Using palaeoclimate information to improve stochastic modelling for water management

Matthew Armstrong

A thesis submitted for the degree of Doctor of Philosophy at The University of Newcastle

September - 2023

This research has been conducted with the support of the Australian Government Research Training Program Scholarship.

Statement of originality

I hereby certify that the work embodied in the thesis is my own work, conducted under normal supervision. The thesis contains no material which has been accepted, or is being examined, for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made. I give consent to the final version of my thesis being made available worldwide when deposited in the University's Digital Repository, subject to the provisions of the Copyright Act 1968 and any approved embargo.

Matthew Armstrong
26/09/2023

Abstract

Various palaeoclimate reconstructions have identified the occurrence of 'megadroughts'. These 'megadroughts' are much longer and more severe than those recorded via the instrumental measurements. Because instrumental measurements are used to infer climate risk when designing water supply infrastructure and management plans, a 'megadrought' invokes a concern for water security. However, such concerns should be viewed within the context of the (a) limitations of using palaeoclimate proxy records as a source of climate information and (b) existing methods used in water management to estimate climate risk (i.e. inferring climate risk using a stochastic model calibrated to instrumental measurements).

The goal of this thesis is to (a) use palaeoclimate proxy records to evaluate stochastic model performance and parameter stationarity under long-term, centennial-scale variability and (b) present a stochastic modelling framework that incorporates proxy centennial-scale variability.

To achieve this goal, this thesis had five objectives:

1. Evaluate the persistence signal in Antarctic ice core records.

It was found that the persistence signal in annual snowfall accumulation and mid-latitude rainfall records are statistically similar. In contrast, ice core Na+ records tended to have slightly higher persistence than mid-latitude rainfall. This analysis informed the subsequent use of ice core information in palaeoclimate-informed stochastic modelling and climate risk assessment.

2. Evaluate different stochastic models using millennium-length palaeoclimate proxy records.

It was found that a stochastic model calibrated to instrumental measurements cannot simulate long-term, centennial-scale climate variability. This means that traditional stochastic modelling approaches (i.e. calibrating to a ~100-year instrumental record) are unable to simulate risk arising from aleatory uncertainty and centennial-scale climate variability. However, several models capable of simulating long-term, centennial-scale climate variability when calibrated to extended, multi-centennial timeseries were identified. Two such models, the ARMA(1,1) and ARFIMA(0,D,0) models, were used for subsequent objectives.

3. Evaluate the role of sampling bias, conditioning error, and likelihood approximation when inferring stochastic model parameters under centennial-scale variability.

Using synthetic timeseries (generated from an ARMA(1,1) or ARFIMA(0,D,0) model) and Bayesian calibration methods, it was found that exact and conditional, approximate likelihoods return similar posteriors.

4. Evaluate stochastic model parameter stationarity using millennium-length palaeoclimate proxy records.

It was found that stochastic model mean and standard deviation are likely (a) non-stationary at multi-centennial and millennial timescales and (b) stationary at centennial timescales. Furthermore, stochastic model persistence is likely stationary over centennial, multi-centennial, and millennial timescales

5. Calibrate stochastic model persistence using ice core information within a Bayesian framework

For the final objective, a Bayesian framework for calibrating a stochastic rainfall model using palaeoclimate proxy data is presented. The framework uses proxy data from an Antarctic ice core and instrumental measurements from southeast Australia to calibrate a catchment-scale stochastic rainfall model. The proxy data is used to define a Bayesian prior for instrumental persistence. This extracts the proxy persistence signal, which is representative of broader regional persistence, without using the proxy to predict catchment-scale rainfall. When validated, the proposed model reproduces the observed drought risk. However, compared with the 'standard' model calibrated using a non-informative persistence prior, the palaeoclimate-informed model can simulate much longer and more severe droughts.

When answering Objectives 4 and 5, it became apparent that centennial-scale variability, aleatory uncertainty, and parameter uncertainty results in irreducibly 'wide' statistical uncertainty. This means that water supply systems must be robust under a future range of drought risk that is irreducibly 'wide'.

In the final discussion, 'wide' uncertainty is discussed within the context of 'deep' uncertainty associated with risk arising from anthropogenic climate change and the 'murky' uncertainty associated with imperfect knowledge, system complexity, and the subjective nature of socio-political values. Approaches for managing water under 'wide, deep and murky' uncertainty are also discussed.

Acknowledgements

I consider myself incredibly lucky to have an opportunity to write a PhD thesis. I feel like I do not deserve this luck; I can certainly think of others more deserving. But, because luck is indiscriminate, at the very least I want to acknowledge and express my gratitude for the various sources of luck that have provided this opportunity. I truly hope I can pay some of this luck forward to others who may need it.

The first source of luck that comes to mind is having Anthony as a supervisor. I initially contacted Anthony when I was an undergraduate about a final year research project. I did this without thinking how important a supervisor will be in my development as a researcher. I am lucky because, without thinking, I'd made a great choice in supervisor. I doubt I could have completed a thesis with a supervisor of different temperament or style too Anthony. Anthony has made me a better thinker, writer, reader, communicator, and learner. This has made me, overall, a better person.

The5econdd source of luck is that my co-supervisor, George, had just retired when my PhD started. This made George very accessible throughout my PhD. To have an accessible and engaged co-supervisor is great. However, I cannot understate how lucky I am to have had someone with George's experience and knowledge so accessible throughout my PhD. Conversations with George always leave me both challenged and inspired, and, like Anthony, George has made me a better thinker, writer, reader, communicator, and learner.

The third source of luck is having such an amazing network of friends and family. They provided the essential circuit breaker from the (occasional) myopia and drudgery of PhD work. The chance to catch up with friends or to spend time with my nephews, who care nothing about the minutiae of stochastic modelling, always left me feeling refreshed.

The fourth source of luck is, in my final 18 months, meeting Chel. Our relationship has been as amazing as it was unexpected. I cannot wait to spend more, and better, time with you now this is finished! I cannot thank you enough for all your help, understanding, and support during the last few months.

The final, and most important, source of luck is having two wonderful parents who have supported me throughout my studies. Regardless of my living (or economic!) situation, there was always a place where I could eat, stay, and relax. I know many people are not so fortunate to have the safety net they have provided throughout my life.

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

To manage water under climate variability and change, climate risk must be quantified. Climate risk can be viewed as a function of climate hazard (e.g. a drought), climate exposure (e.g. the probability of a drought occurring), and climate vulnerability (e.g. the potential impacts of a drought) (Kim et al., 2015).

In water management, the climate hazard and climate exposure is typically quantified using a stochastic model calibrated to instrumental measurements(Loucks and Van Beek, 2017). These stochastic models generate synthetic timeseries with similar statistics to the calibration data but with different, and potentially more severe, droughts (Fiering, 2013). Once generated, the synthetic timeseries are used as inputs into a water system model (Kuczera, 1992). This simulates water system behaviour under the various droughts generated by the stochastic model, which then informs water system operation, design, and adaptation (Vogel, 2017). This makes stochastic model calibration a crucial task for the design and operation of water supply systems.

When using stochastic models to infer climate risk, a key modelling assumption is one of parameter stationarity. (Milly et al., 2008). Parameter stationarity assumes that stochastic model parameters are time invariant (Koutsoyiannis and Montanari, 2015; Montanari and Koutsoyiannis, 2014). By assuming stationarity, there are implicit assumptions that (a) climate risk inferred from instrumental measurements is representative of historic risk and (b) historic risk is representative of future risk.

To better understand the validity of these assumptions, there is a need to consider alternative, longer hydroclimatic timeseries. This is because short instrumental records are subject to considerable parameter uncertainty, making it hard to validate or invalidate the stationarity assumption (Serinaldi and Kilsby, 2015; Thyer et al., 2006). Furthermore, short records contain few multi-decadal cycles of climate variability, and no multi-centennial cycles. This means instrumental records are likely too short to quantify the full range of natural climate variability (Vance et al., 2022) and, crucially, properly evaluate a stochastic models ability to reproduce low-frequency variability.

Palaeoclimate proxy records are a potential source of climate information that can be used to re-evaluate stochastic model performance and assumptions. These records are taken from

physical ‘layers’ with properties sensitive to climate (e.g., tree ring width being sensitive to available moisture), which are typically several hundred years in length (Griffin and Anchukaitis, 2014; Ho et al., 2015a; Verdon-Kidd et al., 2017). These records have been used to reconstruct pre-instrumental climate, with several studies identifying pre-instrumental ‘megadroughts’ (Cook et al., 2022; Helama et al., 2009; Routson et al., 2011; Stevenson et al., 2022). Naturally, a megadrought invokes a concern for water security.

Megadroughts have quite severe social and economic impacts (Fernández et al., 2023; Muñoz et al., 2020), but adapting a water supply system to mitigate the risks posed by megadrought also has social and economic impacts (Gober et al., 2016). Therefore, any water security concerns about megadrought should be viewed with respect to the limitations and assumptions made when using proxy records as a source of climate information.

From a water management and climate risk perspective, potential issues with proxy records include:

1. Proxy data are imperfect recorders of climate information, meaning they have limited skill in predicting climate. This means proxy-based climate reconstructions underestimate instrumental-period variance (Meko et al., 2022). By extension, these reconstructions also underestimate the magnitudes of instrumental-period extremes (Patskoski et al., 2015). These statistics are crucial for any climate risk assessment.
2. There are limited in-situ (i.e., local) proxy data for catchments of interest (Galelli et al., 2021; Tingstad et al., 2014). This is a particularly prevalent issue across the mid-latitude Southern Hemisphere (Croke et al., 2021; Goodwin et al., 2022; O’Connor et al., 2022).
3. Proxy-based statistical reconstructions assume that the proxy-climate relationship is stationary and can be inferred from the instrumental-period. However, numerous factors influence proxy formation and properties; the relative influence of these different factors may change between instrumental and pre-instrumental periods (Cook, 1985; Kiem et al., 2020; Pelletier and Turcotte, 1997; Tozer et al., 2016). This means that the statistical model calibrated from the instrumental period may not be wholly suitable for some pre-instrumental periods (D’Arrigo et al., 2008).

Despite the limitations of proxy records, there are physical explanations, and corresponding evidence from different proxy types, indicating that climate varies at centennial and millennial timescales. In short, megadroughts are possible (Cook et al., 2022). Therefore, there is a clear

need to re-evaluate and, potentially, update the statistical methods/assumptions used when inferring climate risk in water management using palaeoclimate (i.e. stochastic models). Doing so while considering the limitations of proxy records is the key motivation of this thesis.

1.1 Objectives

The goal of this thesis is to (a) use palaeoclimate proxy records to evaluate stochastic model performance and parameter stationarity under long-term, centennial-scale variability and (b) present a stochastic modelling framework that incorporates proxy centennial-scale variability. To achieve this goal, the thesis had five key objectives:

1. Evaluate the persistence signal in Antarctic ice core records.

In the Southern Hemisphere, Antarctic ice cores are a major source of annually-resolved palaeoclimate information. These records can contain persistence signals influenced by the El Nino Southern Oscillation and the Interdecadal Pacific Oscillation, which also influence hydroclimatic persistence in the mid-latitudes (Tozer et al., 2016; Vance et al., 2013). It is important to understand if and how ice core persistence differs from hydroclimatic persistence because the major population centres of the mid latitude Southern Hemisphere, such as southern Australia, southern South America, South Africa, experience a highly variable climate (Grimm et al., 2000; Mason and Jury, 1997; Verdon et al., 2004).

2. Evaluate different stochastic models using millennium-length palaeoclimate proxy records.

Palaeoclimatology and stochastic models have an overlapping goal: characterising climate variability. Palaeoclimatology pursues this goal by statistically modelling physical relationships between climate and climate proxies; stochastic models pursue this goal by statistically modelling the inherent randomness and persistence of the climate system. Considering the relative simplicity of stochastic models and the availability of independent proxy records that represent pre-instrumental climate variability, it is of interest to assess the ability of stochastic models to simulate pre-instrumental variability when calibrated to the instrumental record. It is also of interest to identify whether stochastic models can reproduce low-frequency climate variability captured by proxy data.

3. Evaluate the role of sampling bias, conditioning error, and likelihood approximation when inferring stochastic model parameters under centennial-scale variability.

Quantifying stochastic model parameter uncertainty may be necessary to ensure that climate risk estimates derived from stochastic models, which inform water management, are reliable (Berghout et al., 2017; Stedinger and Taylor, 1982a). To quantify parameter uncertainty, Bayesian calibration methods (which require a likelihood function) are often used (Gelman et al., 2013). However, accurately quantifying parameter uncertainty with Bayesian methods may be difficult because hydrological processes can exhibit centennial-scale variability (i.e. long-term persistence). Instrumental rainfall and streamflow records used in stochastic model calibration are only ~100-years long; meaning that, with respect to centennial-scale variability, instrumental records have a sampling bias. Furthermore, conditional and approximate likelihoods are often used (for ease of computation) (Beran, 2017; Haslett and Raftery, 1989). Under centennial-scale variability, the errors associated with the initial conditioning and approximation could, potentially, bias parameter inference (Box et al., 1970). Therefore, for timeseries exhibiting centennial-scale variability, it is of interest to evaluate if and how sampling bias, conditioning error, and likelihood approximation impacts Bayesian inference of stochastic models.

4. Evaluate stochastic model parameter stationarity using millennium-length palaeoclimate proxy records.

When using stochastic models to infer climate risk, a key modelling assumption is one of parameter stationarity. Parameter stationarity assumes that stochastic model parameters are time invariant (Koutsoyiannis and Montanari, 2015; Milly et al., 2015). However, validating the stationarity assumption is difficult. This is because instrumental rainfall and streamflow records are relatively short. Short records are subject to considerable statistical uncertainty (making it hard to identify clear statistical change, even under global warming) and may not capture long-term climate variability (Serinaldi and Kilsby, 2015; Thyer et al., 2006). Palaeoclimate proxy records, which span hundreds/thousands of years, can better assess stationarity because longer record lengths will (a) reduce statistical uncertainty; and (b) contextualise if any recent hydroclimatic changes are consistent with historic climate variability. Validating or invalidating the stationarity assumption under historic variability has implications for estimating drought risk and, subsequently, determining appropriate water management decisions and infrastructure design.

5. Calibrate stochastic model persistence using ice core information within a Bayesian framework

Any palaeoclimate-informed stochastic modelling framework must preserve instrumental-period variance and extremes during calibration. Existing frameworks have done this by using proxy data to inform the resampling and/or stochastic modelling of the instrumental record (Erkyihun et al., 2016; Gangopadhyay et al., 2009). Although these methods preserve variance, the sampled wet or dry values are limited to either (a) those contained in instrumental data; or (b) those derived from stochastic models calibrated to wet or dry instrumental periods. Regarding (a), this means extrapolation to larger, unrecorded extremes is not possible (such extremes are possible and should be accounted for when quantifying drought risk). Regarding (b), given that these wet/dry periods are a subset of an already short instrumental record, parameter uncertainty will be substantial. Large parameter uncertainty will propagate through a water system model, which makes it hard to identify optimal management rules and infrastructure (Berghout et al., 2017; Stedinger and Taylor, 1982a). Considering these limitations, a palaeoclimate-informed stochastic modelling approach that extrapolates to unobserved values while minimising parameter uncertainty is desirable.

These objectives are primarily concerned with understanding and quantifying historic 'baseline' risk. This 'baseline' risk can be used to contextualise future risks posed by anthropogenic climate change (Armstrong et al., 2020). In the final discussion, the implications of the thesis objectives are discussed within the context of the various sources of uncertainty and risk posed by anthropogenic climate change (e.g. Maier et al. (2016)) and the 'wicked' nature of water resource management (e.g. Kwakkel et al., 2016a; Wu et al., 2023)

1.2 Thesis structure

In this thesis, each chapter is organised as a stand-alone paper investigating a significant gap in the literature. The literature review, knowledge gaps and research significance are embedded within the introduction of each paper (as opposed to presenting these in a stand-alone thesis chapter). At the end of each paper is a section describing how the current and previous papers are linked with the succeeding paper. This structure was chosen as an admittedly imperfect trade-off between the need to produce a coherent thesis and the desire to produce potential journal papers.

Chapter 2. Evaluating hydroclimatic persistence signals in Antarctic ice cores

2.1 Abstract

Antarctic ice cores are an important source of high resolution palaeoclimate information in the Southern Hemisphere. More recently, Antarctic ice cores have provided opportunities to study long-term hydroclimatic persistence (i.e. the tendency for wet and dry years to cluster) and drought risk in Australia. However, due to a lack of long-term, in-situ precipitation/snowfall measurements, the fidelity of the ice core persistence signal is unknown. In this study, the persistence signal in 57 annual snowfall accumulation records and 48 ice core sodium Na+ records was evaluated against extended annual rainfall records from the mid-latitude Southern Hemisphere (23.5°S to 50°S). Hydroclimates in Antarctica and the mid-latitude Southern Hemisphere are both influenced by Southern Ocean synoptic systems, the Southern Annular Mode and the El Nino Southern Oscillation, allowing ice core persistence to be evaluated against mid-latitude persistence. We found the persistence signal in annual snowfall accumulation and mid-latitude rainfall records to be statistically similar. This indicates that annual snowfall accumulation persistence is not significantly biased by post-deposition processes (such as wind erosion and ice advection) and statistical processing after core collection. In contrast, ice core Na+ records tended to have slightly higher persistence than mid-latitude rainfall. Results from this study suggest that annual snowfall accumulation records can be used to characterise hydroclimatic persistence and regional drought risk. In contrast, before being used to characterise regional drought risk, ice core Na+ records require site-specific assessments that identify the constituent climate signal. These insights can guide future research in palaeoclimate-informed drought risk assessment and water management in the mid-latitude Southern Hemisphere.

2.2 Introduction

Hydroclimatic persistence refers to the tendency of wet and dry years to cluster (Hurst, 1951; Markonis and Koutsoyiannis, 2016). There are numerous statistical methods for estimating and modelling persistence which are, in turn, used for estimating drought risk (Hosking, 1984; Thyer and Kuczera, 2000). These drought risk estimates inform how water supply systems are designed and managed under climate variability (Loucks and Van Beek, 2017). Calculating regional hydroclimatic persistence using accurate data is, therefore, crucial for successful water management. In this study, we evaluate the persistence signal in an alternative source of hydroclimatic data – palaeoclimate data from ice core records.

Typically, hydroclimatic persistence is calculated using instrumental rainfall and streamflow records – but there remains an open question whether these records have sufficient temporal coverage. At best, instrumental records are ~100-150 years long in the Southern Hemisphere (although many regions have shorter records, Menne *et al.*, 2012; Higgins *et al.*, 2023). Calculating persistence requires long records; around 100-years is the minimum length (Koutsoyiannis, 2003). However, even with longer records, there is still considerable statistical uncertainty (Thyer et al., 2006; Weron, 2002a). This uncertainty makes it hard to identify optimal water system designs and management rules (Berghout et al., 2017).

From a water management perspective, alternative, longer sources of hydroclimatic information would be useful for calculating regional persistence. One such data source is from palaeoclimate archives which preserve proxy climate records. These proxy records are derived from climatically sensitive, physical 'layers', - such as tree-ring widths and ice core physical and chemical properties (such as snowfall accumulation rates) - which (a) can be dated and (b) have formed over hundreds to thousands of years (Weedon, 2003). This means that proxy records can extend observational records and, for hydroclimatically sensitive proxies, provide additional information about long-term climate variability and drought risk (Cook et al., 2015; Palmer et al., 2015)

Although proxy records are potentially useful for studying persistence and drought risk, note that proxy records are imperfect recorders of climate information. Numerous processes can confound the climate signal of interest (Weedon, 2003) Therefore, from a climate risk

perspective, any study of proxy data should be prefaced by an assessment of proxy bias. In this study, we evaluate the persistence signal in Antarctic ice cores.

This evaluation is necessary because:

1. It is important to understand if and how proxy persistence differs from hydroclimatic persistence because the major population centres of the mid-latitude Southern Hemisphere, such as southern Australia, southern South America, South Africa, experience a highly variable climate (Grimm et al., 2000; Mason and Jury, 1997; Verdon et al., 2004).

For these regions, the full extent of climate variability and, by extension, climate risk is hard to quantify using instrumental measurements (Ho et al., 2015b; Kiem et al., 2020; Mundo et al., 2012). To better understand the climate risks facing mid-latitude water security, palaeoclimate proxy information is useful (Armstrong et al., 2020; Fernández et al., 2018; Sauchyn et al., 2015).

2. Typical climate risk assessments (e.g., evaluating water supply system performance) require annual/sub-annual resolution climate data (Fowler et al., 2022; Kuczera, 1992; Ren et al., 2023).

Antarctic ice cores are one of the primary sources of annually-resolved palaeoclimate data in the Southern Hemisphere. Tree-rings, another common source of annually-resolved palaeoclimate data, are not widely available near major mid-latitude Southern Hemisphere populations, such as east Australia and southern Africa (Dixon et al., 2017; Flack et al., 2020; Goodwin et al., 2022).

3. Previous work evaluating proxy persistence has exclusively focussed on tree-ring records.

These studies found that tree-rings exhibit larger persistence than instrumental rainfall (Franke et al., 2013a; Zhang et al., 2015). This is because the relationship between rainfall and tree growth is mediated by temperature and soil moisture, which exhibit stronger persistence than rainfall. The stronger persistence signal will influence subsequent drought risk estimates(Ludescher et al., 2020; Yuan et al., 2021). Because proxy data can contain a mix of climate signals, which can influence the persistence signal, an evaluation of the persistence signal in Antarctic ice cores is needed before it can be used to estimate drought risk and inform water management.

Various properties are measured in ice cores – in this study, we evaluate the persistence signal from two common measurements: annual snowfall accumulation and annual sodium (Na+) concentration.

Snowfall accumulation is a measure of how much snowfall was incorporated into the ice sheet over a given period (typically annual). Clearly, this is closely related to regional snowfall, which in turn is driven by regional weather/synoptic systems (Thomas et al., 2017; Wang et al., 2017).

Various processes could potentially confound the persistence signal in annual snowfall accumulation records preserved in ice cores. Wind deposition or erosion can remove or redistribute snowfall from one area to another, distorting the eventual annual layer (Thomas et al., 2017; Wang et al., 2017). Advection of ice upslope from where the ice core is drilled, combined with different climate conditions at the deposition site (e.g. a lower annual snowfall rate), may introduce bias (Huybrechts et al., 2007). Finally, annual ice layers also thin over time. This is due to the vertical strain caused by new snowfall and downslope movement of the ice sheet (Dansgaard and Johnsen, 1969). When inferring ice accumulation, this thinning can be accounted for using a statistical or physical model (Nye, 1963; Parrenin et al., 2004). It is unknown if and how the thinning model changes the snowfall accumulation persistence signal, although different thinning models produce proxy records with similar persistence (Roberts et al., 2015). However, the fidelity of these thinning models with respect to observations is unknown.

The second ice core measurement type evaluated in this study is the concentration of sodium ions (Na+) in the ice itself. For high annual snowfall sites, ice core Na+ is generally wet deposited (e.g. associated with snowfall events) (Wolff et al., 2006). The sodium originates from the open ocean or sea ice and is scoured as a direct function of surface wind speed, then transported to the ice core site by synoptic scale tropospheric processes (e.g. mid-latitude cyclonic circulation). These broader-scale synoptic processes also influence regional hydroclimate (Udy et al., 2022, 2021). However, the use of ice core Na+ as a hydroclimatic proxy is relatively new; links between ice core Na+ and regional hydroclimate have not been studied to the same extent as snow accumulation.

Various processes could also confound the persistence signal in ice core Na+ records. Similar to snow accumulation records, post-deposition transport or erosion could distort ice core Na+ concentration. However, unlike annual snowfall accumulation records, sea salt concentration records undergo less statistical post-processing (because thinning does not impact layer concentration) (Sigl et al., 2016; Winski et al., 2019). Furthermore, because ice core Na+ deposition is primarily mediated by wind, relationships between ice core Na+ and regional snowfall are somewhat indirect. Other factors that influence wind and atmospheric circulation, both local and global, may also influence ice core Na+ persistence.

