GISS-E2-1-G p3*	9			
	GISS-E2-1-G p5	9			
	GISS-E2-1-G-CC p1	1			
	GISS-E2-1-H p1	15			
	GISS-E2-1-H p3	5			
	GISS-E2-1-H p5	5			
	GISS-E2-2-H p1	5			
NCAR	CESM2 p1	11	2	3	3
	CESM2-FV2 p1	3			
	CESM2-WACCM p1	3			
	CESM2-WACCM-FV2 p1	3			
NCC	NorCPM1 p1	30			
	NorESM2-LM p1	2/3	3/1	3/0	1
	NorESM2-MM p1	3			
NOAA-GFDL	GFDL-CM4 p1	1			
	GFDL-ESM4 p1	3	1	3	1
	Total Runs Used	402	74/2	169/51	93/95
	Total Models Used	34/3	12	13/9	12
	Total Institutions Used	11	11	11/9	11

Tbl. B.3: Fully-coupled CMIP5 simulations used in this chapter for precipitation. When the simulations for SST differ, they are presented after a slash. CESM1-CAM5-1-FV2 p1 are excluded from SST data because there was no appropriate mask available in the fixed data. Two of the CESM1-CAM5 p1 simulations contain a couple NaN values around 1960. Precipitation from CCSM4 r6i1p14 is excluded because of a downloading error.

Institution	Model	Number of Runs Used			
		ALL	AA	NAT	GHG
CAS	FGOALS-g2 p1	5	1	3	1
	FGOALS-s2 p1	3			
CCCma	CanESM2 p1	5		5	5
	CanESM2 p4		5		
CSIRO	CSIRO-Mk3-6-0 p1	10		5	5
	CSIRO-Mk3-6-0 p4		5		
IPSL	IPSL-CM5A-LR p1	6/5		3	3 (0)
	IPSL-CM5A-LR p2				2 (5)
	IPSL-CM5A-LR p3		1		
	IPSL-CM5A-MR p1	3		3	
	IPSL-CM5A-MR p2				3
	IPSL-CM5B-LR p1	1			
NASA-GISS	GISS-E2-H p1	6		5	5
	GISS-E2-H p107		5		
	GISS-E2-H p2	5			
	GISS-E2-H p3			5	
	GISS-E2-H p310		5		
	GISS-E2-H-CC p1	1			
	GISS-E2-R p1	6		5	5
	GISS-E2-R p107		5		
	GISS-E2-R p2	5			
	GISS-E2-R p3	5		5	
	GISS-E2-R p310		5		
	GISS-E2-R-CC p1	1			
NCAR	CCSM4 p1	6		4	3
	CCSM4 p10		3		
	CCSM4 p14		2/0		
	CESM1-BGC p1	1			
	CESM1-CAM5 p1	3		3/0	3
	CESM1-CAM5 p10		3		
	CESM1-CAM5-1-FV2 p1	4/0			
	CESM1-FASTCHEM p1	3			
	CESM1-WACCM p1	1			
	CMCC-CESM p1	1			
NCC	NorESM1-M p1	3	1	1	1
	NorESM1-ME p1	1			
NOAA-GFDL	GFDL-CM3 p1	5	3	3	3
	GFDL-CM2 p1	0/10			
	GFDL-ESM2G p1	3			
	GFDL-ESM2M p1	1		1	1
	GFDL-ESM2M p5		1		
	Total Runs Used	94/99	45/43	51/48	40
	Total Models Used	26	14/13	14/13	13
	Total Institutions Used	8	8	8	8