Complicating an evaluation of Antarctic ice core persistence is a lack of long-term, in-situ precipitation observations that can be used for validation. However, some mid-latitude regions have long-term observations, and these regions are hydroclimatically linked to Antarctica. For example, there are common synoptic systems influencing east Antarctic snowfall and southern Australian rainfall (Udy et al., 2022, 2021; van Ommen and Morgan, 2010; Zheng et al., 2021). Southern Ocean storm fronts also drive rainfall in western South Africa (Stager et al., 2012). Furthermore, variability in the mean latitudinal positions of Southern Ocean storm tracks are influenced by the El Nino Southern Oscillation (ENSO) (Crockart et al., 2021; Dätwyler et al., 2020; Vance et al., 2013). ENSO is also a primary driver of climate variability in the mid-latitude Southern Hemisphere (Kiem and Franks, 2004; Power et al., 1999; Westra et al., 2015). This means climate variability in mid-latitudes (i.e. 23.5°S to 50°S) and high latitudes are influenced by the same large-scale drivers and synoptic systems. Due to these links between Antarctic and mid-latitude climate, extended rainfall observations from Southern Hemisphere subtropical and temperate zones that are climatologically linked to the Southern Ocean and Antarctica can be used to validate the persistence signal in ice cores.

In this study, we evaluate the persistence signal in hydroclimatically sensitive, Antarctic ice core proxies using extended rainfall timeseries from the mid-latitude Southern Hemisphere. Due to a lack of in-situ ice core measurements, it is not possible to compare ice core measurements and rainfallon a site-by-site basis. Instead, statistical distributions should be compared to assess if the ice core and rainfall sampling distributions are similar. Further inferences about ice core bias (or lack thereof) are made based on two assumptions:

1. That Antarctica and the mid-latitudes exhibit similar hydroclimatic persistence. This assumption is based on studies which demonstrate that hydroclimatic persistence

exhibits no broad-scale spatial dependence (Fatichi et al., 2012; Iliopoulou et al., 2018; Markonis et al., 2018).

2. Any difference between ice core and rainfall persistence is due to non-hydroclimatic factors impacting the ice core (e.g., wind deposition and erosion, ice sheet thinning, statistical processing of original measurements, additional climate signals). Conversely, any similarity is due to the accurate recording of hydroclimatic persistence in the ice core proxies.

2.3 Data

Rainfall stations and ice core records used in this study are shown in Figure 2-1. In total, 57 annual snowfall accumulation records, 48 ice core Na+ records, and 886 rainfall stations (32 from South Africa, 48 from New Zealand, and 806 from Australia) were included in the final analysis (no stations from South America met the selection criteria). All annual data is based on calendar year.

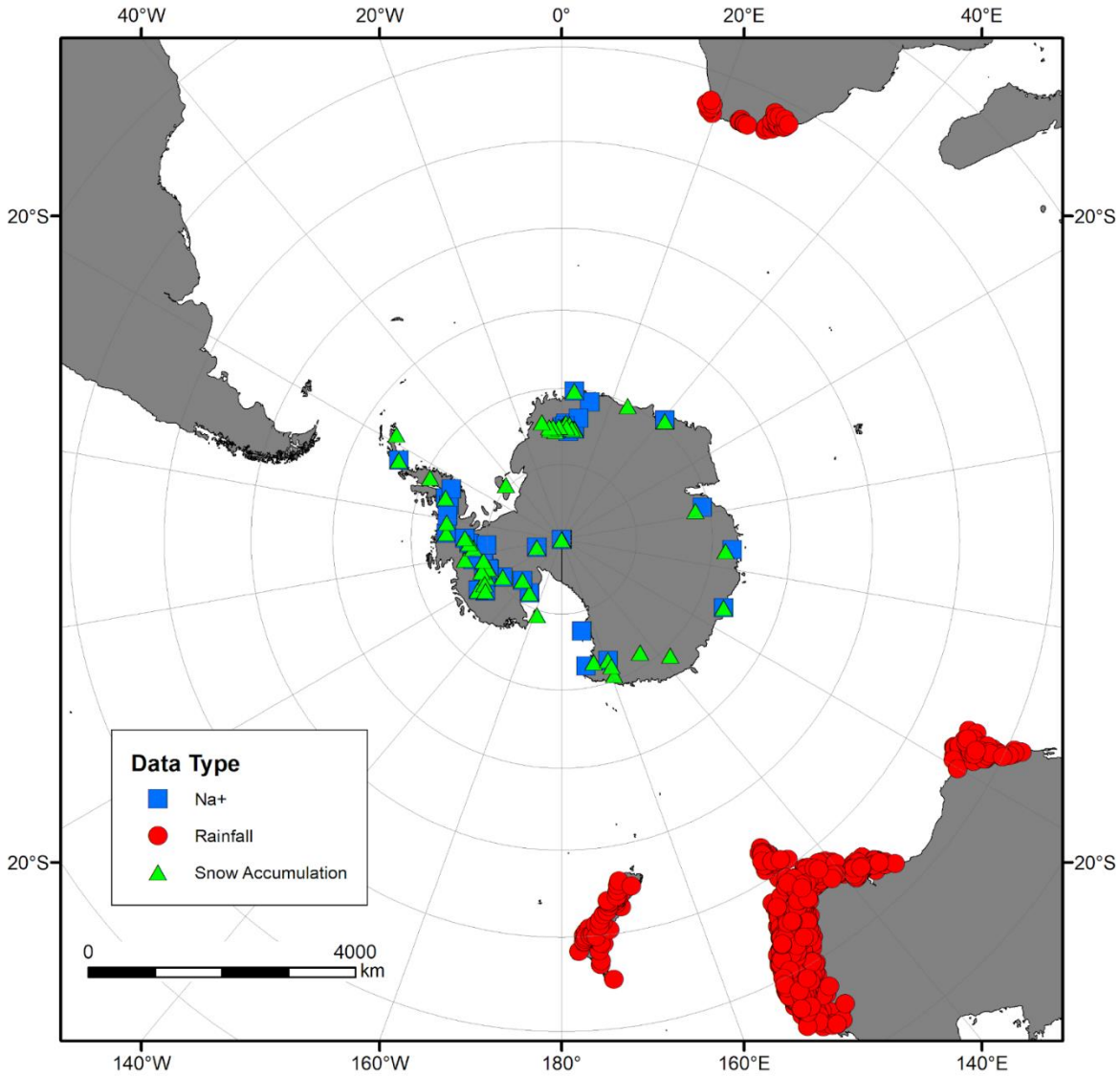

Figure 2-1: Data used in this study

2.3.1 Rainfall data

Rainfall data was taken from three different sources. For Australian stations, rainfall data was taken from the SILO dataset. SILO is a publicly available climatological dataset (https://www.longpaddock.qld.gov.au/silo/) which provides station data throughout Australia (Jeffrey et al., 2001). SILO was selected over globally focussed datasets because missing daily values had already been infilled via multivariate interpolation (referred to as 'patched point' data). Such infilling is not typical in globally focussed datasets. Other stations were taken from the Global Historical Climatological Network Daily (GHCN-D). GHCN-D comprises daily climate data from ~80,000 stations worldwide and has undergone an automated quality assurance check (Durre et al., 2010; Menne et al., 2012). This data has been used in previous

studies examining long-term persistence in rainfall (Iliopoulou et al., 2018; Tyralis et al., 2018). Finally, because extended rainfall stations from New Zealand are not included in the GHCN-D dataset, data for this country was downloaded from the National Institute of Water and Atmospheric Research (NIWA) CliFlo database (https://cliflo.niwa.co.nz/). Note that there is little documentation describing the CliFlo database, meaning quality control/infilling methods are unknown to the authors. Rainfall stations were selected based on record length, proportion of non-missing years, and location/climate zone - specific criteria will now be discussed.

Regarding rainfall record length, although instrumental measurements contain limited information about long-term persistence, this can be somewhat remediated by examining long observational records (e.g. over 100-years) (Koutsoyiannis, 2003). As such, for SILO; GHCN-D; and CliFlo datasets, only stations with observations spanning 100 years were selected. In the case of the SILO data, where missing data has been infilled via spatial interpolation, stations were included if they comprised 80% observations. Annual values for SILO stations were calculated as an average of daily rainfall totals. For GHCN-D stations, stations with a minimum of 90% complete years were selected. A complete year comprised of 90% observations. A sensitivity analysis was conducted which varied these specific percentage thresholds, but no difference was found in the distributions of subsequent persistence statistics. GHCN-D station annual values were calculated as the average daily total for non-missing days. Finally, CliFlo stations (for which annual rainfall totals were downloaded directly) were selected provided they had a minimum of 90% non-missing years.

Missing values in annual GHCN-D and CliFlo timeseries were infilled using the K-Nearest Neighbour method.

Regarding location, rainfall stations were considered provided they (a) were located below -23.5°S (a rough approximation for the Tropic of Capricorn); and (b) were in a Temperate Koppen-Geiger climate zone with winter rainfall influences (i.e., Csa; Csb; Csc; Cfa; Cfb; and Cfc climate subgroups). These regions were selected to reduce the influence of tropical synoptic systems on station rainfall. The climate zones were selected to ensure station rainfall was somewhat influenced by Southern Ocean synoptic systems, which have their strongest influence on mid-latitude rainfall during winter (Meneghini et al., 2007; Reason and Jagadheesha, 2005; Risbey et al., 2009).

2.3.2 Snowfall accumulation records

Snowfall accumulation records were mainly taken from Thomas et al. (2017). Along with this dataset, more recent ice core records were also included: the South Pole Ice Core of Winski *et al.* (2019) and the updated Law Dome record of Jong et al. (2022). Only annually resolved records were chosen, with most records having average accumulation rates > 100 kg m^2 $year^{-1}$. For high accumulation sites like these, the impact of post-deposition processes on the final accumulation should be small (i.e. most of the initial snowfall will be accumulated into the ice sheet, regardless of post-deposition redistribution).

From the initial dataset, accumulation records were selected for further analysis if (a) they contained at least 100 observations in the period overlapping with rainfall records (which was 1853 onwards); and (b) had no more than 10% missing data from the first year to the last years. 57 records met these requirements, with all records having <5% missing years. Missing years were infilled using the K-Nearest Neighbour method.

2.3.3 Sea salt (Na+) records

Sea salt concentration records were taken from Thomas et al. (2023), who collated various ice core geochemistry records (including Na+ concentration). Na+ records were screened and infilled following the same method used for accumulation. Many of the selected Na+ records were close to the coast and located at or near high snow accumulation sites, which means Na+ is more likely to be deposited with snowfall (i.e. wet deposition) than

Note that there are other ice core proxies available, such as Na+ deposition flux and various sulphate records. However, Na+ deposition flux is calculated based on multiplying Na+ measurements by snowfall accumulation rate (Thomas et al., 2023). This multiplication introduces a dependence between the deposition flux and accumulation and is likely only necessary at annual snowfall accumulation rates of less than 100 kg m^2 $year^{-1}$ (~10 cm ice equivalent per year) (Wolff et al., 2006). Because most of the study sites were from high accumulation areas, Na+ concentration was evaluated instead of deposition flux. Sulphate records were also not considered. These records are closely linked to volcanic activity and marine biogenic emissions (Curran et al., 2003; Plummer et al., 2012; Sigl et al., 2016) Links with hydroclimate are less clear and require further research (Thomas et al., 2023)

2.4 Methods

In this study, we calculated and compared distributions of the Hurst coefficient - a common measure of timeseries persistence (Hurst, 1951). The Hurst coefficient is a dimensionless measure of how timeseries variance changes across aggregation scale, which is indicative of how similar timeseries values cluster. Stationary timeseries have Hurst coefficients ranging from 0 to 1. Timeseries with no persistence (i.e. white noise) have a coefficient of 0.5. In contrast, persistent timeseries (e.g. geophysical timeseries) have a Hurst coefficient >0.5. The Hurst coefficient has been used extensively to study and characterise persistence in hydroclimatic timeseries - for detailed reviews, refer to O'Connell et al. (2016) and Graves et al. (2017).

There are various methods of estimating the Hurst coefficient. To account for this uncertainty, we calculated and compared Hurst distributions using the following methods:

- Rescaled range (R/S) (Mandelbrot and Wallis, 1969);
- Least squares based on standard deviation (LSSD) (Tyralis and Koutsoyiannis, 2011);
- Whittle estimator (Beran, 2017);
- Detrended fluctuation analysis (DFA) (Peng et al., 1995);
- Maximum likelihood estimator (MLE) (McLeod and Hipel, 1978);
- Periodogram regression via the Geweke and Porter-Hudak method (GPH) (Geweke and Porter-Hudak, 1983).

For a description of the different persistence estimators (and a thorough assessment of their differences), refer to Taqqu et al. (1995); Tyralis and Koutsoyiannis (2011); and Weron (2002).

Although there are numerous methods of estimating the Hurst coefficient, these persistence estimators can be biased when performed on small sample sizes (Cannon et al., 1997; Hamed, 2007; Kendziorski et al., 1999; Wallis and Matalas, 1970). To assess potential biases in the different Hurst coefficient estimators, an analysis using synthetic data was conducted. Synthetic timeseries 125 years long (the mean length of the rainfall records used in this study) were created via a Fractional Gaussian Noise (FGN) model (Mandelbrot, 1971). The FGN model can generate timeseries with a specific Hurst coefficient (subject to sampling uncertainty). For Hurst coefficients ranging from 0.5-0.7 (increments of 0.02), 1,000 synthetic timeseries were generated and the Hurst coefficient estimated.

Figure 2-2 shows Hurst coefficient sampling distributions for the different estimators, calculated from the synthetic data. We can see that:

- All methods, aside from GPH, tended to underestimate the Hurst coefficient.
- The R/S estimator had the largest bias in underestimating the Hurst coefficient.
- The DFA estimator consistently underestimated the Hurst coefficient, however, this was offset by larger sampling variability.
- The GPH estimator had much larger sampling variability than the other estimators.

We excluded the R/S estimator due to its bias. We excluded the GPH estimator due to its large sampling variability (GPH sampling distributions consistently spanned the entire 0-1 stationary range). Large sampling variability reduces statistical power (i.e., the ability of a statistical test to detect differences).

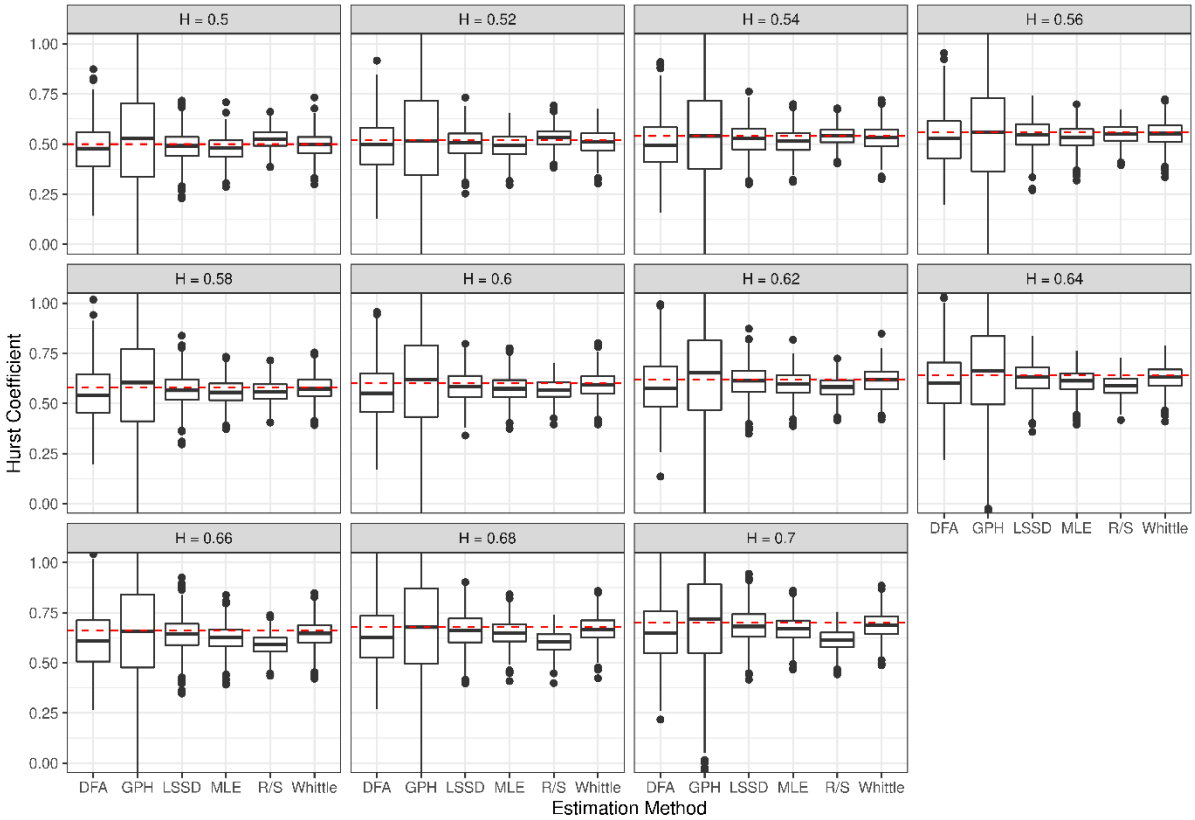

Figure 2-2: Hurst coefficient sampling distributions calculated from synthetic data. Red lines show the Hurst coefficient used when generating synthetic data with Fractional Gaussian Noise model.

Aside from the estimation method, comparing Hurst distributions is also complicated by considerable sampling uncertainty. For timeseries containing ~100 values (i.e, the timeseries used in this study), Hurst coefficients estimated from white noise can cover most of the stationary 0-1 range (Weron, 2002). Sub-sampling different periods of the same record can also result in different Hurst coefficient estimates (Markonis and Koutsoyiannis, 2016). To minimise how differences in rainfall and ice core record length and period could influence the analysis, we used the following method to compare rainfall and ice core persistence signals:

1. Randomly assign a rainfall station to each of the ice core records (e.g. randomly choose 57 rainfall records and assign each to an ice accumulation record).
2. Subset corresponding ice core and rainfall records to cover the same period.
3. Calculate the Hurst coefficient of each record using different estimation methods.
4. Calculate the difference between corresponding ice core and rainfall Hurst coefficients for each estimation method.
5. Evaluate the difference distribution via the Student's T-test and Wilcoxon Rank Sum test respectively. An insignificant test result indicates that ice core and rainfall persistence is statistically similar.
6. Repeating steps 1-4 until all rainfall records had been assigned to an ice core record. Note that there were 57 ice accumulation records, 48 ice core Na records, and 886 rainfall records. This meant that there were 15 ice accumulation/rainfall samples and 19 ice core Na+/rainfall samples. For the 15th and 19th accumulation and ice core Na samples, some rainfall records had to be resampled.

Aside from accounting for sample size sensitivity, Step 1 and Step 2 served two other purposes. First, Step 1 (randomly sub-sampling/pairing rainfall records with ice core records) increased the likelihood of independent rainfall samples. Although various studies have indicated that hydroclimatic persistence does not change across the high and mid-latitudes (Fatichi et al., 2012; Iliopoulou et al., 2018; Markonis et al., 2018), there is evidence for spatial correlation at scales less than 200km (Tyralis et al., 2018). This should be accounted for because both the Student's T-test and Wilcoxon test assume independent samples. Second, Step 2 (i.e. subsetting the records to cover the same period) ensured that ocean-atmospheric processes influencing Antarctic and mid-latitude climate variability (e.g., ENSO, Southern Annular Mode) were in the same relative state for both records.

When interpreting results from the proposed method, note that there is large statistical uncertainty when estimating the Hurst coefficient from ~100-year timeseries. Naturally, a question arises: what differences in persistence statistics can we meaningfully detect with the proposed method? In other words, what is the statistical power of the proposed method? To assess statistical power, we conducted a similar analysis using synthetically generated data.

As with the bias analysis (Figure 2-2), the statistical power analysis involved generating ice core/rainfall timeseries via a Fractional Gaussian Noise (FGN) model (Mandelbrot, 1971). Synthetic 125-year ice core and rainfall timeseries were generated using the FGN model with combinations of Hurst coefficients ranging from 0.54-0.64 (in increments of 0.02). 125-years was selected because this was the mean length of the rainfall timeseries.

For each combination of ice core and rainfall Hurst coefficient, 50 synthetic ice core and rainfall timeseries were generated (similar to the number of Na+ records used). For each timeseries, the Hurst coefficients were calculated. The distribution means and medians were then compared using a Student T-test and Wilcoxon test. Note that, for the synthetic experiment, pairing/subsetting was not necessary. This was repeated 500 times, with the proportion of statistically significant p-values (i.e., p-value less than 0.05) calculated. For synthetic timeseries with the same Hurst coefficients, approximately 5% of the p-values will be significant.

Results from the statistical power analysis are shown in Figure 2-3. Rows (columns) display the Hurst coefficient of synthetic ice core (rainfall) data. We can see that all Hurst coefficient estimation methods and statistical tests correctly identify differences when Hurst coefficients differ by 0.1. However, statistical power decreases as the difference between synthetic Hurst coefficients become smaller. For the DFA estimator, statistical power decreases when the difference is 0.08 or smaller. For the other estimators, statistical power is low when the difference is 0.04 or smaller. This suggests that the study method is unable to detect small differences between ice core and rainfall persistence. Instead, any potential bias in Hurst coefficient will be at most ~0.04 (for the LSSD, MLE, and Whittle estimators) and ~0.08 for the DFA estimator. However, compared with the large sampling variability evident in Figure 2-2, such biases are small.

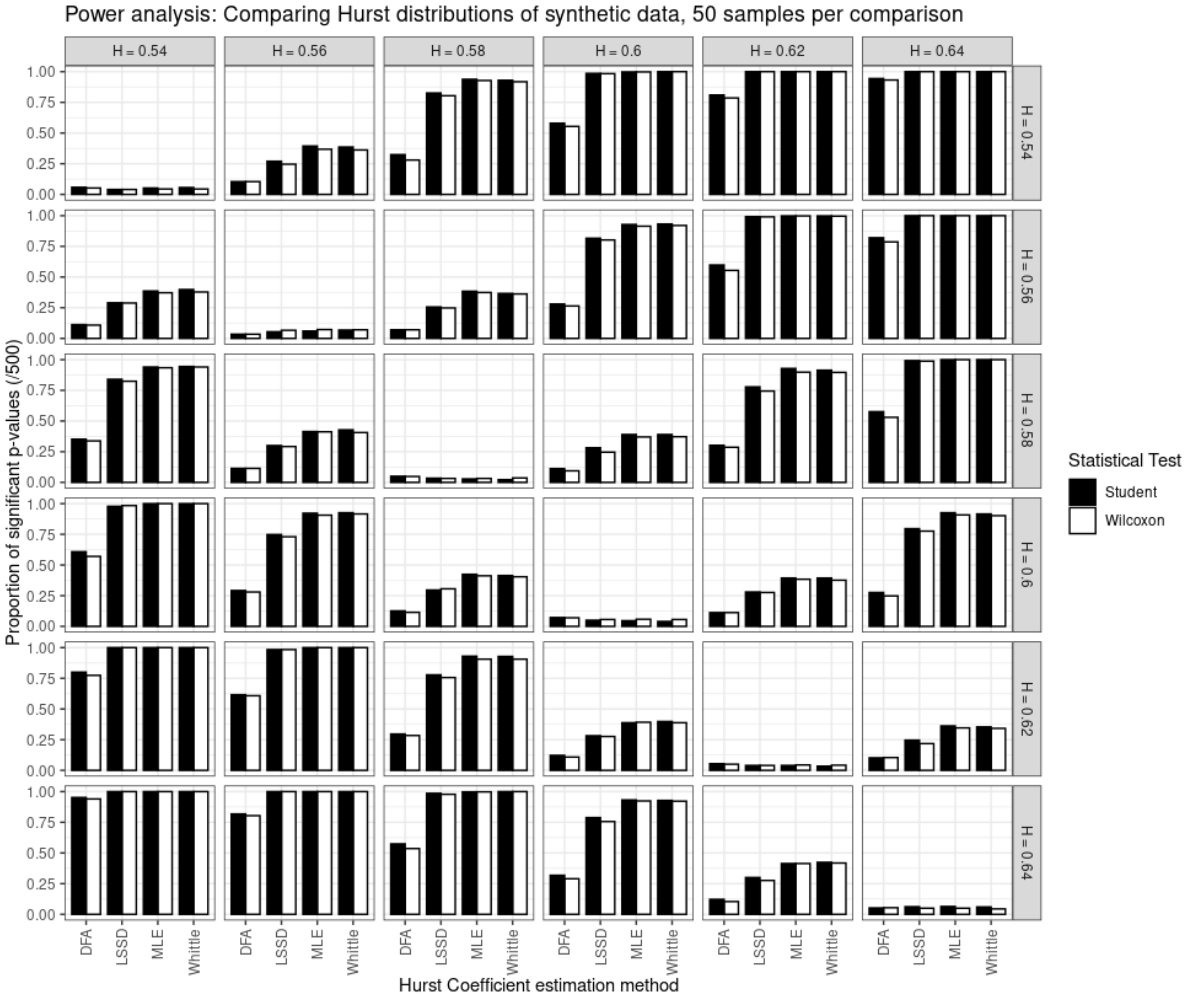

Figure 2-3: Proportion of statistically significant p-values from 500 iterations of the synthetic experiment. A single iteration involved generating 61 synthetic rainfall/ice core timeseries with a specific Hurst Coefficient and comparing distributions. Synthetic timeseries were 125 years long, which was the average length of the observed rainfall records. Rows (columns) display the coefficient of synthetic ice core (rainfall) data.

Aside from comparing Hurst coefficients, we also compared ice core and rainfall power spectrum. The global wavelet power spectrum was calculated for each record using the Morelet convolution (Torrence and Compo, 1998). For rainfall, ice accumulation, and ice core Na+, the median and 90% sample intervals of the global average power were calculated for different periods. This was to evaluate if ice core and rainfall records had similar power for interannual frequencies. Because Antarctic and mid-latitude hydroclimate is influenced by ENSO, we expect similar spectral power across the 3-7 year period associated with ENSO variability.

Prior to the wavelet analysis, both the ice core and rainfall records were detrended. Either linear or quadratic detrending was used. Detrending method was chosen by fitting and comparing two regression models – one with time as a single predictor (i.e linear detrending), another with

time and time squared as a predictor (i.e. quadratic detrending). Detrending was performed using the model with the lowest Akaike Information Criterion. Note that in the absence of a statistically significant trend, this approach is equivalent to subtracting the mean from the record. This subtraction is already necessary prior to wavelet analysis. Therefore, detrending was applied to all ice core and rainfall records.

Note that detrending is necessary prior to spectral analysis, but not prior to calculating the Hurst coefficient. There are two key reasons for this difference. First, some Hurst coefficient estimation methods have specific detrending steps (e.g. DFA), so explicit detrending was not necessary (Peng et al., 1995). Second, a key purpose of detrending prior to spectral analysis is to remove any low-frequency 'leakage' that might confound the higher frequency signals of interest. In contrast, to estimate the Hurst coefficient, any low-frequency signal should be preserved, not removed. Furthermore, for Hurst coefficient estimation methods without an explicit detrending step, there is limited guidance on if and how trends should be considered. Therefore, methods were applied as described in the relevant journal article.

2.5 Results

Figure 2-4 shows the mean and median difference between ice core and rainfall Hurst coefficients. We can see that:

- Annual snowfall accumulation and rainfall records had statistically similar Hurst coefficients, regardless of rainfall sample and Hurst coefficient estimator.
- Ice core Na+ records had higher Hurst coefficients than rainfall records (around 0.05 higher, on average). This was sensitive to rainfall sample and Hurst estimator.

In Figure 2-4, note that some significant and insignificant results were not perfectly separated along the x-axis. This is because different samples had different variance (Figure 2-5) and sample variance will also influence the statistical significance.

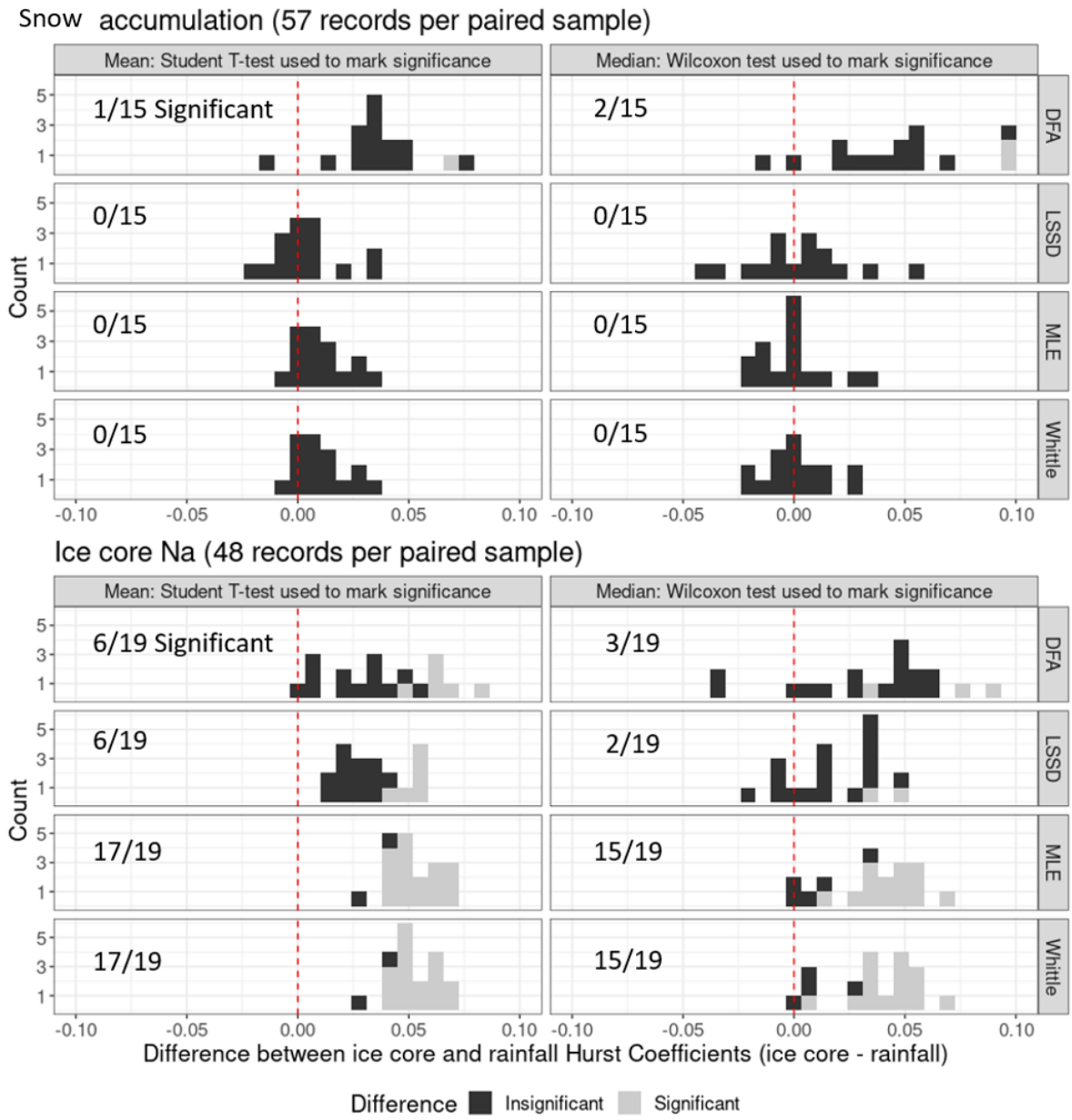

Figure 2-4: Mean and median difference between ice core and rainfall samples. Significant and insignificant results are shown.

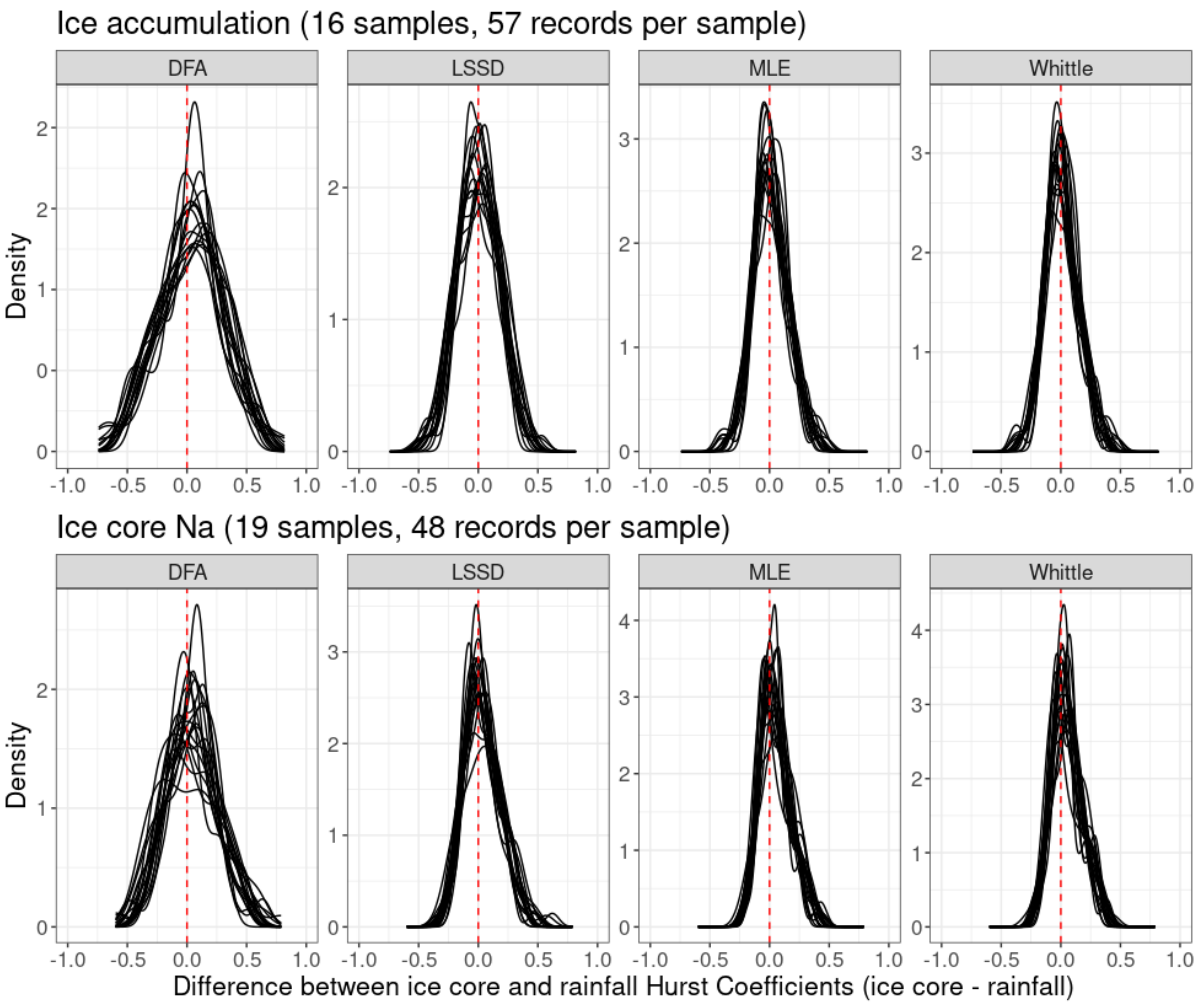

Figure 2-5: Difference distributions used for tests in Figure 2-4. Note that Figure 2-4 shows the mean and median of the difference distribution, shown here are the actual difference distributions.

Results presented in Figure 2-4 were derived from a sampling method designed specifically for this study. It was also of interest to see if a more generic comparison would return similar results. This involved calculating and comparing Hurst coefficients from the entire sample of ice core and rainfall records (i.e. no sub-sampling and sub-setting of rainfall data). Figure 2-6 shows that the results from this are similar to the results shown in Figure 2-4. Although similar results were returned, the sub-sampling approach is still useful for evaluating the sensitivity of the analysis to rainfall sample.

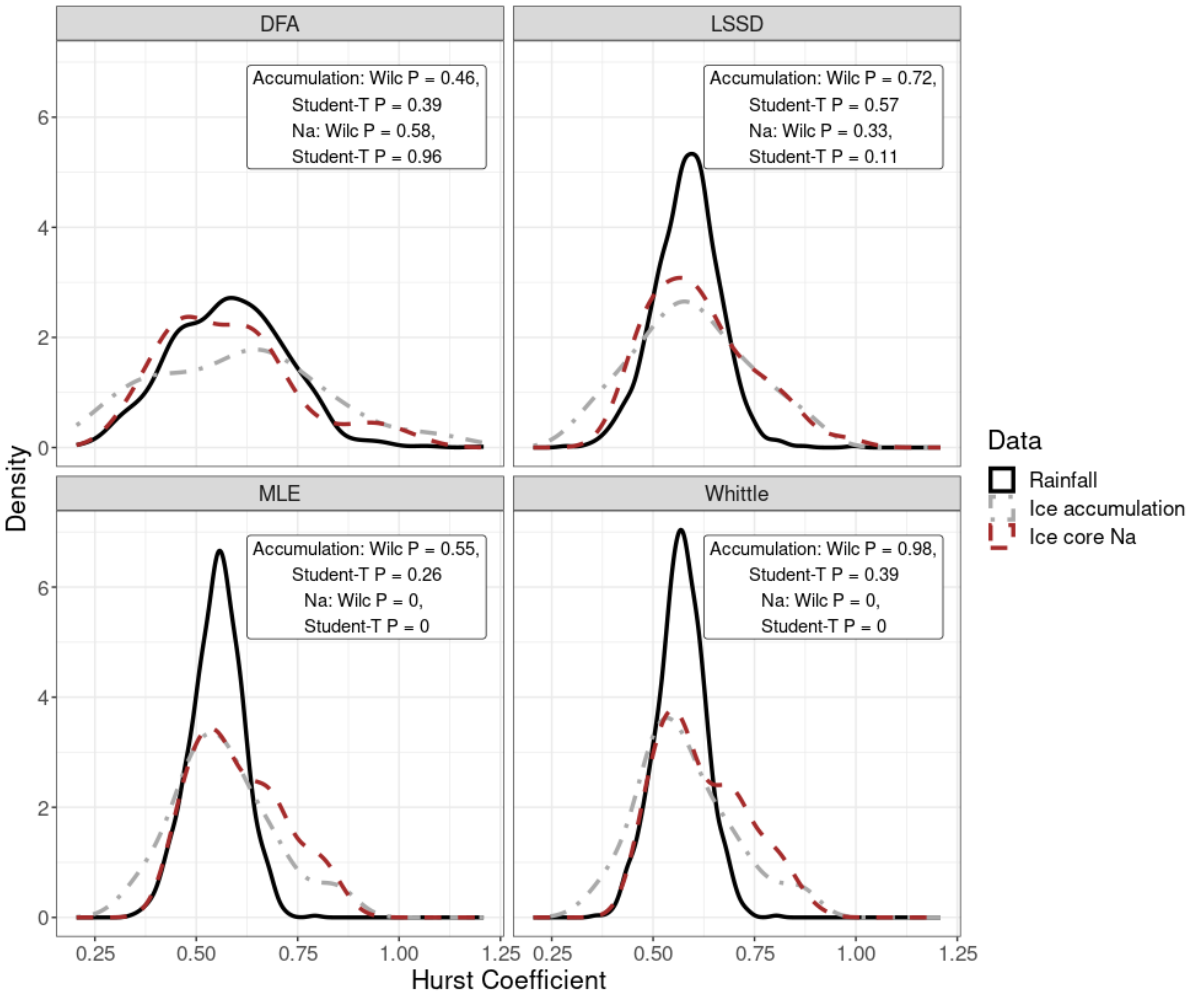

Figure 2-6: Comparison of ice core and rainfall Hurst coefficients, but without the sampling method used for Figure 2-4

A comparison of ice core and rainfall spectral power is shown in Figure 2-7. Compared with rainfall, both annual snowfall accumulation and ice core Na+ had similar spectral power over interannual frequencies (i.e. 3-7 years). However, ice core Na+ records had higher spectral power for the 10 to 20-year frequencies. The ice accumulation records also had higher spectral power for the 30 to 40-year frequencies.

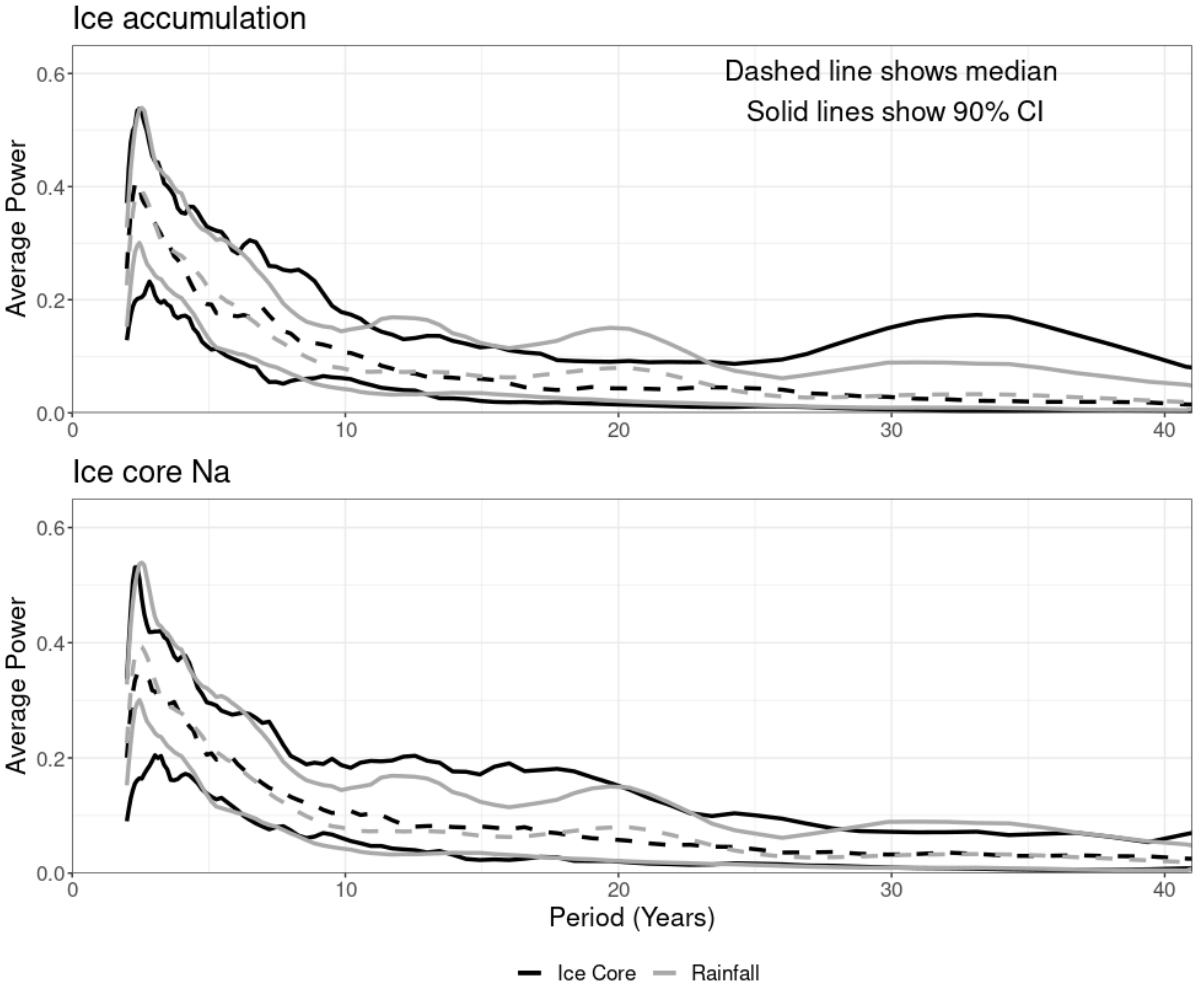

Figure 2-7: Wavelet average global power for rainfall and ice core datasets. Median (dashed line) and 90% confidence intervals (solid lines) are shown.

2.6 Discussion

The two key findings from this study are:

- Ice core accumulation records have similar persistence to mid-latitude rainfall records. This result was robust to rainfall sample used and persistence estimator.
- Ice core sea salt records have higher persistence than mid-latitude rainfall records. This result was sensitive to rainfall sample and persistence estimator.

Before discussing these findings further, note the key assumption made in this study: high and mid-latitudes exhibit similar hydroclimatic persistence. Various studies suggest that this assumption is valid (Fatichi et al., 2012; Iliopoulou et al., 2018; Markonis et al., 2018). This allows high-latitude ice cores and mid-latitude rainfall records to be compared fairly. However, these studies are still limited by the high sampling uncertainty that comes with estimating

persistence from ~100 years of data. This makes identifying any spatial coherence using instrumental measurements challenging because any potential signal could be dominated by noise.

The large sampling uncertainty of the Hurst coefficient also limited the statistical power of the analysis. Small differences in annual snowfall accumulation and rainfall persistence cannot be detected. However, other studies analysing the Hurst coefficient using different rainfall datasets have found a median value of ~0.55-0.6 (Fatichi et al., 2012; Iliopoulou et al., 2018; Taqqu et al., 1995; Tyralis et al., 2018), consistent with the annual snowfall accumulation records evaluated in this study (Figure 2-6). This also suggests that potential differences in ice core records are small compared with the sampling uncertainty inherent in persistence statistics. Future work should explore if these small differences significantly impact drought risk estimates.

With these limitations and assumptions in mind, this study shows that ice core accumulation records contain realistic representations of Southern Hemisphere hydroclimatic persistence. This indicates that, on average, post-deposition processes, additional non-hydroclimate influences, and statistical processing of collected cores does not bias the underlying persistence signal. However, these findings pertain to the ice core accumulation sample, not individual records. Biases may still be present in individual accumulation records.

These findings should be considered in conjunction with the key motivating factors for this study. These were (a) the potential benefits of including palaeoclimate information in climate risk assessments and water management and (b) the limited annually resolved, in-situ proxy records in the Southern Hemisphere (especially in Australia and Southern Africa). For ice core accumulation records, a realistic persistence signal, a clear link with regional hydroclimate, and much longer record lengths make them ideal for studying climate risks posed by low-frequency climate variability (particularly in the mid-latitude Southern Hemisphere). However, future work is needed to bridge the gap between the information we can reasonably extract from proxy records and the requirements of operational climate risk assessment (Galelli et al., 2021). Based on this study, any such work can use annual snowfall accumulation records with increased confidence in the corresponding persistence signal.

In contrast to the ice accumulation records, the ice core Na+ records had higher persistence than rainfall; what might cause this discrepancy? Na+ records undergo minimal statistical post-processing, which makes confounding climate signals the most likely cause.

With respect to potential confounding climate signals, ice core aerosol concentration (which include Na+) is negatively correlated with ice core oxygen isotope ratios, which are a temperature proxy (Buizert et al., 2015; Lambert et al., 2008). Markle *et al.* (2018) proposed that the mediating factor between ice core aerosols and oxygen isotope ratios is mid-latitude 'rainout' of aerosols – isotope ratios are high when temperature is high, high temperatures increase atmospheric moisture capacity (governed by Clausius-Clapeyron scaling) and increased moisture capacity leads to more aerosols being 'rained' out of the atmosphere in the mid-latitudes and less being aerosols being transported to Antarctica. Following this theory, lower oxygen isotope ratios and temperature means that more aerosols are transported to Antarctica, meaning temperature can influence ice core Na+. Because temperature timeseries exhibit larger persistence than rainfall (Franke et al., 2013a), this relationship could explain the study results. However, this proposed relationship occurs over much longer timescales than those considered in this study. Moreover, in the coastal high-resolution records used in this study, event-scale moisture intrusions to the Antarctic ice sheet from the mid-latitudes (e.g. Pohl et al., 2021; Wille et al., 2021) can overwhelm the annual water isotope signal (Jackson et al., 2023).

Na+ concentration and persistence can, potentially, be influenced by multiple climate variables. When considering these climate influences, what persistence structure might we expect? The answer is not immediately clear. Therefore, unlike annual snowfall accumulation, it is not appropriate to label a difference between ice core Na+ and rainfall persistence a 'bias'. Instead, ice core Na+ will likely contain a mixture of climate signals, similar to tree-rings.

Considering ice core Na+ records in water management will require a more in-depth understanding of the climate signal in individual records. There may be ice core Na+ records which are clearly linked to hydroclimate and accurately record hydroclimatic persistence. For example, in East Antarctica, clear physical links between synoptic patterns and ice core Na concentration have been demonstrated. In contrast, other ice core Na+ records may be primarily influenced by temperature, which would result in a different persistence structure. However, this requires extensive, site-specific analysis and modelling. In lieu of these specific analyses,

we are restricted to either (a) statistical analyses that correlate ice core Na+ with climate or (b) sample-scale evaluations (i.e. this study).

2.7 Conclusion

In this study, the hydroclimatic persistence signal in Antarctic ice core records was evaluated. We evaluated two ice core measurements – annual snowfall accumulation and ice core Na+ concentration. We found that ice accumulation records contain a relatively unbiased persistence signal. This indicates that the post-deposition processes and statistical processing of ice accumulation measurements do not confound the underlying persistence signal. In contrast, the ice core Na+ records typically overestimated hydroclimatic persistence. Reasons for this difference were outside the scope of this study, however, it seems likely that other climate variables (e.g. temperature) also influence Na+ persistence. Results from this study will guide future research on paleoclimate-informed climate risk assessment in the mid-latitude Southern Hemisphere

2.8 Links with following chapters

Because ice core records are, currently, the primary source of annually resolved, palaeoclimate information available for Australia, results from Chapter 2 will inform the incorporation of ice core information in water management-based climate risk assessment. Results from Chapter 2 are useful because any palaeoclimate-informed climate risk assessment should consider the fidelity of the climate signal contained in potential proxies. Therefore, results from Chapter 2 are used to justify the subsequent incorporation of ice core persistence in a palaeoclimate-informed stochastic modelling framework (Chapter 6).

Chapter 3. Evaluating different stochastic models using a global network of millennium-length hydroclimatic proxy records

3.1 Abstract

Stochastic models are used by water managers/hydrologists to generate long synthetic hydroclimate time series with statistics consistent with the input record (usually the instrumental record). These series contain droughts more severe than found in the instrumental record, making them useful for characterising drought risk. However, instrumental records are short (~100 years long) and contain limited information about low-frequency climate variability. This makes stochastic model validation with respect to low-frequency climate variability difficult. Hydroclimatic proxy records (e.g. tree rings and ice cores) are often several hundred years in length and, being representative of local/regional precipitation, provide an opportunity to validate different stochastic models using observed data of sufficient length to characterise low-frequency climate variability. In this study, we investigated the performance of nine commonly used stochastic models (AR(1), ARMA(1,1), ARFIMA(0,D,0), ARFIMA(1,D,0), Symmetric Moving Average, two and five-state Hidden Markov, k-Nearest Neighbour Bootstrap, and Wavelet Autoregressive models) in capturing different statistics related to low-frequency climate variability contained in proxy records. These models were validated on 45 millennium length hydroclimatic proxies located across both hemispheres. We found that (a) proxy data from the last 100 years (i.e. the instrumental period) is consistent with an AR(1) model; (b) stochastic models calibrated to the instrumental record do not reproduce pre-instrumental statistics; and (c) when calibrated to the entire proxy record, only the ARFIMA(0,D,0), ARFIMA(1,D,0), ARMA(1,1), Symmetric Moving Average, and five-state HMM models were able to reproduce statistics from the entire proxy record. Critically, the AR(1) model – widely used in operational hydrology - was unable to capture low-frequency climate variability when calibrated to the entire proxy record. This research highlights potential limitations associated with using stochastic models calibrated to the instrumental record to characterise baseline climate risk.

3.2 Introduction

Water supply should be resilient to drought. To design water supply systems that are drought resilient, drought risk must first be estimated. Drought risk can be viewed in terms of the drought hazard (e.g. the duration and severity of potential droughts), drought exposure (e.g. the probability of experiencing a drought), and drought vulnerability (e.g. the social, economic, and environmental impacts of a drought) (Kim et al., 2015). In this study, we evaluate key tools used by water managers to estimate drought hazard and exposure, which we refer to as 'drought risk'.

Estimating drought risk is difficult for three compounding reasons. First, measurements of rainfall and streamflow have only been taken since ~1900 (D. Jones et al., 2009; Menne et al., 2012). Second, hydroclimatic processes have, in a practical sense, random elements (Koutsoyiannis, 2010). This means that observed records represent just one possible realisation of past climate (McKinnon and Deser, 2021; Sivakumar, 2000). Third, at an annual timescale, hydroclimatic processes are persistent (i.e. wet and dry years tend cluster). This means that, in an instrumental record, only a few droughts are record, which reduces the effective sample size from which drought risk can be characterised (Hu et al., 2017; Koutsoyiannis and Montanari, 2007). For these three reasons, quantification of drought risk from a short, somewhat random record that contains a limited number of droughts is challenging.

To address short record length, randomness, and persistence when estimating drought risk, stochastic models can be used (Loucks and Van Beek, 2017). These models typically represent rainfall/streamflow as a weighted sum of previous timesteps plus random noise (Box et al., 1970). Stochastic models are first calibrated to instrumental measurements, then used to generate synthetic timeseries of arbitrary length (Matalas, 1967; Stedinger and Taylor, 1982b). These synthetic timeseries have similar statistics to the calibration data (i.e. instrumental measurements), but can contain droughts of greater severity. These synthetic timeseries can be used to characterise an extreme drought. However, when using a stochastic model to estimate drought risk, there is an implicit assumption that the instrumental record (i.e. the calibration data) is representative of the full range of climate variability that is possible. But, is this a valid assumption?

Estimating drought risk using an 'instrumental-period' stochastic model is further complicated by low-frequency (i.e., multi-decadal/centennial climate variability). Short instrumental records contain few multi-decadal cycles and no multi-centennial cycles. This means instrumental records are likely too short to (a) quantify the full range of natural climate variability (Vance et al., 2022); and, crucially, (b) validate a stochastic models ability to reproduce low-frequency variability. These limitations make it hard to (i) falsify simple stochastic models that do not account for low-frequency climate variability (Thyer et al., 2006); and (ii) justify the use of more complicated stochastic models that can account for low-frequency variability (Markonis et al., 2018). This inability to properly validate stochastic models highlights potential limitations with using a stochastic model calibrated to the instrumental record to characterise climate risk (Armstrong et al., 2020).

One option to overcome issues of short record length is to validate stochastic models using palaeoclimate proxy records. These records are taken from physical 'layers' with properties sensitive to climate (e.g., tree ring width being sensitive to available moisture), which are typically several hundred years in length (Griffin and Anchukaitis, 2014; Ho et al., 2015a; Verdon-Kidd et al., 2017). When examining the climate variability contained in these proxy records, they can either be mapped to a related climate variable via a statistical model (e.g., linear regression) or studied as is (Gangopadhyay et al., 2009; P. Jones et al., 2009; Razavi et al., 2015). However, although various studies have highlighted that palaeoclimate records contain signals that indicate droughts of greater severity than those experienced in the instrumental record, limited work has been conducted on using proxy records to validate stochastic models.

Koutsoyiannis (2003) and Iliopoulou *et al* (2018) demonstrate how palaeoclimate proxy data can be used to evaluate a stochastic model. Both studies demonstrated that the AR(1) model, which is commonly used in hydrology/water management, is unable to reproduce the autocorrelation structure of millennium-length hydroclimatic proxy timeseries. However, these studies were primarily focussed on explaining and validating a statistical framework that describes the Hurst phenomenon (a term for low-frequency climate variability). There is a need to validate different stochastic models using proxy data. In this study, we will build on insights from these studies and evaluate various stochastic models using several proxy records and statistics relevant to water resource management.

Although limited research has been conducted evaluating stochastic models using proxy data, various studies have used proxy data to (a) understand historic climate variability; and (b) inform water supply system adaptation. For example, reconstructions of southeast Australian climate variability have identified extended drought periods – some greater than 30 years, which is far longer than any instrumental record drought – that would significantly impact regional water security (Flack et al., 2020; Tozer et al., 2018; Vance et al., 2015). Moreover, when evaluating water supply systems using palaeoclimate informed streamflow scenarios in North America, various studies have demonstrated that system performance is unsatisfactory (meaning operational requirements were not met) and identified adaptations that improved system performance (Sauchyn et al., 2015; Tingstad et al., 2014).

All studies looking at palaeoclimate reconstructions are impacted by assumptions, uncertainties and potential biases introduced by the statistical reconstruction model. For example, underpinning all statistical reconstructions is an assumption that the proxy-climate relationship is stationary and can be inferred from the instrumental-period. However, numerous factors influence proxy properties; the relative influence of these factors may change between instrumental and pre-instrumental periods (D'Arrigo et al., 2008; Kiem et al., 2020). This means that the statistical model calibrated from the instrumental period may not be wholly suitable for some pre-instrumental periods.

Considering these limitations, for the purpose of evaluating stochastic models or making general inference about climate risk, it may be preferrable to use the original proxy records This alleviates potential biases introduced by a statistical reconstruction model and instead makes a more general assumption that the proxy record contains some hydroclimatic signal. Razavi *et al.* (2015) adopted this approach of examining the original proxy records when examining hydroclimatic non-stationarity in Canadian tree-ring chronologies. However, even with this approach, proxy records are still imperfect recorders of climate information (discussed in Section 3.3.2). As such, when making inference about past climate from proxy records there is an implicit assumption that any unexplained variability is not systematic with respect to the proxy records (i.e. the unexplained variability in instrumental and pre-instrumental periods is serially independent and has no positive or negative bias).

Regardless of whether statistical reconstructions or proxy records are examined, palaeoclimatology and stochastic models have an overlapping goal: characterising climate

variability. Palaeoclimatology pursues this goal by statistically modelling physical relationships between climate and climate proxies; stochastic models pursue this goal by statistically modelling the inherent randomness and persistence of the climate system. Considering the relative simplicity of stochastic models and the availability of independent proxy records that represent pre-instrumental climate variability, it is of interest to assess the ability of stochastic models to simulate pre-instrumental variability when calibrated to the instrumental record. It is also of interest to identify whether stochastic models can reproduce low-frequency climate variability captured by proxy data. As such, in this study we will explore the following research question:

1. Can stochastic models calibrated to the instrumental-period capture pre-instrumental variability?
2. Which stochastic models best reproduce low-frequency climate variability?

Both questions have important implications for the use of stochastic models to quantify baseline climate risk in water management. Addressing question 1 has implications for assessing the adequacy of current approaches that base the design of robust water supply systems on historic climate variability. Question 2 will help inform the future selection of stochastic models when accounting for climate variability in hydroclimatic risk analysis. This has implications for evaluating water supply system performance under pre-instrumental climate variability and any subsequent system adaptation.

3.3 Data

3.3.1 Proxy records used

There are numerous proxy records that could be used to validate stochastic models. In this study, we selected proxies based on the following criteria:

- Publicly available
- Described in a peer-reviewed journal article
- Annual temporal resolution
    - This allowed easier comparison with existing stochastic model validation studies – many of these studies use annual rainfall/streamflow data. Note that this effectively ruled out coarser resolution speleothem, lake sediment, and pollen proxy records.
- At least 1,000 years long

    - This ensured that proxy records covered periods associated with the Medieval climate anomaly and the Little Ice Age, plus some very large and well-dated volcanic eruptions that have had significant effects on lobal and regional climates.

This screening limited proxy types to be either tree-ring records, ice core snow accumulation, and ice core geochemistry records. Given the widespread availability of tree ring records, we applied additional screening criteria. These were:

- Tree ring records are made publicly available as a processed chronology or as a raw, unprocessed ring widths (referred to as 'rwl' format). If a potential record was available in rwl format, there had to be sufficient information in the relevant papers on how to standardise into a chronology.
- The chronology must have been used to reconstruct at least 1,000 years of hydroclimate. This ensured that the chronology had a statistically significant relationship with hydroclimate. Furthermore, for these chronologies, only the reconstruction period was used for stochastic model evaluation. This is because, when producing a tree-ring based reconstruction, the overall reconstruction period is selected based on the signal coherence of predictor chronologies and constituent tree samples (referred to as the Expressed Population Signal, or EPS - Wigley, Briffa and Jones (1984)). This ensures that only pre-instrumental periods with some common, and most likely climate related, signal were analysed.
    - However, some selected chronologies did not fully meet this criteria. Although the chronology was > 1,000 years in length, the associated climate reconstruction was not quite 1,000 years long. For these chronologies (morc014, oro062, nv516, fl001), the most recent 1,000 years was analysed. The selection criteria was relaxed for these chronologies to improve the spatial coverage of the dataset (e.g. North Africa, eastern North America).

We also included some Antarctic ice core geochemistry records - more specifically, Sodium (Na^+) concentration. These records were included because Antarctic and Southern Hemisphere mid-latitude hydroclimates are influenced by the same synoptic systems (Udy et al., 2021). For example, the Law Dome summer sea salt record and is correlated with east Australian rainfall, with east Australia sustaining a significant population (Kiem et al., 2020; van Ommen and

Morgan, 2010; Vance et al., 2015). As such, analysing these proxies still has implications for climate risk and stochastic modelling in urban water settings.

Figure 3-1 and Table 3-1 present the proxy records used in this study. In total, 45 proxy records were used: 25 tree-ring records, 5 ice accumulation records, and 15 ice core Na+ concentration records. All the tree-ring records were from the Northern Hemisphere - primarily North America (18), but with some in Asia (four), Europe (two), and Africa (1). In contrast, all ice core records were in the Southern Hemisphere. One snow accumulation record was in South America. The other four snow accumulation records and all Na+ records were in Antarctica.

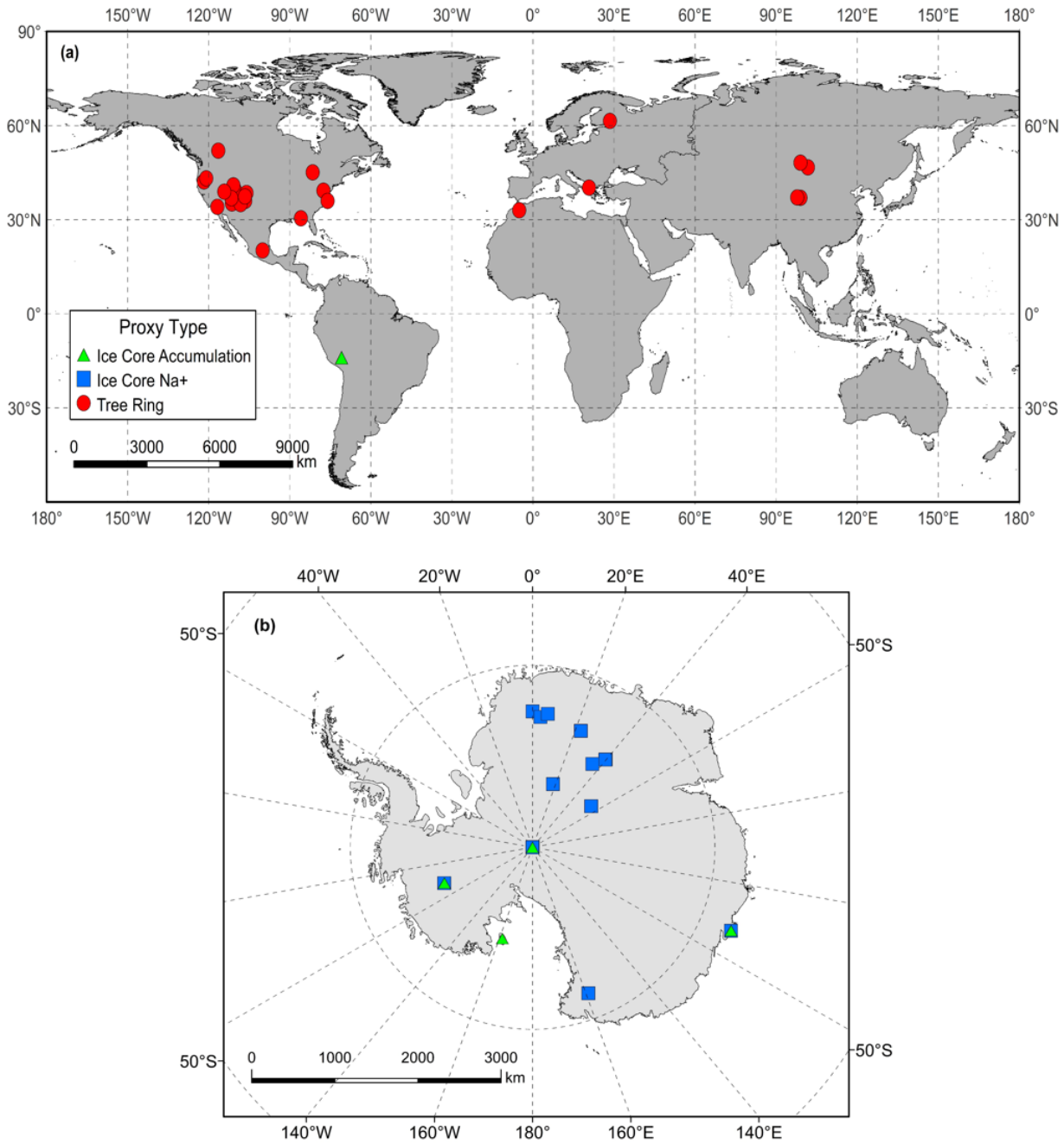

Figure 3-1: Location of proxy records used in this study

Many of the tree-ring proxies were publicly available as raw ring width series (rwl). These series contain age-related growth trends and various non-climatic signals (discussed further in Section 3.3.2). To produce a climate related tree-ring chronology, various pre-processing steps are needed. All tree-ring proxies presented in Table 3-1 were pre-processed following the methods outlined in the corresponding journal article using either ARSTAN software (available at https://www.ldeo.columbia.edu/tree-ring-laboratory/resources/software) or the R package 'dplR' (Bunn, 2008). For a more in-depth review of various millennium-length hydroclimatic tree-ring records, refer to Ljungqvist et al., 2020.

Table 3-1: Proxy records used in this study

Record	Continent	Period Analysed	Proxy Type	Reference	ITRDB Code
Dulan, China	Asia	159-1993	Tree Ring	Sheppard et al., 2004	chin006
Delingha, China	Asia	1000-2003	Tree Ring	Shao et al., 2005	chin050-chin054
Uurgat, Mongolia	Asia	488-2013	Tree Ring	Hessl et al., 2018	mong042
Khorgo, Mongolia	Asia	15-2014	Tree Ring	Hessl et al., 2018	mong041
Southern Finland	Europe	670-2012	Tree Ring	Helama, Meirläinen and Tuomenvirta, 2009	finl030-finl034
Mount Smolikas, Greece	Europe	730-2015	Tree Ring	Klippel et al., 2018	gree013-gree016
Flowerpot, Canada	North America	650-1989	Tree Ring	Buckley et al., 2004	NA
Whirlpool Point, Canada	North America	896-2008	Tree Ring	Case and MacDonald, 2003	cana220
Cedar Knob, USA	North America	950-1998	Tree Ring	Maxwell et al., 2011	wv005
Barranca de Amealco, Mexico	North America	880-2008	Tree Ring	Stahle et al., 2011	mexi047
Tavaputs Plateau, USA	North America	6-2005	Tree Ring	Knight, Meko and Baisan, 2010	ut530
Mount San Gorgonio, USA	North America	651-1998	Tree Ring	MacDonald, 2007	ca051
Southern Colorado Plateau, USA	North America	570-1990	Tree Ring	Salzer and Kipfmuller, 2005	az570
Jemez Mountains, USA	North America	824-2007	Tree Ring	Touchan et al., 2011	nm583
Upper Arkansas Basin, USA	North America	216-2007	Tree Ring	Woodhouse, Pederson and Gray, 2011	Multiple
Upper Klamath Basin, USA	North America	1000-2010	Tree Ring	Malevich, Woodhouse and Meko, 2013	or093
El Malpais,USA	North America	5-2004	Tree Ring	Stahle et al., 2009	nm580
Bear River, USA	North America	916-2013	Tree Ring	DeRose et al., 2015	ut541
Summitville, USA	North America	10-2009	Tree Ring	Routson, Woodhouse and Overpeck, 2011	co656
Atlas Mountains	Africa	985-1984	Tree Ring	Esper et al., 2007	morc014

Choctawhatchee River	North America	993-1992	Tree Ring	Stahle et al., 2012	fl001
Lee's Ferry	North America	760-2005	Tree Ring	Meko et al., 2007	ut529
Colorado River	North America	985-1984	Tree Ring	MacDonald, Kremenetski and Hidalgo, 2008	nv516
Sacramento River	North America	997-1996	Tree Ring	MacDonald, Kremenetski and Hidalgo, 2008	or062
Albermarle Sound	North America	934-1985	Tree Ring	Stahle, Burnette and Stahle, 2013	va021
Law Dome Snowfall	Antarctica	17-2016	Ice Core Accumulation	Jong et al., 2022	NA
Roosevelt Island	Antarctica	13-2012	Ice Core Accumulation	Winstrup et al., 2019	NA
West Antarctic Ice Sheet Divide	Antarctica	8-2007	Ice Core Accumulation	Sigl et al., 2016	NA
SPICE Snowfall	Antarctica	15-2014	Ice Core Accumulation	Winski et al., 2019	NA
Quelccaya Ice Core	South America	683-2009	Ice Core Accumulation	Thompson et al., 2013	NA
Law Dome Sea Salt	Antarctica	7-2016	Ice Core Na	Jong et al., 2022	NA
DF01	Antarctica	607-1903	Ice Core Na	Motizuki et al., 2017	NA
DFS10	Antarctica	10-2009	Ice Core Na	Sigl et al., 2014	NA
DML05	Antarctica	150-1998	Ice Core Na	Traufetter et al., 2004	NA
DML07	Antarctica	454-1996	Ice Core Na	Traufetter et al., 2004	NA
DML17C98_33B33	Antarctica	0-1996	Ice Core Na	Traufetter et al., 2004	NA
NUS072	Antarctica	336-1993	Ice Core Na	Pasteris et al., 2014; Sigl et al., 2014	NA
NUS075	Antarctica	0-1982	Ice Core Na	Pasteris et al., 2014; Sigl et al., 2014	NA
NUS077	Antarctica	8-2007	Ice Core Na	Pasteris et al., 2014; Sigl et al., 2014	NA
NUS085	Antarctica	346-2000	Ice Core Na	Pasteris et al., 2014; Sigl et al., 2014	NA
SP01	Antarctica	905-2000	Ice Core Na	Budner and Cole-Dai, 2003	NA
SP04C5	Antarctica	176-2004	Ice Core Na	Ferris et al., 2011	NA
SPICE Sea Salt	Antarctica	15-2014	Ice Core Na	Thomas et al., 2023	NA

TD05	Antarctica	542-1986	Ice Core Na	Severi et al., 2017	NA
WDC06A	Antarctica	5-2004	Ice Core Na	Sigl et al., 2015	NA

Note that for records longer than 2000 years, analysis was only conducted on the most recent 2000 years, in line with various PAGES2K projects (Emile-Geay et al., 2017).

3.3.2 A cautionary note on the use of proxy data for stochastic model evaluation

In using proxy records as a substitute for observed climate data when evaluating stochastic models, two key assumptions are made:

1. There are no systematic biases in the proxy record with respect to hydroclimate.
2. The various pre-processing steps involved in the creation of proxy records do not distort the underlying climate signal.

Any violation of these assumptions would confound subsequent inferences about stochastic model performance and climate risk. Although the proxy data types used in this study (i.e. tree rings and ice cores) undeniably contain information about historic climate variability, potential violations of both assumptions will now be discussed.

Regarding tree-rings and assumption (1), there are numerous confounding factors that can bias the relationship between tree-ring width and hydroclimate. For example, age-related trends (i.e., the tendency for younger trees to produce wider rings) must be accounted for prior to any climate inference via a process of tree-ring standardisation. Standardisation, which is required when extracting climate information from tree-rings, can also remove information about low-frequency climate variability and introduce additional uncertainty in the proxy-climate relationship (Büntgen et al., 2021; Cook et al., 1995). Aside from age-related trends, integration of climate conditions over multiple years (Meko, 1997, p. 199), low-frequency variations arising from stand dynamics and canopy competition (Cook, 1985), and confounding temperature signals (Franke et al., 2013b; Ludescher et al., 2020; Yuan et al., 2021) can result in tree-ring records displaying greater persistence than corresponding hydroclimate (note that temperature exhibits larger persistence than hydroclimatic variables).

To remove the increased persistence in tree-ring records a statistical technique called pre-whitening is often used. Pre-whitening involves fitting an autoregressive model to the tree-ring

record and using model residuals for subsequent analysis (Hamed, 2009; Yue et al., 2002). However, pre-whitening will remove low-frequency climate information. For example, Razavi and Vogel, (2018) demonstrated that pre-whitened tree-ring width series will underestimate the magnitude and duration of droughts/pluvials. Iliopoulou *et al.* (2018) also demonstrated that, compared with instrumental data and other proxy records, pre-whitened tree-ring records had consistently lower Hurst coefficients (indicating less persistence). Therefore, pre-whitened chronologies should be viewed as a record of high-frequency, interannual climate variability.

In essence, when using tree-ring records for stochastic model evaluation, it is important to note that (a) tree-ring records *may* have higher persistence than corresponding hydroclimate and (b) attempts to remove this higher persistence can also remove low-frequency climate information. These issues have clear implications for this study.

Even with these factors, all tree-ring records used in this study have been used to reconstruct hydroclimate, with the corresponding standardisation method peer reviewed. Furthermore, all study records are comprised of either multiple chronologies or numerous tree samples, meaning that a coherent and replicated hydroclimatic signal was identified. Also, the tree-ring records used in this study were produced using various statistical methods. At the very least, this accounted for the methodological uncertainties introduced by pre-processing and, for this study, a "good" stochastic model should be robust to this uncertainty.

Compared with tree-ring records, systematic biases and potential pre-processing issues in snow accumulation records have been less explored. Nevertheless, several caveats should be mentioned when using snow accumulation as a proxy for hydroclimate. For example, wind deposition and erosion of initial snowfall can confound a regional snowfall signal (Thomas et al., 2017). This can be overcome by taking measurements from multiple ice cores across a region and calculating a composite accumulation record, averaging deposition and erosion disturbances (Wang et al., 2017). However, it's typically infeasible to drill multiple ice cores that contain millennium-scale information. Therefore, when drilling an extended core, additional, shorter cores that cover the instrumental-period are often drilled to evaluate the regional consistency of the extended core.

The advection and thinning of ice core layers over time can also confound inferences about pre-instrumental hydroclimate. However, ice core drilling sites with minimal advection are

typically chosen. In contrast, ice core thinning (occurs due to downward pressure from recent snowfall and horizontal movement of the ice sheet (Dansgaard and Johnsen, 1969)), is often corrected via a mathematical model. Model choice can influence subsequent proxy measurements. However, relative to the magnitude of hydroclimatic variability suggested by ice core records, these differences are small (Roberts et al., 2015), and thinning is negligible for the first few centuries of the ice core.

Unlike tree-rings and ice accumulation, the links between ice core Na+ and hydroclimate have not been explored in detail (Thomas et al., 2023). Ice core Na+ is initially wind-scoured from the open ocean or from 'frost flowers' (brine crystals) formed on the surface of sea ice. The sea salt aerosol is then deposited at the ice core site, primarily via snowfall (but with some dry deposition) (Wolff et al., 2006). This makes wind and atmospheric circulation a mediating variable between ice core Na+ and hydroclimate. Naturally, there are many local and broad-scale physical processes that can influence wind and atmospheric circulation, which may influence the wind-mediated sea salt aerosol production, transport and deposition. For example, global temperature will change the moisture holding capacity of the atmosphere (Visser et al., 2022; Wasko and Sharma, 2015; Westra et al., 2013). Higher temperatures can lead to increased moisture holding capacity and rainfall, which in turn removes aerosols (e.g. Na+) from the atmosphere before they can reach Antarctica (Markle et al., 2018). This highlights that variability in Na+ records can contain a complex mix of different signals (similar to tree-rings).

Despite the multiple, non-hydroclimatic factors that can influence ice core Na+, recent studies have also demonstrated clear links between some Na+ records and regional rainfall. For example, Southern Ocean synoptic systems are quite large, spanning from Antarctica to the mid-latitudes (Udy et al., 2021). Some commonly occurring Southern Ocean synoptic systems are conducive to increased (reduced) salt deposition in east Antarctic and increased (reduced) rainfall in east Australia (Udy et al., 2022). This highlights that Na+ is a potentially reasonable proxy for past hydroclimate. The potential for Na+ to be a reasonable hydroclimatic proxy, combined with a lack of alternative long-term Southern Hemisphere proxy records, were the key reason for the inclusion of Na+ records in this study.

It is important to emphasise that despite the issues associated with different proxy types, proxy records are the best available source of information we have for understanding long-term

climate variability. Multiple lines of evidence from different proxies indicate that pre-instrumental climate, and by extension future climate, may be very different to instrumental climate. Therefore, to better understand potential climate risks to water security, it is important to study proxy variability within a climate risk and water management context. Evaluating stochastic model performance using proxy records is just one way to do so.

3.4 Stochastic Models

There are a wide variety of stochastic models that can be used to simulate timeseries. In this study, we validated a small - but representative - subset of these models. The broad classes of assessed models were Autoregressive Moving Average (ARMA) models, long-term persistence models (also referred to as the Autoregressive Fractionally Integrated Moving Average models, or 'ARFIMA'), Hidden Markov models; non-parametric models, Symmetric Moving Average (SMA) models, and "Component-signal ARMA" models (which will be explained in Section 3.5).

Table 3-2 displays the models used in this study. In total, nine different models from six different model classes were evaluated. More in-depth descriptions of these models (and the corresponding general model class) are included in Sections 3.4.1-3.4.6.

Table 3-2: Stochastic models validated in this study.

Model	Model Class	References
AR(1) or ARMA(1,0)	ARMA	Box et al., 1970
ARMA(1,1)	ARMA	Box et al., 1970
ARFIMA(0,D,0)	Long-term persistence	Granger and Joyeux, 1980; Hosking, 1984
ARFIMA(1,D,0)	Long-term persistence	Hosking, 1984; Montanari et al., 1997
2-State HMM	Hidden Markov	Visser, 2011
5-State HMM	Hidden Markov	Visser, 2011
KNN bootstrap	Non-parametric	Lall and Sharma, 1996
Symmetric Moving Average to Anything	Symmetric Moving Average	Koutsoyiannis, 2000; Tsoukalas et al., 2018
Wavelet-Autoregressive	Component-signal ARMA	Kwon et al., 2007

3.4.1 Autoregressive Moving Average (ARMA) models

Historically, ARMA type models have been used throughout hydrology and water management. Originally developed by (Box et al., 1970), these models are essentially linear regression models; timeseries values are estimated using weighted sums of previous values and/or residuals as weighted predictors. Once a mean estimate is obtained from the regression model, random noise (sampled from the regression residual distribution) is added. The "autoregressive" in ARMA refers to models which use previous values (i.e., the expected value plus a residual) as predictors, whereas "moving average" refers to models which use previous residuals. Because various combinations of previous values/residuals can be used as predictors, specific types of ARMA model are denoted as ARMA(p,q), with 'p' indicating the number of previous values used as predictors and 'q' denoting the number of previous residuals.

In this study, two different ARMA variants were evaluated - an ARMA(1,0) and an ARMA(1,1). The ARMA(1,0), also referred to as an AR(1) or Lag-1 Autoregression model, is used commonly in operational water management. Equations for respective models are shown below - note that all ARMA models assume the data is normally distributed.

$$y_t = \mu + \phi(y_{t-1} - \mu) + \epsilon_t \qquad \text{Equation 3-1}$$

$$y_t = \mu + \phi(y_{t-1} - \mu) + \theta\epsilon_{t-1} + \epsilon_t \qquad \text{Equation 3-2}$$

Where:

$$\epsilon_t \sim N(0, \sigma) \qquad \text{Equation 3-3}$$

Although it is possible to identify and select ARMA models with higher p/q lags, in this study we wanted to focus on ARMA models that have been applied in the water resources industry and shown to successfully reproduce instrumental-period hydroclimate. This enabled an assessment of whether such model structures are appropriate for use in characterising climate risk under observed low-frequency variability.

Given that ARMA models assume the data is normally distributed, the marginal distributions for each proxy record were tested for normality using the Shapiro-Wilks test prior to parameter estimation. If a Shapiro-Wilks test returned a p-value less than 0.05, the data were transformed using a Box-Cox transformation (Box and Cox, 1964). The optimal transformation was identified via Maximum Likelihood Estimation using the R package 'forecast' (Hyndman and Khandakar, 2007). Once replicates of the transformed timeseries were generated, the transformation was reversed (which mapped the transformed data back to the original units).

3.4.2 Long-term persistence models

Considering that current values are estimated as a function of recent previous values, ARMA models are sometimes referred to as "short-term persistence" models. However, hydrological timeseries typically display dependence across extended time periods (e.g., decades/centuries). This dependence has been referred to as long-term persistence, multidecadal/centennial climate variability, or the "Hurst phenomenon". Stochastic models designed to reproduce this phenomenon can be loosely classed as long-term persistence models.

Long-term persistence models were initially developed in response to findings by Hurst (1951). By studying various extended geophysical timeseries - including Nile River flood maximums - Hurst noted a tendency for high/low values to cluster in a way that could not be explained by independent and identically distributed random models, nor any periodicity. This was subsequently called the Hurst phenomenon.

Since the initial discovery, various mathematical frameworks of the Hurst phenomenon have been proposed (Hosking, 1984; Mandelbrot, 1971; Mandelbrot and Wallis, 1969). These typically involve estimating a dimensionless constant that describes how timeseries variance increases with temporal scale. This constant is sometimes referred to as the "Hurst coefficient". Independent and Markovian timeseries have a Hurst coefficient of 0.5, timeseries exhibiting long-term persistence have a Hurst coefficient > 0.5 (Koutsoyiannis, 2002).

Several stochastic models that explicitly model long-term persistence have also been proposed since the initial findings of Hurst. These include Fractional Gaussian Noise models (Mandelbrot, 1971); non-stationary mean models (Boes and Salas, 1978); Broken-Line models (Mejia et al., 1972); and ARFIMA models (Hosking, 1984). For a general history on the

discovery of the Hurst phenomenon and the development of associated models, refer to Graves *et al.* (2017).

In this study, the ARFIMA model class was selected as a representative class of long-term persistence models. Initially proposed by Granger and Joyeux (1980), ARFIMA models extend ARMA type models to also consider long-term persistence. A timeseries y_t that cannot be modelled as a stationary ARMA process can be fractionally differenced using the following equation:

$$y_t = (1 - B)^D \, y_t \qquad \text{Equation 3-4}$$

Where B is the backshift operator such that:

$$B^k = y_{t-k} \qquad \text{Equation 3-5}$$

And

$$(1 - B)^D = \sum_{k=0}^{\infty} \binom{D}{k} (-B)^k \qquad \text{Equation 3-6}$$

The fractionally-differenced timeseries y_t is then modelled as an ARMA(p,q) process with the final model structure denoted as ARFIMA(p, D, q) (with D being the fractional differencing parameter). Note how the differencing in Equation 3-6 results in the current value being explicitly influenced by all preceding values.

In this study, two different ARFIMA model structures were validated - an ARFIMA(0,D,0) model and an ARFIMA(1,D,0) model. An ARFIMA(0,D,0) model does not consider any short-term timeseries persistence - it only models long-term persistence via the 'D' parameter. This model is similar to the Fractional Gaussian Noise (FGN) model, although it is derived from a slightly different philosophical perspective. The 'D' parameter can be used to derive the FGN 'H' parameter (which also describes long-term persistence) via H=D+0.5 (Graves et al., 2017).

In contrast to ARFIMA(0,D,0), the ARFIMA(1,D,0) model considers both long and short-term persistence (Montanari et al., 1997). Although ARFIMA models with higher p and q lags could have been fitted, we elected for the ARFIMA(1,D,0) model to (a) avoid overfitting and (b) replicate existing stochastic model evaluations studies.

Models were calibrated and subsequent replicates generated using the R package 'fracdiff' (Fraley et al., 2012). As with the ARMA models, ARFIMA models assume normally distributed residuals. Therefore, a Box-Cox transformation was applied prior to modelling non-normal proxy records.

3.4.3 Hidden Markov models

Compared with ARMA-type models, Hidden Markov Models (HMMs) offer an alternative approach to modelling interannual and multidecadal climate variability. HMMs are serially dependent mixture models; a mixture model means that the data is drawn from multiple distributions, each with different parameters (Visser, 2011). The different distributions/parameters can be used to represent relative climate state (i.e., wet/dry), and the serial dependence between these states is represented by a transition matrix (Thyer and Kuczera, 2000). The transition matrix contains the probability that the successor to the current value will remain in the same state as the current value, or transition to some other state. Transition probabilities can be stationary or time varying (Hughes et al., 1999) and the number of states can be user or determined by the (Gershman and Blei, 2012; Lambert et al., 2003).

In this study, HMMs with two and five hidden states were evaluated. Distributions for each state were Gaussian, with transition matrices and distribution parameters identified using the Expectation Maximisation (EM) algorithm via the R package 'depmixS4' (Visser and Speekenbrink, 2010).

3.4.4 Non-parametric models

In contrast to parametric models (such as ARMA/HMMs), non-parametric stochastic models make no underlying assumptions about the structure of the data (Lall and Sharma, 1996). This makes non-parametric methods suitable for replicating non-stationary time series (Sharma et al., 1997). Most stochastic non-parametric models are variants of a K-Nearest Neighbour (KNN) bootstrap (Lall and Sharma, 1996; Yates et al., 2003). This bootstrapping approach uses a weighted kernel to assign sampling probabilities to the K-nearest neighbours of each

observation (with neighbours identified based on Euclidean or Mahalanobis distance). For each sample, the successor value is selected based on the successor of the K-nearest samples.

3.4.5 Symmetric Moving Average models

Symmetric Moving Average (SMA) models are another type of stochastic model. Originally proposed by Koutsoyiannis (2000), these models involve two general steps:

1. Estimating the autocovariance function. Typically, this is done by fitting a Cauchy-like distribution to the empirical autocovariance function (ACF) - see Equation 3-7, where ρ(t) is the estimated autocorrelation at lag *t*. These distributions can reproduce a wide variety of positive and monotonically decreasing ACFs (typical of hydroclimatic timeseries), including those arising from short and long-memory processes.

$$\rho_{(t)} = (1 + k\beta t)^{-1/\beta} \qquad \text{Equation 3-7}$$

2. Simulating timeseries that preserve this ACF using a symmetric moving average generating scheme. Theoretically, this scheme estimates timeseries values as a weighted sum of infinite previous/future randomly drawn innovations (these innovations can also preserve the mean and skew of the target timeseries). In practice, only a finite number of innovations (e.g. 500) are used to estimate timeseries values (this only has a slight impact on the ability of the model to reproduce the autocovariance function). The innovation weights are estimated from the corresponding ACF, which ensures that the ACF is reproduced in subsequent timeseries replicates.

Recent work by Tsoukalas et al. (2018) has enabled the SMA model to be applied to timeseries with any marginal distribution/autocovariance structure. This approach initially generates timeseries replicates from a standard normal distribution using an SMA generating scheme. Then, a Nataf transformation is used to reproduce a desired marginal distribution. Because these transformations are non-linear (which distorts the original autocovariance function), the ACF of the standard normal distribution being modelled is estimated as a function of the desired ACF. As with the original SMA model, the desired ACF can be estimated from the observed timeseries. In doing so, the original ACF is reproduced after the Nataf transformation.

In this study, we used the SMA model of Tsoukalas et al. (2018). An appropriate marginal distribution was selected by identifying Maximum Likelihood parameters from Pearson type distributions I-VII and selecting the distribution which returned the smallest Akaike Information Criterion (AIC). These distributions were selected because they subsume distributions typically used in stochastic modelling (e.g., gamma, normal distribution). After an optimal Pearson type distribution was selected, a Cauchy-like function (Equation 7) was fitted to the empirical ACF by minimising the Mean Absolute Error. Timeseries replicates were then generated using the R package 'anySim' (Tsoukalas et al., 2020).

3.4.6 Component-signal ARMA models

"Component-signal" ARMA models refer to models that decompose a time series into orthogonal components, calibrate separate ARMA models to these components; generate replicates of these components, and then recombine replicate signals into a final timeseries. By modelling both high and low-frequency signals separately, these models are better equipped to reproduce low-frequency variability in climate timeseries. These component signals can be extracted in several ways, such wavelet decomposition (Kwon et al., 2007; Nowak et al., 2011) or Empirical Mode Decomposition (Lee and Ouarda, 2012; McMahon et al., 2008).

In this study, we evaluated the Wavelet-Autoregressive Model (WARM) of Kwon et al. (2007). Significant frequencies were identified using a Morlet wavelet transformation (Torrence and Compo, 1998). Significant wavelet frequencies were identified by comparing the observed global power spectrum with the global power spectrums derived from 1,000 AR(1) replicates of the corresponding proxy record. Frequencies were considered significant if they had a p-value of 0.05 with respect to the AR(1) replicates. Component signals were then modelled as an AR(p) process, with the optimal lag identified using the AIC. In this study, optimal parameter values for each lag were identified by maximising the log-likelihood using the Nelder-Mead method (Nelder and Mead, 1965).

Some minor modifications from the original method were necessary to improve model performance:

- Initial evaluation found that WARM replicates did not reproduce the skew of the marginal distribution. This is because component signals are (a) assumed to be orthogonal, resulting in orthogonal timeseries replicates; and (b) the AR(p) model assumes a marginal Gaussian distribution. The sum of independent Gaussian variables

produces a Gaussian variable (which has a skew of zero). As such, a Box-Cox transformation was applied to non-normal proxy records prior to fitting the model, with the transformation reversed after replicates were generated.

- For many low-frequency component signals (which are highly smoothed), calibrating high lag AR(p) models typically resulted in optimiser degeneracy. When the optimiser did not fail, high lag AR(p) models offered minimal, if any, improvement over more parsimonious models. Therefore, in order to reduce model complexity and simplify the model calibration, a forward stepwise selection approach was used. For each component signal, this involved iteratively increasing the lag of the AR(p) model being fitted and only accepting the model if it produced an Akaike Information Criterion value smaller than the previous lag. If not, the current model is selected for subsequent replicate generation.
- In most cases, wavelet signals were not orthogonal (i.e., they were correlated). In such cases, recombining correlated component signals assuming independence will produce a timeseries with reduced variance (relative to the original timeseries) (Nowak et al., 2011). To recover this lost variance, the replicate timeseries were scaled to match the variance of the original timeseries. Although various multivariate stochastic models could explicitly preserve this correlation (and, by extension, the variance of the recombined signals), we considered such modifications to the original WARM as too extensive for the purpose of this study.

3.5 Methods

Three different calibration/validation experiments were conducted for all models and proxy records. For all experiments, the "instrumental-period" refers to the most recent 100-years of the proxy record. This is a simplifying assumption – instrumental records often vary in length and the final 100 years of each proxy record may not perfectly overlap with the corresponding instrumental record. For each model/record/experiment, 1,000 stochastic replicates of equal length to the corresponding validation period were generated.

The following experiments were conducted:

- Experiment 1: Calibration/validation on the instrumental-period of each proxy record. This experiment replicates a stochastic model calibration study where only instrumental data is available.

- Experiment 2a: Calibration on the instrumental-period of each proxy record, validation on the full pre-instrumental period of each proxy record. This experiment identifies whether the instrumental-period contains sufficient information for respective models to reproduce pre-instrumental low-frequency climate variability. Note that because different records have different lengths, validation period lengths are not consistent across respective records.
- Experiment 2b: Calibration on the instrumental-period of each proxy record, validation on the most recent 400-years in the pre-instrumental period.
    - Results from Experiment 2 were then contrasted with results from Experiment 1. This identified whether the high performing models from Experiment 1 were also able to reproduce pre-instrumental climate variability.
- Experiment 3: Calibration/validation on the full proxy record.
    - Results from Experiment 3 were then contrasted with Experiment 1 and Experiment 2. This identified (a) if models that perform well during the instrumental-period also perform well for longer records and (b) if poor performing models from Experiment 2 were able to reproduce low-frequency climate variability when provided with more calibration data.

For each experiment, models were evaluated based on their ability to reproduce validation data statistics across all proxy records. For each model/record combination, this involved calculating the corresponding statistic for each stochastic replicate (i.e., the sampling distribution was derived). The percentile rank of the validation statistic was then calculated with respect to the stochastic sampling distribution. A statistic was considered reproduced, or "captured", if the percentile rank was greater than 5 and less than 95 (i.e., the statistic was within the 90% confidence intervals of the sampling distribution). For each model, the proportion of records for which the validation statistic was captured, had a percentile rank < 5 (indicating that the stochastic model overestimated the statistic), or had a percentile rank > 95 (indicating that the stochastic model underestimated the statistic) was calculated.

A limited set of statistics was used for model evaluation. These statistics are typically used in stochastic model evaluation studies or are related to low-frequency climate extremes; both of which are of interest to climate risk analysts and water resource managers. These statistics included:

- Mean

- Lag-1 Autocorrelation
- Standard Deviation
- Hurst Coefficient
    - In this study, the Whittle estimator of the Hurst coefficient was used (Beran, 2017). Although different methods can yield very different Hurst coefficients for short timeseries (e.g. 100 values or less), Weron (2002) demonstrated that for timeseries longer than 1000 years this uncertainty is significantly reduced, regardless of estimator. As such, this choice is not expected to influence the results of Experiments 2a/b and 3 but could influence Experiment 1. However, Chapter 2 found that the Whittle estimator is suited for smaller samples.
- Skew
- Minimum
- Maximum
- Minimum and maximum cumulative sums for overlapping 2, 5, 10, 30, 50, and 100-year windows.
    - Due to the limited record length, for Experiment 1 the 50 and 100-year minimum/maximum cumulative sums were not used as validation statistics.

A schematic of the method is Figure 3-2, which highlights how results for individual records, statistics, and models were aggregated for the different experiments.

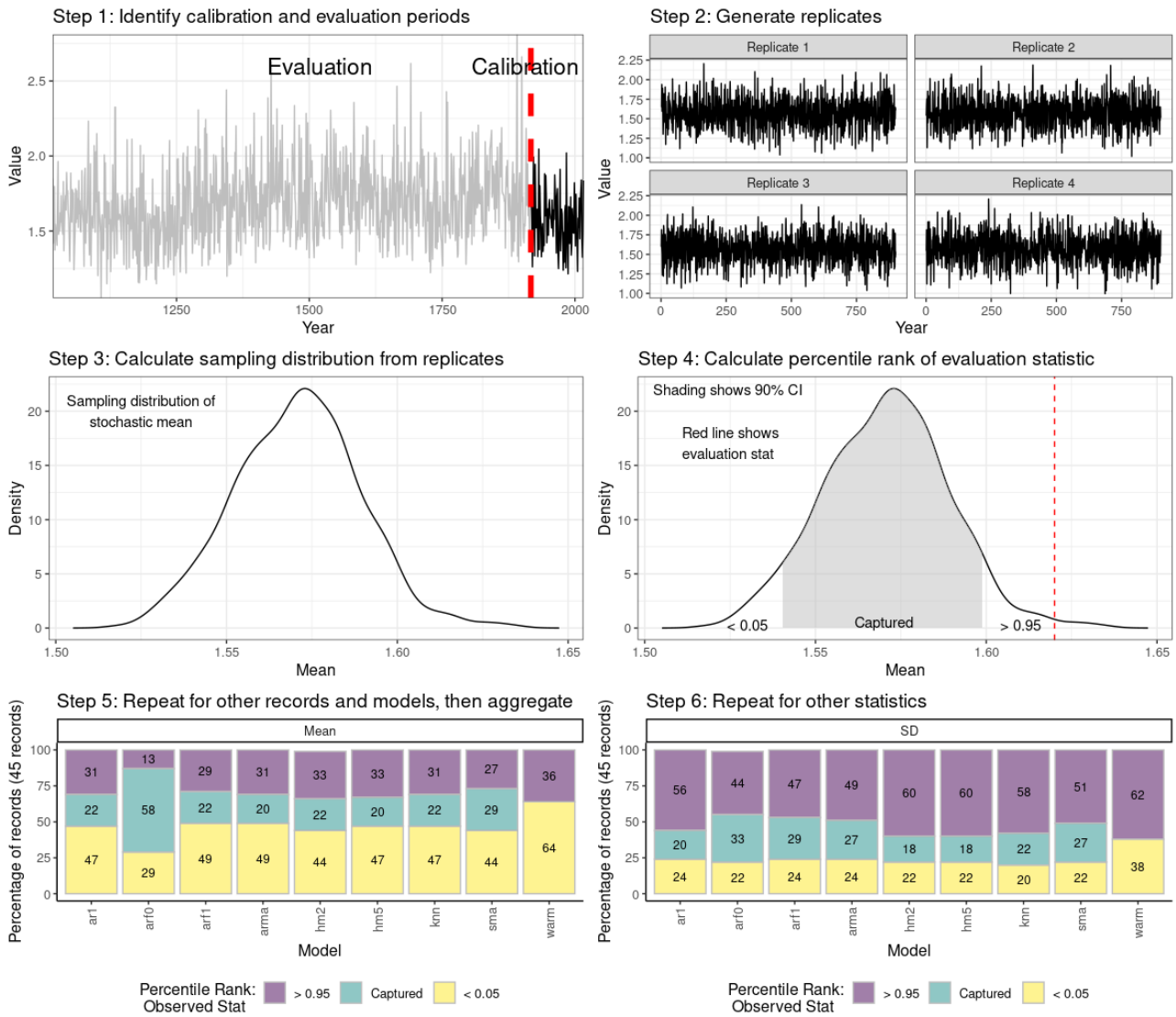

Figure 3-2: Schematic of method

3.6 Results

3.6.1 Experiment 1 – Calibrating and validating on the instrumental-period

Results from Experiment 1 (i.e., calibration/validation on the instrumental-period of the proxy records) are shown in Figure 3-3. We can see that all models were consistently able to reproduce all statistics. Note that the AR(1) model typically used in water management performed well over the instrumental-period.

Although the plot indicates that the KNN model was unable to reproduce the observed maximum, this is misleading. Because the KNN model is non-parametric it does not extrapolate beyond the calibration period. This means that, in most cases, the stochastic replicate maximum

is the observed maximum. The percentile rank is calculated as the proportion of replicate maximums greater than or equal to the observed maximum - because the KNN model cannot extrapolate, this results in a percentile rank of 1. So even though the model is reproducing the maximum, this is not reflected in the percentile rank. This issue is also apparent in Experiment 3.

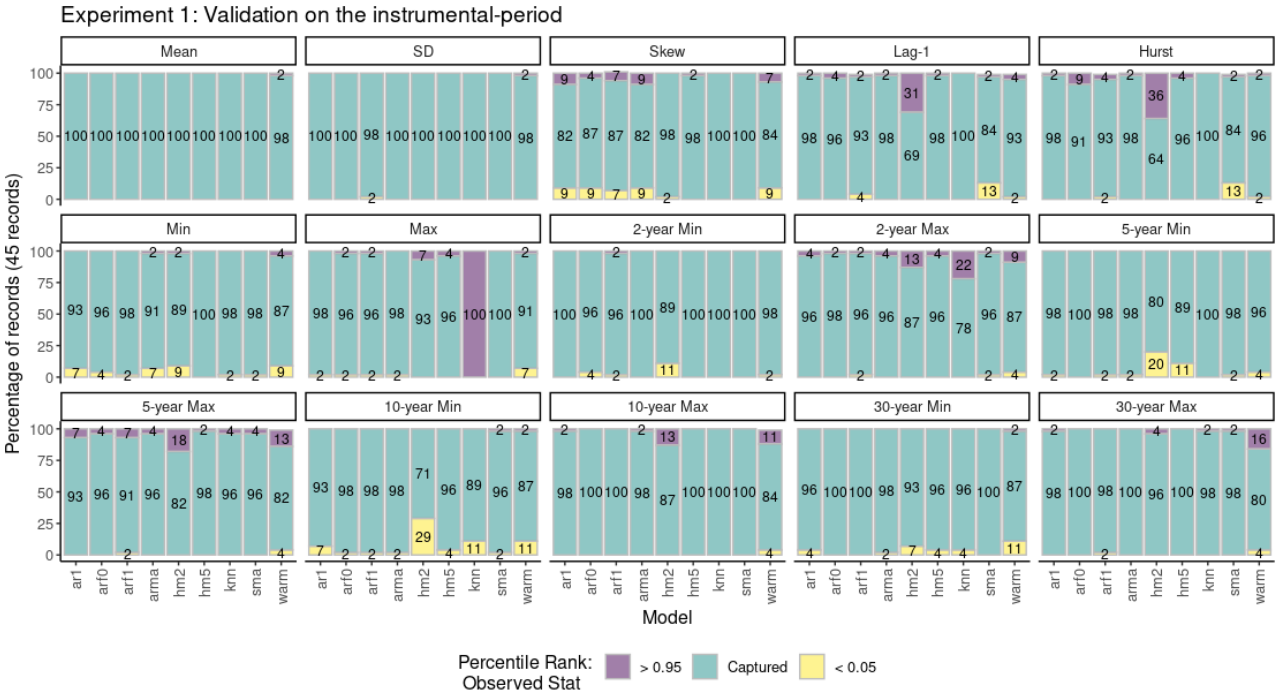

Figure 3-3: Stacked bar charts showing the proportion of records for which the observed statistic had a percentile rank that was either within the 90% confidence intervals (i.e., "captured") or outside the 90% confidence intervals (i.e., percentile rank either < 0.05 or > 0.95) of the stochastic sampling distribution. Numbers in individual bars show the proportion of proxy records that, for each statistic, the model either captured or had a percentile rank < 0.05 or > 0.95

3.6.2 Experiment 2 – Calibrating on the instrumental-period and validating on the pre-instrumental period

Results from Experiment 2 (i.e., calibration on the instrumental-period, validation on the full pre-instrumental period) are shown in Figure 3-4. We can see that:

- Validation results are generally poor for all models. Typically, statistics were reproduced in <50% of the proxy records analysed.
- Across all models, the Hurst exponent, Lag-1 autocorrelation, skew, and standard deviation were more likely to be underestimated than overestimated.
- Across all models, minimum statistics were more likely to be overestimated.

- Across all models, maximum statistics were more likely to be underestimated.

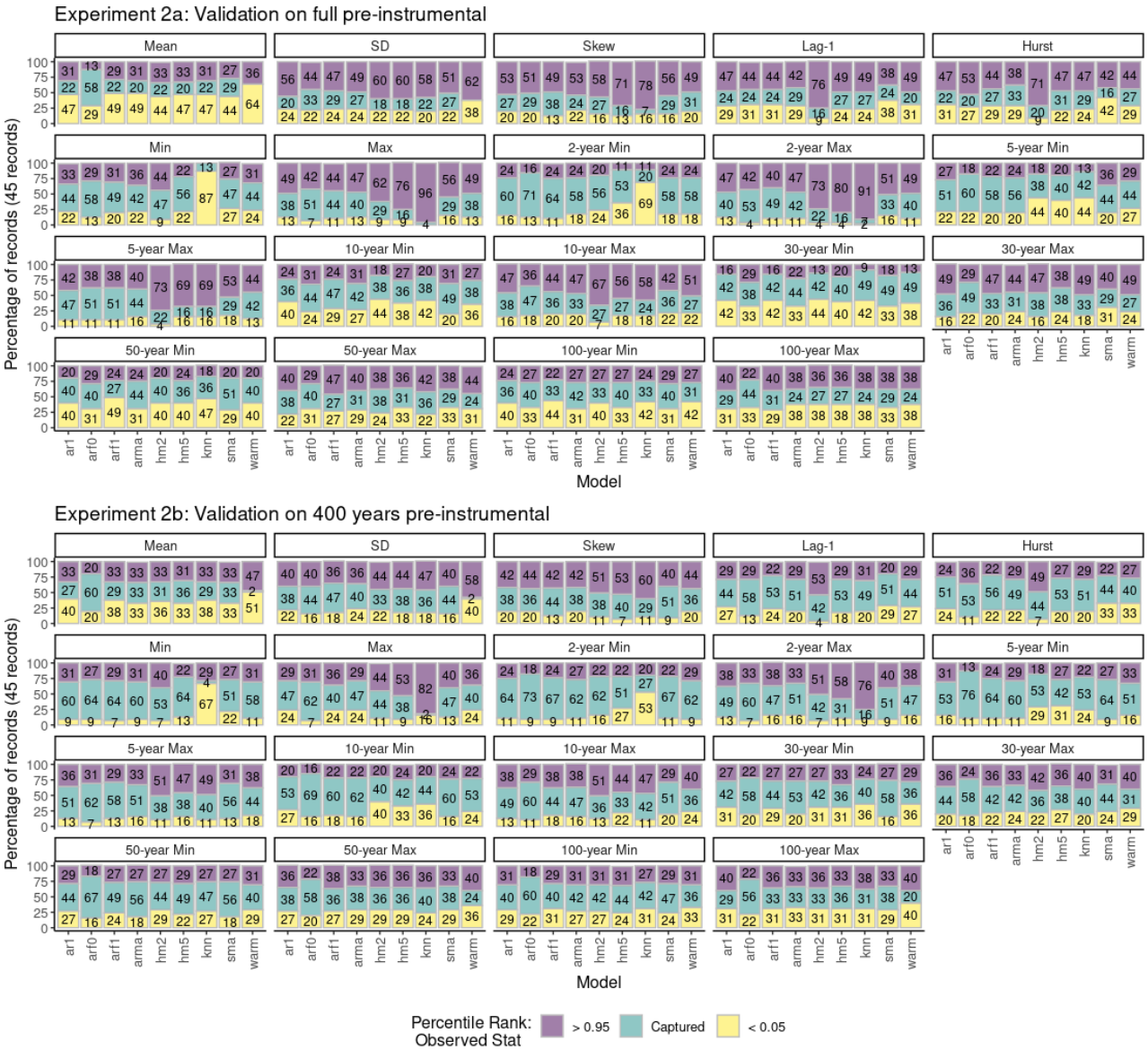

Figure 3-4: Same as Figure 3-3, but for Experiment 2 (i.e. calibration on the instrumental-period of the proxy record, validation on the full pre-instrumental period or validation on the most recent 400-year pre-instrumental period).

Results from Experiment 2b (i.e., calibration on the instrumental-period, validation on the most recent 400-year pre-instrumental period) are also shown in Figure 3-4. As with Experiment 2a, validation results were generally poor for all models (although there is a slight improvement over Experiment 2a). Again, across all models, maximum statistics were more likely to be overestimated than underestimated. However, Experiment 2b differed from Experiment 2a in that minimum statistics were not consistently overestimated.

3.6.2.1 Assessing proxy records for consistency with pre-instrumental statistics

Results from Experiment 2 highlight that a stochastic model calibrated to the instrumental-period may not simulate pre-instrumental variability. To further explore the potential reasons for this result, rolling 100-year statistics were calculated for the proxy pre-instrumental period. Then, the percentile rank of the corresponding instrumental-period statistic was calculated. This allowed instrumental-period statistics to be put into context when compared with statistics from the pre-instrumental period.

For Experiment 2, this allowed any poor performance to be explained by either (1) an instrumental-period climate regime shift outside the confines of pre-instrumental variability; or (2) the inability of models to extrapolate to pre-instrumental climate regimes outside the confines of instrumental variability.

The percentile ranks of different instrumental statistics (with respect to pre-instrumental statistics) are shown in Figure 3-5. Pre-instrumental statistics predominantly capture instrumental statistics. As such, any poor model performance identified in Experiment 2 is likely because the stochastic model cannot extrapolate to pre-instrumental climate.

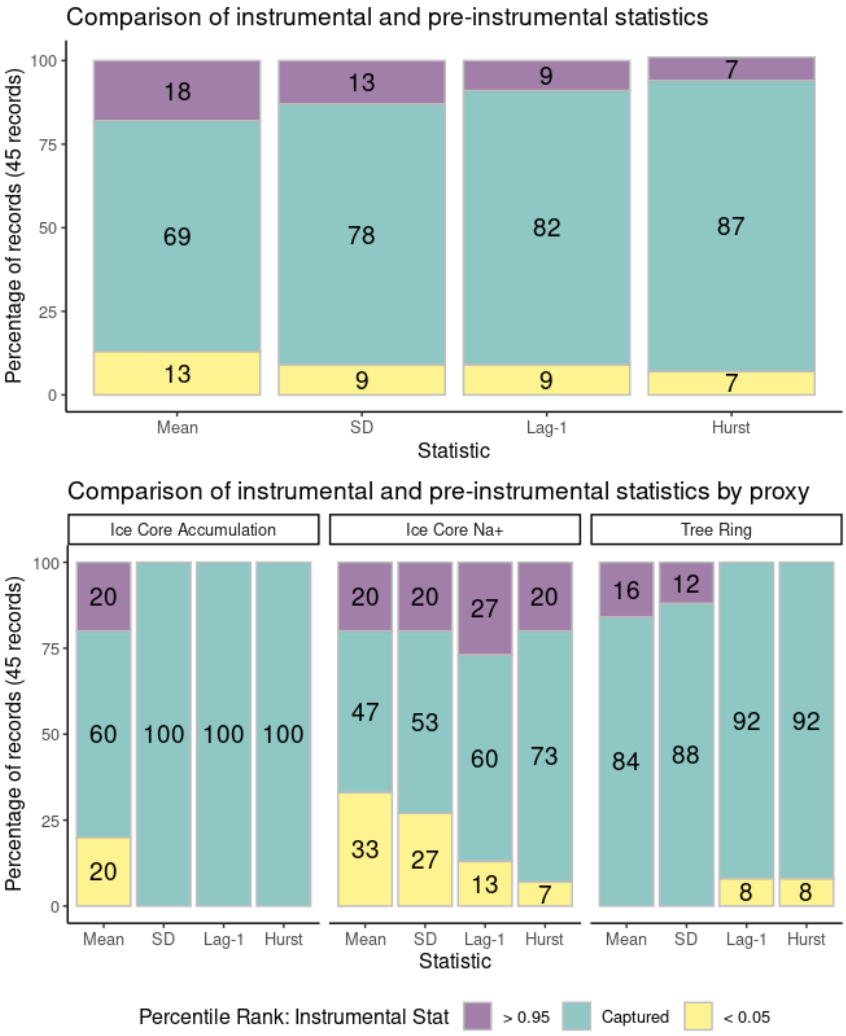

Figure 3-5: Summary of instrumental statistic percentile ranks when compared against rolling 100-year pre-instrumental statistics.

3.6.3 Experiment 3 – Calibrating and validating on the entire proxy record

3.6.3.1 Validation Results

Results from Experiment 3 (i.e., calibration and validation on the full proxy records) across all records are shown in Figure 3-6. We can see that:

- ARMA(1,1), ARFIMA(0,D,0) ARFIMA(1,D,0), 5-state HMM, and the SMA models performed best with respect to 50 and 100-year extremes, reproducing statistics in ~80% of records.
- In contrast, the two-state HMM and AR(1) models performed poorly with respect to 50 and 100-year extremes. These models typically overestimated minimum statistics and underestimated maximum extremes.

- All models tended to perform better on minimum statistics than maximum statistics.

Note that the WARM model performed relatively poorly in simulating low-frequency statistics. However, to calibrate the AR model to low-frequency wavelet signals (which were very smooth and typically monotonically increasing/decreasing for 100s of years), we had to make various simplifications. Further work is needed to determine if these simplifications explain the relatively poor performance or if the WARM model itself is limited to simulating shorter timeseries.

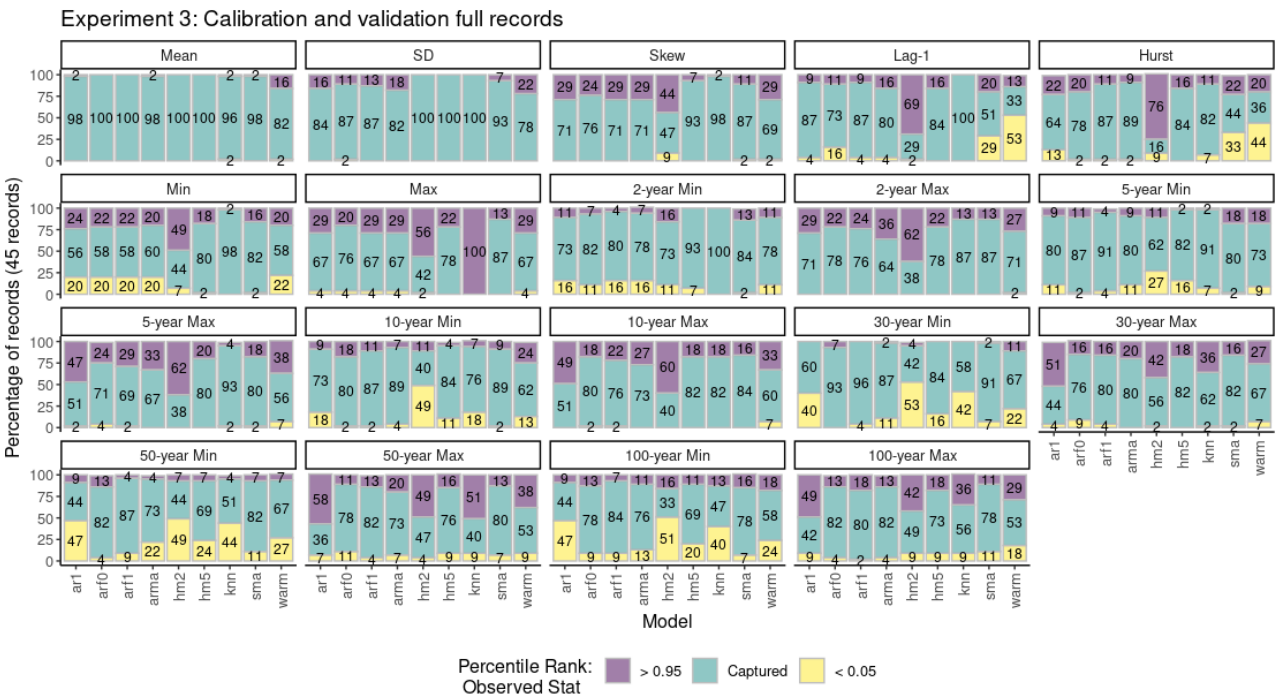

Figure 3-6: Same as Figure 3-3, but for Experiment 3 (i.e. calibration and validation on the full proxy record).

3.6.4 Summary of results

A summary of the results for the different experiments is shown in Figure 3-7. It shows that:

- All models performed well when calibrated and validated on the instrumental-period (Experiment 1). Note that the AR(1) model performed as well as more complicated stochastic models.
- All models performed poorly when calibrated to the instrumental-period and validated on the pre-instrumental period (Experiment 2a/b). This poor performance was apparent even when the pre-instrumental validation period was reduced to the most recent 400-years.

- When calibrated/validated on the full proxy records, the ARMA(1,1); ARFIMA(0,D,0); ARFMA(1,D,0); 5-state HMM; and SMA models performed best, with >80% of statistics captured across all records.

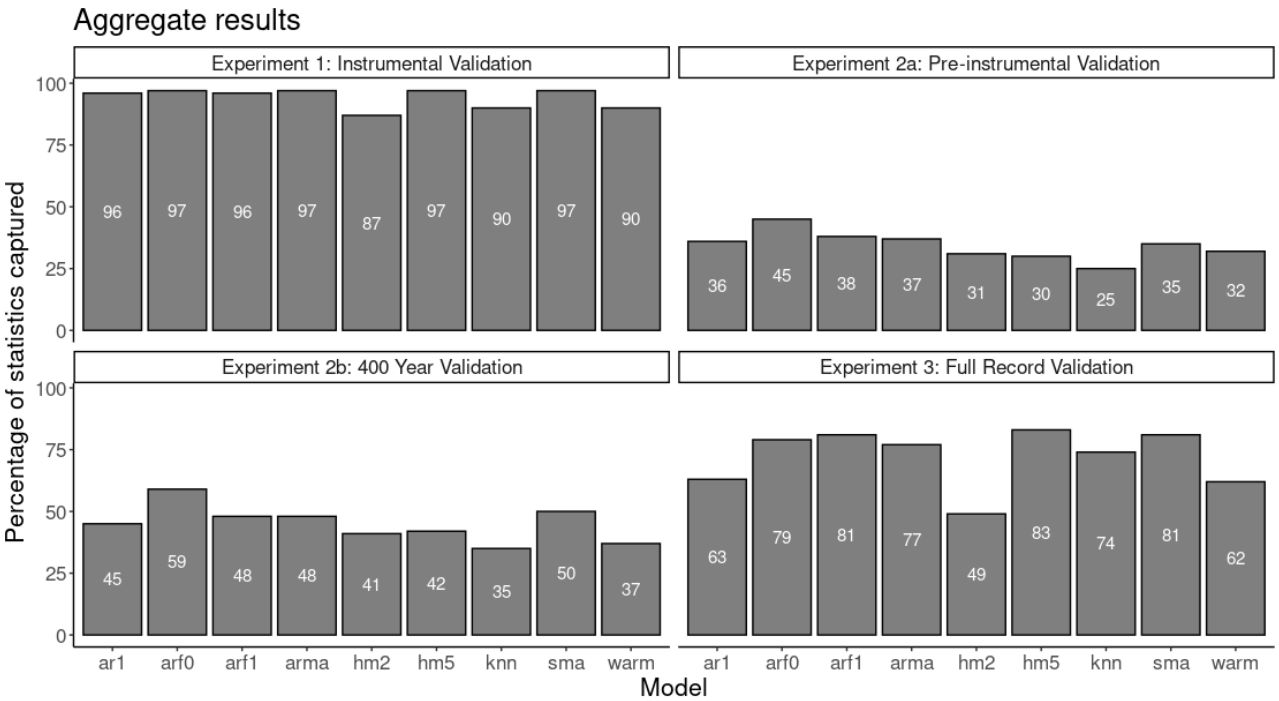

Figure 3-7: Aggregated results from all experiments, showing the total proportion of statistics captured across all records for each model. A statistic was considered 'captured' if the percentile rank was >0.05 and <0.95.

Figure 3-8 shows the aggregated results for each proxy type. Of note, we can see that:

- For Experiment 2a and 2b, stochastic models performed better on tree-ring records than ice core accumulation and Na+.
- For Experiment 3, all stochastic models performed worse on Na+ records than ice core accumulation and tree-ring records (except for KNN).
- The ARMA(1,1) and ARFIMA(0,D,0) models performed well for ice core accumulation and tree-ring records (capturing ~85-90% of statistics), but poorly for Na+ record (capturing ~55% of statistics).

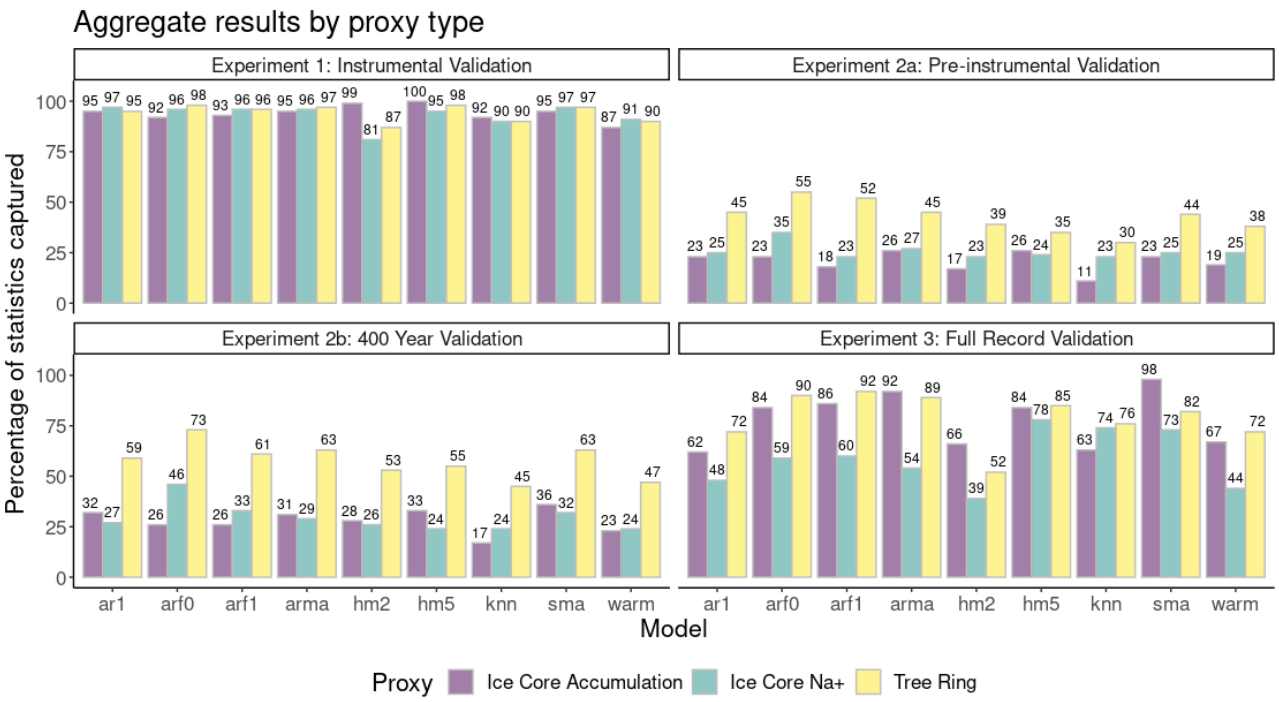

Figure 3-8: Same as, but with aggregated results presented for each proxy type

3.6.4.1 The influence of pre-whitening on validation results

Although all proxy records contain hydroclimatic information, prior to the analysis some tree-ring width records were pre-whitened (discussed in Section 2.2). This can influence subsequent representations of proxy climate variability (Iliopoulou et al., 2018; Razavi and Vogel, 2018), which in turn could influence stochastic model performance. Considering the potential influence of pre-whitening on model performance, the aggregated results from each experiment for pre-whitened and 'standard' tree-ring records are shown in Figure 3-9.

We can see that for Experiments 2a and 2b, models performed better on pre-whitened records. This can be attributed to the pre-whitening removing the low-frequency climate variability not captured in the instrumental-period. The removal of low-frequency climate variability via pre-whitening is also apparent in results for Experiment 3. Models with no explicit mechanism for reproducing long-term persistence (e.g., AR(1), two-state HMM, and KNN models) performed much better on pre-whitened records. This reflects that pre-whitened records contain information about high-frequency, interannual variability – which these models can reproduce.

With respect to Experiment 3, the SMA model performed worse on pre-whitened records. This is because in some cases the pre-whitened records had negative autocorrelation functions. Our implementation of the SMA model is unable to reproduce such autocorrelation functions – only

positive, monotonically decreasing autocorrelation functions can be simulated. In contrast, non-pre-whitened records had consistently positive autocorrelation functions. The SMA model can reproduce these autocorrelation functions, which explains the improved performance on non-pre-whitened records.

Finally, although pre-whitening can influence stochastic model performance, note that for Experiment 3 the better performing models were the same across pre-whitened and non-pre-whitened records. This means that, qualitatively, study results were not impacted by including pre-whitened records. The only exceptions were the K-Nearest Neighbour and AR(1) models, which performed reasonably well for pre-whitened records only.

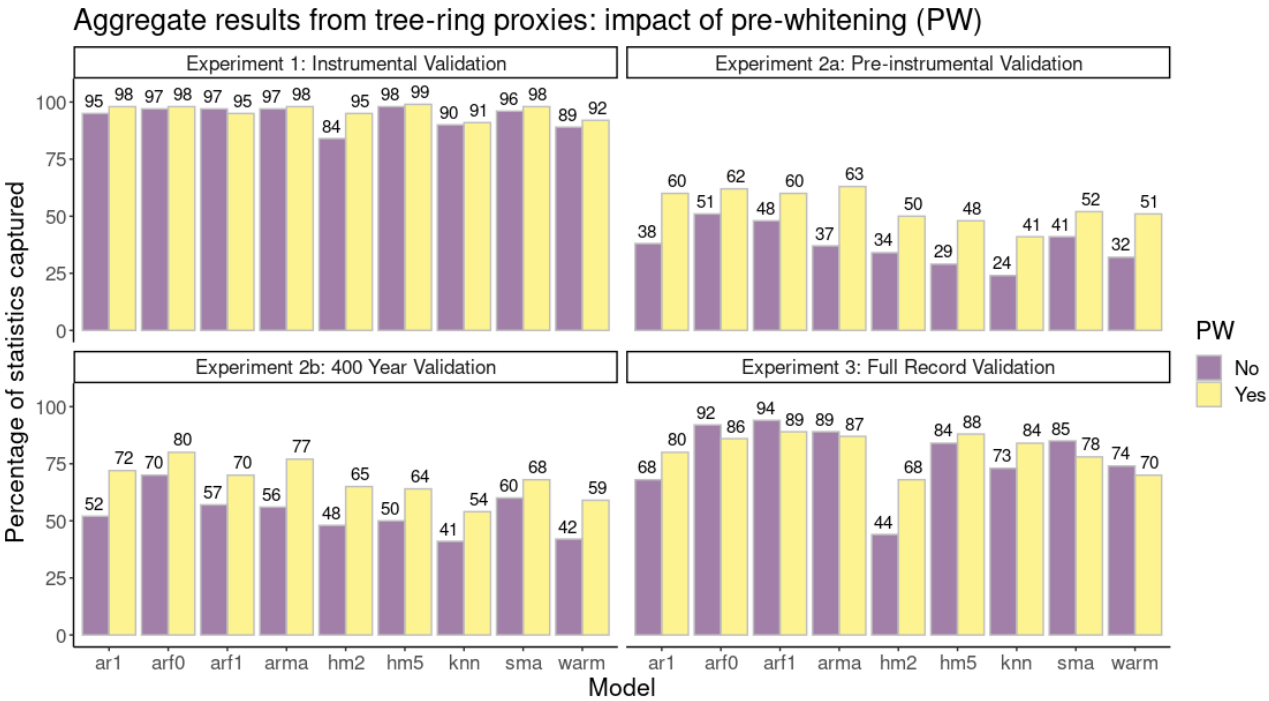

Figure 3-9: Total proportion of tree-ring statistics captured by each model, accounting for pre-whitening of records as a pre-processing step.

3.6.4.2 Why did models perform poorly on Na+ records? Some exploratory analysis

Influence of pre-whitening aside, another interesting result from Experiment 3 is that all stochastic models performed poorly on Na+ records (even models that performed well on other proxy types, such as ARMA(1,1) - Figure 3-8). Thorough exploration of this poor performance requires detailed examination of individual models and records, which is outside the scope of this study. However, we did conduct some exploratory analysis comparing key timeseries features across proxy records.

To inform this exploratory analysis, note that stochastic models aim to replicate a timeseries persistence structure and marginal distribution, typically under an assumption of stationary variance. For each proxy type, Figure 3-10 compares timeseries persistence via a standardised climacogram, marginal distributions via an L-moment diagram, and examines stationary variance via a somewhat heuristic measure of timeseries volatility (the specifics of this heuristic are explained later).

The climacogram displays the relationship between a timeseries aggregation scale (x-axis) and the aggregated timeseries variance (y-axis) – this relationship is a diagnostic for timeseries persistence (Dimitriadis and Koutsoyiannis, 2015). When interpreting the standardised climacogram, steeper negative slopes indicate shorter-term persistence, whereas flatter slopes indicate longer-term persistence. We can see in Figure 3-10 that Na+ records do not have noticeably different climacograms to other proxy types. In contrast, we can see that pre-whitened tree-ring records had steeper, negative slopes than other proxy types, indicating that these records exhibit weaker persistence (consistent with Razavi and Vogel (2018)).

The L-Moment diagram displays the relationship between proxy L-skew and L-kurtosis (note that L-moments exhibit less bias than standard probability weighted moment - Vogel and Fennessey (1993)). We can see that some Na+ records exhibit much larger L-skew and L-kurtosis than other records. Therefore, Na+ records are more likely to have different marginal distribution shapes to other proxy records. Note that although the skew and kurtosis of Na+ records are quite large, various stochastic models, probability distributions, or transformations can handle these features. Different marginal distributions alone may not explain why models performed poorly on Na+ records.

Aside from different marginal distributions, visual inspection of Na+ records also suggested the presence of timeseries 'volatility' (i.e. sudden changes in timeseries variance). This volatility is indicative of non-stationary variance – most of the stochastic models we validated assume stationary variance. Of note is that the KNN model, which makes no assumption of stationary variance, was the only stochastic model which performed better on Na+ records than accumulation records.

To further explore differences in timeseries volatility, we fitted an Autoregressive Conditional Heteroscedasticity (ARCH(p)) model to standardised, differenced proxy timeseries (Engle, 1982). ARCH(p) models are often used in economics to simulate the non-stationary variance of stock prices (large price changes tend to cluster, meaning the variance is non-stationary) (Engle and Bollerslev, 1986). These models simulate timeseries variance as a function of prior timeseries variance, similar to how ARMA models simulate timeseries values as a function of prior values.

For this analysis, we fit an ARCH(1) model to standardised, differenced proxy timeseries. Differencing removed any variations in the proxy mean, resulting in a better representation of underlying changes in variance.

From Figure 3-10, we can see that some Na+ records returned much higher ARCH(1) parameters than other proxy types. This suggests that these records are more likely to have non-stationary variance, which violates an underlying assumption of some stochastic models validated in this study (e.g. ARMA and ARFIMA models).

The ARCH(1) model results in Figure 3-10 should be viewed with caution. Accurately diagnosing and modelling non-stationary variance is more complicated than fitting a simple ARCH(1) model to a standardised, differenced timeseries. Instead, this analysis was conducted to confirm/refute what was suggested by visual inspection of proxy timeseries (i.e. that some Na+ records had non-stationary variance).

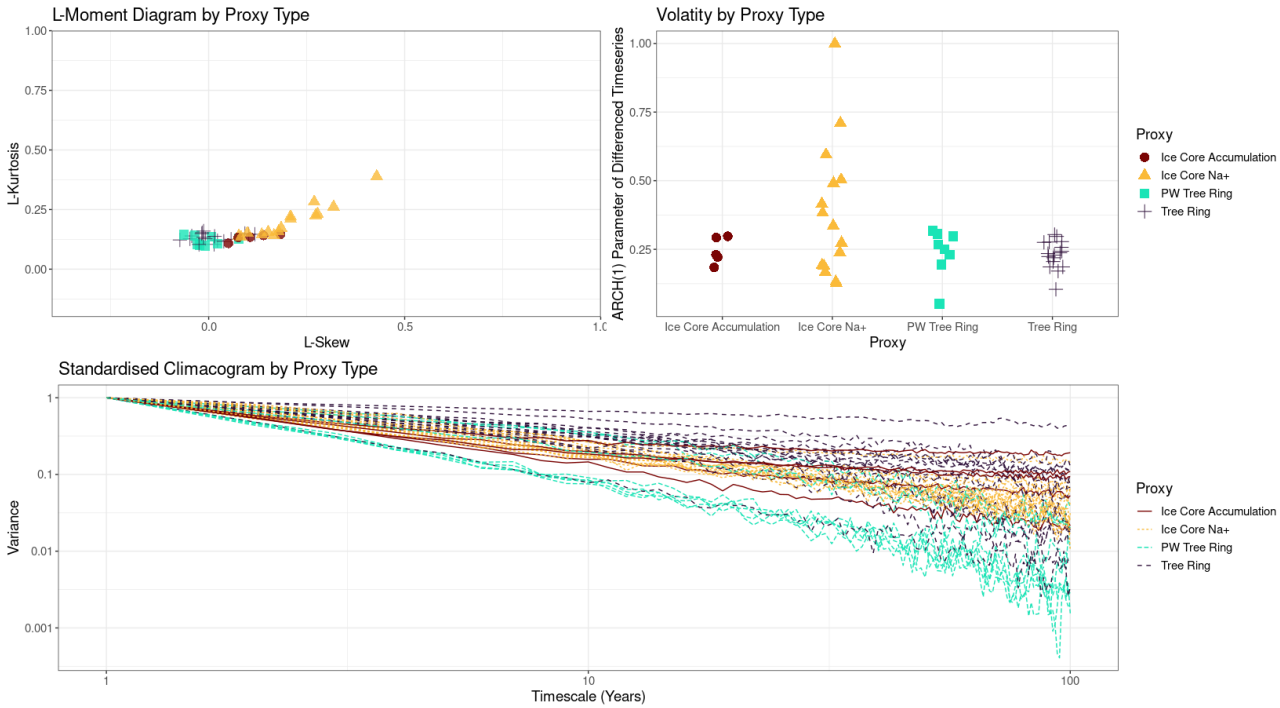

Figure 3-10: L-Moment diagram (top right), ARCH(1) P parameter of differenced timeseries (top right) and climacogram (bottom) for each proxy type

With this in mind, we re-examined Experiment 3, but will all Na+ records removed (Figure 9-2 in the Appendix), which resulting in substantial improvements in stochastic model performance in simulating low-frequency extremes (in particular, the ARMA(1,1), ARFIMA variants, and the SMA model).

3.7 Discussion

In summary the key findings of each experiment were:

- Experiment 1: The instrumental-period is consistent with an AR(1) process, with the AR(1) model reproducing ~95% of evaluated statistics.
- Experiment 2: Stochastic modelling of the instrumental-period will not capture pre-instrumental variability. This result was robust to stochastic model, pre-instrumental period, and proxy type. When validated on either the entire pre-instrumental period or the most recent 400-year pre-instrumental period, models reproduced ~25-50% of evaluated statistics.
- Experiment 3: When calibrated to the full proxy records, the ARMA(1,1), ARFIMA(0,D,0), ARFIMA(1,D,0), five-state HMM, and SMA were the best performing models, reproducing ~80-85% of statistics.

The key finding from Experiment 1 is consistent with previous studies (e.g., Markonis et al., 2018; Srikanthan and McMahon, 2001; Sun et al., 2018) which have demonstrated the ability of the AR(1) model to reproduce instrumental-period statistics. However, given that results from Experiment 3 (which demonstrated that the AR(1) model is unable to reproduce low-frequency climate variability), such findings can be attributed to short instrumental records containing insufficient information to characterise climate variability. Therefore, they struggle to properly identify stochastic models that can reproduce this variability. This should not be viewed as a criticism of these studies, which were limited to using only instrumental records (applications of palaeoclimate proxy data in water management are relatively new developments). Rather, it highlights how instrumental measurements are subject to a significant sampling bias with respect to low-frequency climate variability.

From a climate risk and water management perspective, results from Experiment 2 are particularly concerning. This is because climate risk, and the plans/infrastructure designed to be robust under such risk, is typically quantified using a stochastic model calibrated to instrumental records. Not only were instrumental-period models unable to reproduce climate variability in the full proxy records (which, given the length of these proxy records, may not be surprising), they were also unable to reproduce climate variability in the last 500 years.

Considering water authorities are often legally required to provide water during a 1-in-10,000-year drought, the inability of stochastic models to capture pre-instrumental climate is concerning. It indicates that existing water supply systems have been inadequately designed/optimised and, crucially, that existing water supply systems are exposed - and vulnerable to - much greater climate risk than currently assumed. Further work is needed to explore vulnerabilities arising from this misrepresentation of hydroclimatic risk, which will be system-specific. However, water supply systems are designed under inherently conservative risk estimates.

Although results from Experiment 2 are concerning, a key limitation of the study design is not considering parameter uncertainty during model calibration/replicate generation (only a single, optimal parameter set was used for each model). For 100-year records, parameter uncertainty is large (Thyer et al., 2006) - considering this uncertainty (e.g. through Bayesian calibration methods) results in sampling distributions with much larger variance (Berghout et al., 2017). This increase in variance may result in pre-instrumental statistics being captured by the

instrumental-period model – provided instrumental/pre-instrumental parameters are similar (Figure 3-5 suggests this may be the case). However, given (a) the wide variety of different stochastic models considered in this study - with some not easily calibrated in a way that considers parameter uncertainty, and (b) that typical industry applications of stochastic models use a single optimal parameter set, revisiting Experiment 2a/b while considering parameter uncertainty is left for future work.

Considerations of parameter uncertainty aside, given that results from Experiment 2 were consistent across a wide range of locations regardless of proxy archive type, there is a global need to incorporate palaeoclimate information when characterising climate risk and designing/adapting water supply systems. Previous incorporations have mainly occurred in North America (e.g. Gober et al., 2016; Sauchyn et al., 2015; Tingstad et al., 2014), with these studies demonstrating that system adaptation was necessary to meet operational requirements under palaeoclimate variability. However, such applications of palaeoclimate data will have to address and overcome the various biases and uncertainties introduced when mapping the proxy data to rainfall or streamflow - issues that were not relevant to this study because we examined the proxy data directly. These biases and uncertainties include, but are not limited to, reduced variance in reconstructions relative to the observed climate data (Galelli et al., 2021; Meko et al., 2022; Prairie et al., 2008); non-linearities in the climate-proxy relationship (which are not accounted for using standard linear models) (Geay and Tingley, 2016; O'Donnell et al., 2021); and potential non-stationarity in the climate-proxy relationship (D'Arrigo et al., 2008; Kiem et al., 2020).

Considering a need to incorporate palaeoclimate data in water management, Experiment 3 identified models which may be useful for palaeoclimate-informed stochastic modelling; however, these results come with caveats. For example, both the ARMA(1,1) and ARFIMA(0,D,0) models reproduced ~85% of statistics across all proxy records but, in most cases, produced residuals that were not normally distributed (and, in a few instances, serially dependent – see Figure 9-1 in the Appendix). When model assumptions are not met, identified model parameters may be biased, uncertainty in model parameters is difficult to quantify, and the prediction intervals of the model outputs may be inaccurate (Kavetski et al., 2006). These issues would naturally influence the fidelity of climate risk metrics derived from a stochastic ensemble. However, in most cases the only residual assumption violation was normality – for

linear models, non-normal residuals have minimal impact on parameter bias (Knief and Forstmeier, 2021).

Some minor modifications to the ARMA and ARFIMA models may be necessary to ensure residual assumptions are met. Such modifications have been discussed extensively in the other studies so will only be briefly mentioned here. Potential modifications include:

- Modifying the likelihood function to remove dependence between the Box-Cox parameter and the mean and standard deviation parameters. Thyer et al. (2002) demonstrated that a strong dependence between parameters is introduced when transforming prior to calibration. Strong dependence between parameters, also referred to as collinearity, makes the model hard to optimise and increases parameter uncertainty. To remove this dependency, approximate likelihood functions can be used (Thyer et al., 2002)
- Modelling the autocorrelation and skew of the residuals by calibrating an additional timeseries model to the residual timeseries (Wang et al., 2012). ARCH type models may be particularly useful for modelling skew and non-stationary variance (which may improve the poor Na+ record performance).
- Relaxing the assumption of normally distributed residuals and instead model/draw residuals as/from a skewed distribution (Koutsoyiannis, 2000).

The success of the ARMA(1,1) model may be surprising - given that (a) it is a Markovian (i.e. short-term persistence, or Lag-1) model and (b) multi-decadal/centennial variability suggests long-term persistence is typical for hydroclimatic data (Koutsoyiannis, 2006; Ljungqvist et al., 2016; Pelletier and Turcotte, 1997). However, numerous studies have shown this model can reproduce low-frequency climate variability. Boes and Salas (1978) demonstrated that certain ARMA(1,1) parameter combinations produce identical autocorrelation functions to non-stationary mean stochastic models (i.e. autocorrelations exhibit a power law decay with increasing lag, as opposed to the exponential decay typical of short-term persistence). Koutsoyiannis and Montanari (2007) also demonstrated that, at an annual scale, the ARMA(1,1) model was able to reproduce the Hurst coefficient of an extended temperature reconstructions (the model Hurst coefficients ranged from ~0.79-0.83, which is indicative of low-frequency variability). However, the ability of the model to reproduce low-frequency climate variability was attributed to a sampling size bias. ARMA(1,1) replicates can reproduce Hurst coefficients in timeseries that contain several thousand values, however, as the sample

size approaches infinity replicate Hurst coefficients go to 0.5 (i.e., a Markovian process). Nevertheless, the ability of this model to reproduce annual-scale low-frequency climate variability in records that contain several thousand values (and its relative simplicity) indicates that this model could still be considered for palaeoclimate-informed stochastic modelling.

Regarding palaeoclimate-informed stochastic model selection, the 5-state HMM, which performed well in Experiment 3, has some potential caveats. First, the selection of five hidden states is arbitrary and may not be necessary for all records. Non-parametric HMMs, whereby no assumptions are made about the number of hidden states, can rectify this issue (Van Gael and Ghahramani, 2011). However, for this study, the key motivation for selecting 5 states was to see if additional states improve performance over the 2-state model sometimes used for instrumental rainfall (Thyer and Kuczera, 2000). Second, HMMs with large numbers of hidden states also have many parameters. For instance, the 5-state HMM has ten state parameters (i.e., the mean/standard deviation of the corresponding state normal distribution) and a 5x5 state transition matrix. Given that more parsimonious models can give similar performance (e.g. the ARMA, ARFIMA, and SMA models), the HMM may not be the most appropriate choice for palaeoclimate-informed stochastic modelling. As an aside, the principle of model parsimony can also be applied to selecting the ARFIMA(0,D,0) model over the ARFIMA(1,D,0) model.

This study also demonstrates the usefulness of proxy records in stochastic model validation. In particular, the dataset used in this study provides an opportunity to validate new stochastic models using real-word data that has reasonable representations of low-frequency climate variability and, in the case of Na+ records, potentially non-stationary variance. Typically, stochastic models are either validated using instrumental records, which comes with a sampling bias, or using synthetic data, which may not contain realistic examples of climate variability. As such, we recommend that newly developed stochastic models are validated/justified using proxy data (along with instrumental records and synthetic data).

Aside from stochastic model validation, the dataset used in this study could also help determine how much calibration data is needed to (a) reproduce climate variability from a pre-instrumental period of specific length; and (b) unambiguously identify stochastic model parameters. Regarding (a), this can inform the identification/collection of future proxy records for use in climate risk analysis (based on the historic risk one wishes to quantify). Regarding (b) - this would revisit Thyer et al. (2006), who applied Bayesian methods to synthetic data to

identify data lengths needed to produce AR(1) persistence parameter posteriors that did not contain zero. Naturally, the same methods/questions can be explored using the proxy dataset and appropriate models identified in this study.

These recommendations, and this study in general, have implications for the accurate quantification of historic climate risk. However, there is an additional point that needs addressing: future climate risk posed by anthropogenic climate change.

Considering atmospheric $C0_2$ concentrations and temperatures are increasing to levels for which there is no recent analogue, there is debate about whether it is even reasonable to use historic risk as a proxy for future risk (Stephens et al., 2020). Rather, future risk may have to be quantified using hydroclimatic projections from climate models - not palaeoclimate data. However, model simulations of recent hydroclimate are unable to reproduce observed low-frequency climate variability at regional scales (Rocheta et al., 2017, 2014), nor the drivers of low-frequency climate variability such as the Interdecadal Pacific Oscillation (Henley et al., 2017). This means that palaeoclimate records are currently the best source of information about regional low-frequency variability.

Future work should focus on (a) robust characterisation of baseline climate risk using palaeoclimate records; then (b) comparison of this risk with risk derived from climate model projections, as it is currently unknown whether the climate risk indicated by climate model projections is greater or less than the historic baseline risk. Regardless of the limitations and uncertainties associated with future projections of rainfall/streamflow under climate change, the results from this study indicate that baseline risk has not been properly characterised by stochastically modelling instrumental records.

3.8 Conclusion

In summary, in this study we validated different stochastic models, the main tool used in water management for characterising climate risk, using a global network of millennium-length, hydroclimatically sensitive proxy records. The key findings of this study are:

- Instrumental records, which are typically ~100 years in length, contain insufficient information to identify stochastic models capable of reproducing low-frequency climate variability.

- It is likely that a stochastic model calibrated to a 100-year instrumental record will not extrapolate to pre-instrumental climate conditions. This means that such models cannot properly characterise historic climate risk. Historic risk is often used by in water system design and operation as a proxy for future risk. Therefore, water systems have been designed and operated under mischaracterisations of risk.
- When calibrated to the full proxy record, stochastic models capable of reproducing low-frequency climate variability were identified. These were the ARMA(1,1) model, the ARFIMA(0,D,0) model, the ARFIMA(1,D,0) model, the 5-state HMM, and the Symmetric Moving Average to Anything model. These models are potential candidates for palaeoclimate-informed stochastic modelling.

3.9 Links with following chapters

Chapter 3 highlights potential limitations with inferring climate risk using a stochastic model calibrated to instrumental measurements. Chapter 3 also demonstrates which stochastic models are capable of simulating proxy low-frequency, centennial-scale variability, provided sufficient calibration data is used. These models are suitable candidates for the palaeoclimate-informed stochastic modelling framework proposed in Chapter 6.

Chapter 3 is also used to justify the choice of ARMA(1,1) and ARFIMA(0,D,0) models in subsequent chapters. These models were specifically chosen because they are parsimonious and can be calibrated using a likelihood function. The likelihood function enables Bayesian inference, meaning parameter uncertainty can be quantified and studied in the proxy records. However, before examining ARMA(1,1) and ARFIMA(0,D,0) parameter uncertainty in proxy records, Chapter 4 examines potential issues with posterior inference on synthetic timeseries exhibiting centennial-scale variability.

Chapter 4. Inferring stochastic model parameter uncertainty under centennial-scale climate variability: the role of sampling bias, conditioning error, and likelihood approximation

4.1 Abstract

Quantifying stochastic model parameter uncertainty may be necessary to ensure that climate risk estimates derived from stochastic models, which inform water management, are reliable. To quantify parameter uncertainty, Bayesian calibration methods (which require a likelihood function) are often used. However, accurately quantifying parameter uncertainty with Bayesian methods may be difficult because hydrological processes can exhibit centennial-scale variability (i.e. long-term persistence). Instrumental rainfall and streamflow records used in stochastic model calibration are only ~100-years long; meaning that, with respect to centennial-scale variability, instrumental records have a sampling bias. Furthermore, conditional and approximate likelihoods are often used (for ease of computation). Under centennial-scale variability, the errors associated with the initial conditioning and approximation could, potentially, bias parameter inference. Therefore, this study evaluates the ability of a sophisticated Bayesian calibration method (the No U-Turn Sampler) to reliably infer stochastic model parameter uncertainty from 100-year timeseries using exact and conditional, approximate likelihood functions under high persistence (i.e. centennial-scale variability) and moderate persistence. Synthetic timeseries were generated using the ARMA(1,1) and ARFIMA(0,D,0) models. It was found that, for 100-year high and moderate persistence timeseries, exact likelihoods and conditional, approximate likelihoods return qualitatively similar posteriors. Furthermore, moderate persistence can be inferred from short, 100-year timeseries using ARMA and ARFIMA models with conditional likelihoods. For the ARFIMA timeseries, high persistence can be inferred from 100-year timeseries, however, this posterior will be biased towards underestimating persistence. For the ARMA model, high persistence cannot be reliably inferred from 100-year timeseries using the ARMA model (due to the complex joint posterior of the persistence parameters). This issue can be alleviated using longer timeseries (that contain ~1,000-2,000 values).

4.2 Introduction

In water management and hydrology, stochastic model parameters are inferred from chaotic and, in a practical sense, random hydrological timeseries (Koutsoyiannis, 2010; Loucks and Van Beek, 2017). Because hydrological timeseries are random, stochastic model parameters are in turn random variables subject to uncertainty. Quantifying this parameter uncertainty may be necessary to ensure that the climate risk estimates derived from stochastic models, which inform water management, are robust (Berghout et al., 2017; Stedinger and Taylor, 1982a).

Bayesian methods are often used to infer stochastic model parameter uncertainty (Bezerra et al., 2017; Frost et al., 2007; Thyer and Kuczera, 2000). These methods treat a stochastic model parameter as a random variable with a probability distribution, known as the posterior distribution. Starting with some probabilistic prior beliefs about parameter values, Bayesian inference combines these priors with a likelihood function to calculate the posterior distribution (Gelman et al., 2013). This posterior distribution provides a range of plausible parameter values and their associated probabilities. Naturally, the fidelity of the posterior distribution is dependent on the use of an appropriate likelihood function.

In this study, we examine the influence of various potential confounding factors on accurate posterior inference. These factors can be broadly classified as:

1. Short record sampling bias for timeseries exhibiting centennial-scale variability.

Hydrological processes can exhibit centennial-scale variability (i.e. long-term persistence) (Koutsoyiannis, 2003, 2002). Because most hydrological timeseries are ~100-130 years long, they may be too short to properly identify and characterise long-term persistence (Thyer et al., 2006). This may bias parameter inference.

2. Conditioning error and approximate likelihoods

For computational efficiency, approximate conditional likelihood functions are used instead of exact likelihood functions. For timeseries exhibiting strong serial dependence, the initial conditioning error will be substantial (Box et al., 1970). Furthermore, the conditional likelihood is also approximated based on a finite number of previous values (Haslett and Raftery, 1989). This means that the conditional likelihood is subject to (a) an initial 'conditioning' error and (b) a potential 'approximation' error. This may bias parameter inference.

3. Timeseries models typically assume that the data is normally distributed, hydrological timeseries are often skewed

For example, Autoregressive Moving Average (ARMA) and Autoregressive Fractionally Integrated Moving Average (ARFIMA) models assume that model errors are normally distributed (Box et al., 1970). Because hydrometereological timeseries are typically skewed, they are transformed prior to calibration (Srikanthan and McMahon, 2001). This transformation introduces a strong correlation between the power transformation parameter and the transformed mean and standard deviation, which makes it hard to properly explore the posterior space (Thyer et al., 2002). To remove this dependence, first-order approximations of the mean and standard deviation can be used when estimating the likelihood (Thyer et al., 2002). This may bias parameter inference.

To properly identify and disentangle the relative influence of these confounding factors, in this study stochastic model posteriors are inferred from synthetically generated timeseries. Two parsimonious stochastic models capable of simulating a wide variety of persistence structures, the ARMA(1,1) model and the ARFIMA(0,D,0) model, are used. These timeseries are generated from an underlying 'true' ARMA or ARFIMA model, which removes the need to consider alternative model hypotheses.

4.3 Stochastic models

The ARFIMA(0,D,0) and ARMA(1,1) models were used in this study. These models were selected based on Chapter 3, which found that these models can reproduce centennial-scale variability when calibrated to extended, hydrologically sensitive timeseries.

For a timeseries y of length n, both the ARMA(1,1) and ARFIMA(0,D,0) models represent individual observations at time t as a conditional mean $\bar{\mathbf{y}}_t$ plus a normally distributed residual $\boldsymbol{\varepsilon}_t$ with zero mean and variance $\boldsymbol{\sigma}^2_\varepsilon$

$$y_t = \overline{y}_t + \epsilon_t \qquad \text{Equation 4-1}$$

$$\epsilon_t \sim N(0, \sigma_\epsilon{}^2) \qquad \text{Equation 4-2}$$

For the ARMA(1,1) model, the conditional mean is calculated as:

$$\overline{y}_t | y_{t-1}, \epsilon_{t-1} = \mu + \phi(y_{t-1} - \mu) + \theta\epsilon_{t-1} \qquad \text{Equation 4-3}$$

Where $\boldsymbol{\mu}$ is the timeseries mean, $\boldsymbol{\phi}$ is the lag-1 autoregressive parameter, and $\boldsymbol{\theta}$ is the lag-1 moving average parameter. For a stationary ARMA(1,1) model, both $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are restricted to values between -1 and 1.

For the ARFIMA(0,D,0) model, the conditional mean can be calculated as:

$$\overline{y}_t | \epsilon_{t-1}, \dots, \epsilon_{t-k} = \mu + \sum_{k=1}^{\infty} a(k)\epsilon_{t-k} \qquad \text{Equation 4-4}$$

Where

$$a(k) = \frac{\Gamma(k+D)}{\Gamma(k+1)\Gamma(D)} \qquad \text{Equation 4-5}$$

However, as k →∞, we can approximate the weights *a(k)* as:

$$a(k) \sim \frac{1}{\Gamma(D)} k^{D-1} \qquad \text{Equation 4-6}$$

Considering all previous timeseries values when estimating the conditional mean can significantly increase calibration time. To reduce calibration time, when estimating the conditional mean, the number of previous weights considered can be reduced to 100 (following Haslett and Raftery (1989)). This results in a conditional mean approximated as:

$$\overline{y}_t | \epsilon_{t-1}, \dots, \epsilon_{t-k} = \mu + \sum_{k=1}^{\min(100,t)} a(k)\epsilon_{t-k} \qquad \text{Equation 4-7}$$

For stationary models, D is restricted to values between -0.5 and 0.5.

For both the ARMA and ARFIMA models, the likelihood of some parameter set $\boldsymbol{\theta}_p$ can be calculated as the product of conditional likelihoods:

$$P(\theta_p | y) \propto \prod_{t=2}^{n} y_t^{\lambda-1} * N(\bar{y}_t, \sigma_\epsilon^2, lb, ub) \qquad \text{Equation 4-8}$$

The prior errors used to calculate each $\bar{\mathbf{y}}_t$ are derived by y_{t-k} - $\bar{\mathbf{y}}_{t-k}$. To infer the error at t_1, we assume that the error at t_0 is 0 and that $\bar{\mathbf{y}}_1 = \mu$. For timeseries exhibiting long-term persistence, this initial conditioning error will propagate and influence many future estimates of the conditional mean.

For both the ARFIMA(0,D,0) and ARMA(1,1) models, the likelihood is calculated assuming the data is normally distributed. However, hydroclimatic data typically have skewed marginal distributions and a finite lower bound of zero (i.e., are non-normal). To remove this skew prior to model calibration, the timeseries can be transformed via the parametric Box-Cox transformation (Box and Cox, 1964).

$$\begin{aligned} if\ \lambda \neq 0: z_t &= \frac{y^\lambda{}_t - 1}{\lambda} \\ if\ \lambda = 0: z_t &= log(y_t) \end{aligned} \qquad \text{Equation 4-9}$$

This approach requires inference on the additional Box-Cox parameter, plus additional approximations to account for the dependence between the Box-Cox transformation and the sample mean and standard deviation. The modified likelihood for both the ARMA(1,1) and ARFIMA(0,D,0) models is derived in Section 9.2.1.

4.4 Selection of synthetic timeseries parameters

The need for a thorough synthetic analysis was, in part, identified by calibrating ARMA and ARFIMA models to palaeoclimate proxy records. Various proxy records from Chapter 3 were initially calibrated and, in some cases, returned persistence parameters close to the non-stationary zone (indicative of high persistence – see Box et al. (1970)). Figure 4-1 provides two examples. In the left column, the ARMA(1,1) model calibrated to the Law Dome sea salt

record of Jong et al. (2022). In the right column, the ARFIMA(0,D,0) model calibrated to the Southern Finland tree-ring chronology of Helama et al. (2009). The Law Dome **ϕ** posterior had a mode of approximately 0.98, and the Finland D parameter had a mode of approximately 0.49, both close to non-stationarity (i.e. for ARMA models, **ϕ** or **θ** parameters with an absolute value > 1 and for ARFIMA models, a D parameter with an absolute value > 0.5).

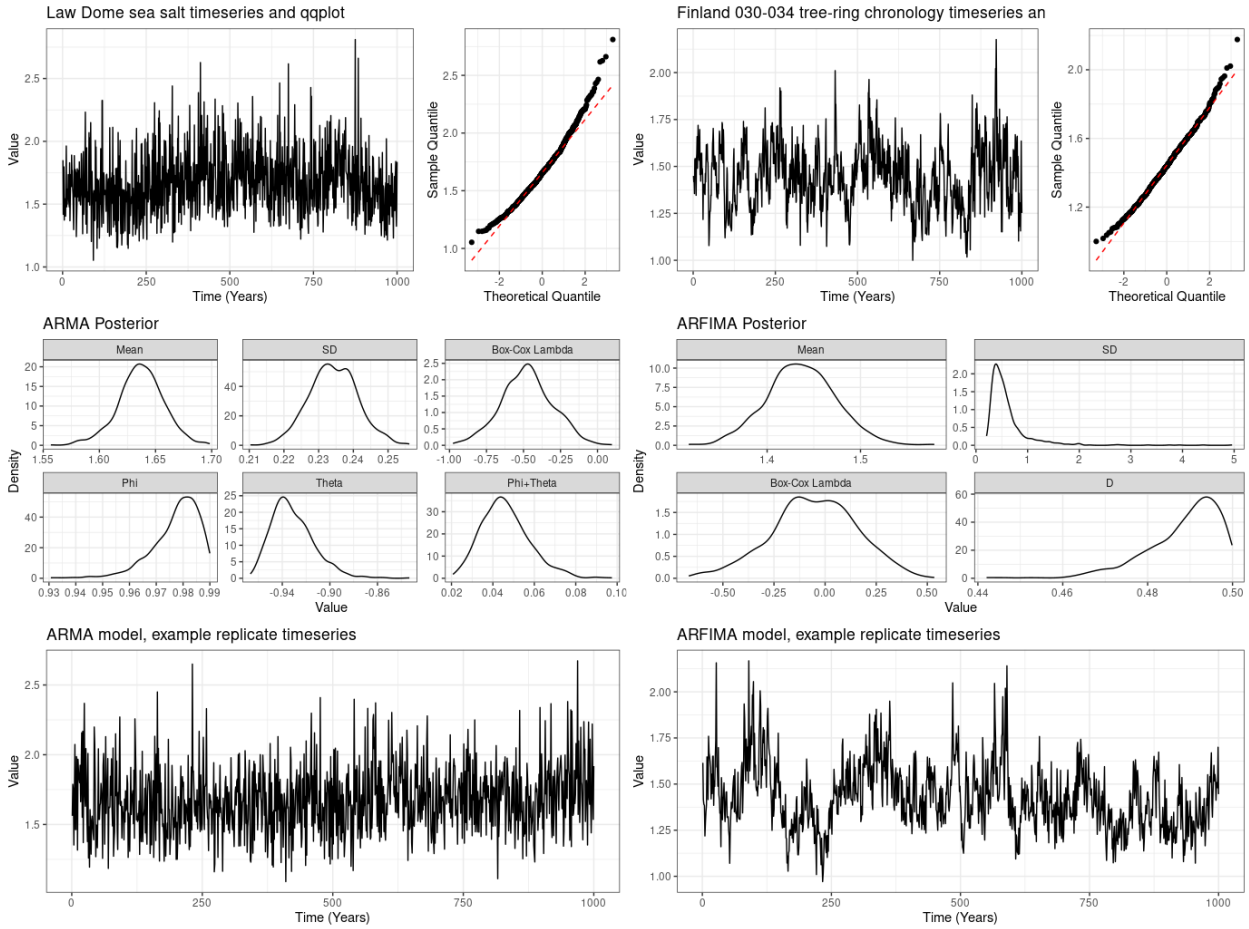

Figure 4-1: Example proxy timeseries, posteriors, and synthetic replicate for two different proxy records. Left column: Example from the Law Dome summer sea salt record of Jong et al. (2022), with inferred ARMA(1,1) persistence parameters close to the non-stationary zone (i.e. Phi and Theta parameters with an absolute value greater than 1). Right column: Example from the Southern Finland tree-ring chronology of Helama et al. (2009), with a persistence parameter close to the non-stationary zone (i.e. a D parameter with an absolute value greater than 0.5).

Based off this preliminary work, 'high' and 'moderate' persistence parameters were identified for the ARMA(1,1) and ARFIMA models. These parameters were then used in subsequent analyses. Note that not all proxy records produced persistence parameters close to the non-stationary zone. However, it was a frequent enough occurrence to warrant further exploration.

Table 4-1: Synthetic model parameters used for this study

Model Type	**ARMA(1,1)**	**ARFIMA(0,D,0)**
High Persistence	$\boldsymbol{\phi}$ = 0.98, $\boldsymbol{\theta}$ = -0.95	D = 0.48
Moderate Persistence	$\boldsymbol{\phi}$ = 0.75, $\boldsymbol{\theta}$ = -0.5	D = 0.25

The theoretical Autocorrelation Functions (ACFs) for the high and moderate ARMA and ARFIMA models are shown in Figure 4-2. For both models, the high persistence parameters produce ACFs with high values, even as the lag approaches 100. Figure 4-2 also highlights a key difference between ARMA and ARFIMA models, the ARMA model ACF decays exponentially, the ARFIMA ACF decays hyperbolically (Dimitriadis and Koutsoyiannis, 2015).

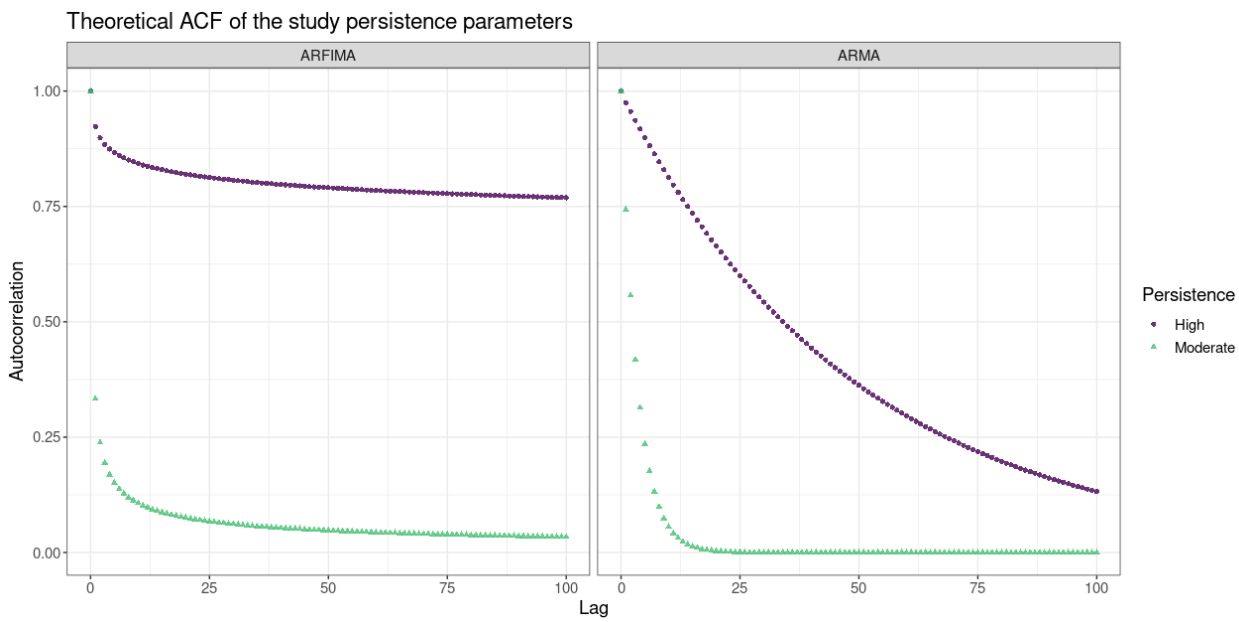

Figure 4-2: Theoretical Autocorrelation Function for the ARFIMA(0,D,0) and ARMA(1,1) models examined in this study (see Table 4-1 for definition of high and moderate persistence).

4.5 Methods

4.5.1 Inferring posteriors using the No U-Turn Markov Chain Monte Carlo Algorithm

There are various Bayesian algorithms that can be used to infer posterior distributions – in this study, the No U-Turn sampling (NUTS) algorithm (Homan and Gelman, 2014) was used. NUTS is an efficient algorithm that can sample from complex posterior distributions. It is a

variant of the Hamiltonian Monte Carlo (HMC) algorithm, which automatically adapts to the geometry of the target posterior distribution during sampling (Upadhyay et al., 2015). These algorithms work by simulating the motion of a particle through a high-dimensional space (each dimension corresponds to a parameter of the probability distribution being sampled) (Betancourt, 2018). A combination of random and deterministic steps are used to guide the particle towards regions of high probability – this 'guidance' makes these algorithms more efficient than standard random walk/Metropolis sampling algorithms used in Bayesian inference (Betancourt, 2018).

For each model calibration, 20,000 posterior samples were taken from eight chains (2,500 samples per chain, plus a prior 500 samples as a burn-in period). After simulation, the 20th sample from each chain was extracted and combined into the final posterior sample. This maximised independence between posterior samples (Jones and Qin, 2022). Chains were initialised at the calibration data's sample mean; sample standard deviation; a $\boldsymbol{\lambda}$ value of one (corresponding to no transformation); and persistence parameters of zero (corresponding to no persistence).

4.5.2 Evaluating posterior inference

To evaluate posterior inference, followed the same general process for different likelihood functions and timeseries lengths:

1. Generate a synthetic timeseries.
2. Infer posteriors from synthetic timeseries.
3. Calculate the p-value of the true parameter with respect to each posterior.
4. Repeat 1-3 20 times to derive a 'p-value' distribution.

Over repeated analyses, this p-value should follow a uniform distribution (because the p-value under the null hypothesis is uniformly distributed) (Talts et al., 2020). Skewed p-value distributions indicate that the posteriors are biased.

Aside from examining the p-value distribution, the number of posteriors where the p-value was either outside or within the 90% posterior credible interval was also calculated. The true value is expected to lie within the 90% credible interval ~90% of the time (subject to sampling uncertainty). This is a less stringent evaluation than p-value uniformity and can indicate whether potentially biased posteriors still consistently capture the true parameter.

4.5.3 Variations of the likelihood function

For both the ARMA and ARFIMA models, three different variants of the likelihood function were used:

1. The exact likelihood.

This involved using known, prior residual errors to calculate the likelihood function from the synthetic timeseries. For example, when making inference on a synthetic 100-year ARMA(1,1) timeseries, this required a 101-year timeseries to be simulated from pre-generated residual errors. The first simulated value and residual error is then used to condition the following 100 values used in the likelihood calculation.

2. The conditional likelihood.

The exact likelihood, as described above, assumes that the residual errors prior to the first observation are known. These errors are unknown. Although exact likelihood functions that do not require known prior innovations exist, calculations are time consuming. Therefore, conditional likelihoods are often used to approximate the exact likelihood. This involves assuming that (a) the 0th residual error is zero and (b) the 1st conditional mean is equal to the unconditional mean, which is then used to estimate the 1st residual error. In this study, the first timeseries value was not included in the final likelihood calculation.

3. The truncated conditional likelihood.

This used the same method as the conditional likelihood, but with the first 10 values removed from the final likelihood calculation. This was a heuristic approach to remove biases caused by a potentially large initial conditioning error.

4.5.4 Experiment 1: Normally distributed timeseries

Before assessing the influence of power transformation and parameter approximation on posterior inference, we conducted analysis using normally distributed timeseries. This was to evaluate posterior inference under 'ideal' conditions, without the added complexity of timeseries skew and transformation.

The first experiment involved evaluating the different likelihood variants on 100-year timeseries, generated from the high and moderate persistence models in Table 4-2. Then, longer timeseries were evaluated. For the ARFIMA model, 500-year timeseries were evaluated.

Timeseries longer than 500 years were not considered because (a) posterior inference for the ARFIMA model was reasonable for these shorter timeseries and (b) calibrating the ARFIMA model to longer timeseries is time consuming. In contrast, for the ARMA(1,1) model, we evaluated timeseries continuing 500; 1,000; and 2,000 values (reasons for this are explained in the Results section).

4.5.5 Experiment 2: Skewed timeseries

After evaluating posterior inference for normally distributed timeseries, we then examined skewed timeseries. This required additional inference on the Box-Cox transformation $\boldsymbol{\lambda}$ parameter. The modified, approximate likelihood function for Box-Cox transformed ARMA and ARFIMA models is shown in the Appendix.

For the high and moderate persistence ARMA and ARFIMA models, four different Box-Cox $\boldsymbol{\lambda}$ parameters (0, 0.2, 0.5, and 1) were used to generate a single synthetic timeseries containing 100; 500; 1,000; and 2,000 values. These $\boldsymbol{\lambda}$ are typical of climate timeseries, with a value of 0 indicating a log transformation and a value of 1 indicating the identity (i.e. no) transformation (McInerney et al., 2019). Timeseries had a mean of 5 and a residual variance of 1 (note that Box-Cox transformations can only be applied to positive data). Because it was infeasible to generate and evaluate 20 timeseries for each persistence parameter, timeseries length, and $\boldsymbol{\lambda}$ combination, a single timeseries was generated for each and results aggregated across model persistence and timeseries length.

Table 4-2: Model parameters and timeseries length used to evaluate parameter inference on skewed timeseries.

$\boldsymbol{\mu}$, $\boldsymbol{\sigma}$ and persistence parameters	**Timeseries Length**	**Box-Cox $\boldsymbol{\lambda}$**
$\boldsymbol{\mu}$ = 5 $\boldsymbol{\sigma}$ = 1 High Persistence Moderate Persistence	100 500 1,000 2,000	0 0.2 0.5 1

4.6 Results

4.6.1 Results from 100-year analyses

Results for the 100-year ARFIMA analysis are shown in Figure 4-3. For the moderate persistence model, all p-value distributions are approximately uniform. In contrast, for the high persistence model, the D parameter p-value distribution is concentrated below 0.5, indicating a tendency to underestimate the true D parameter.

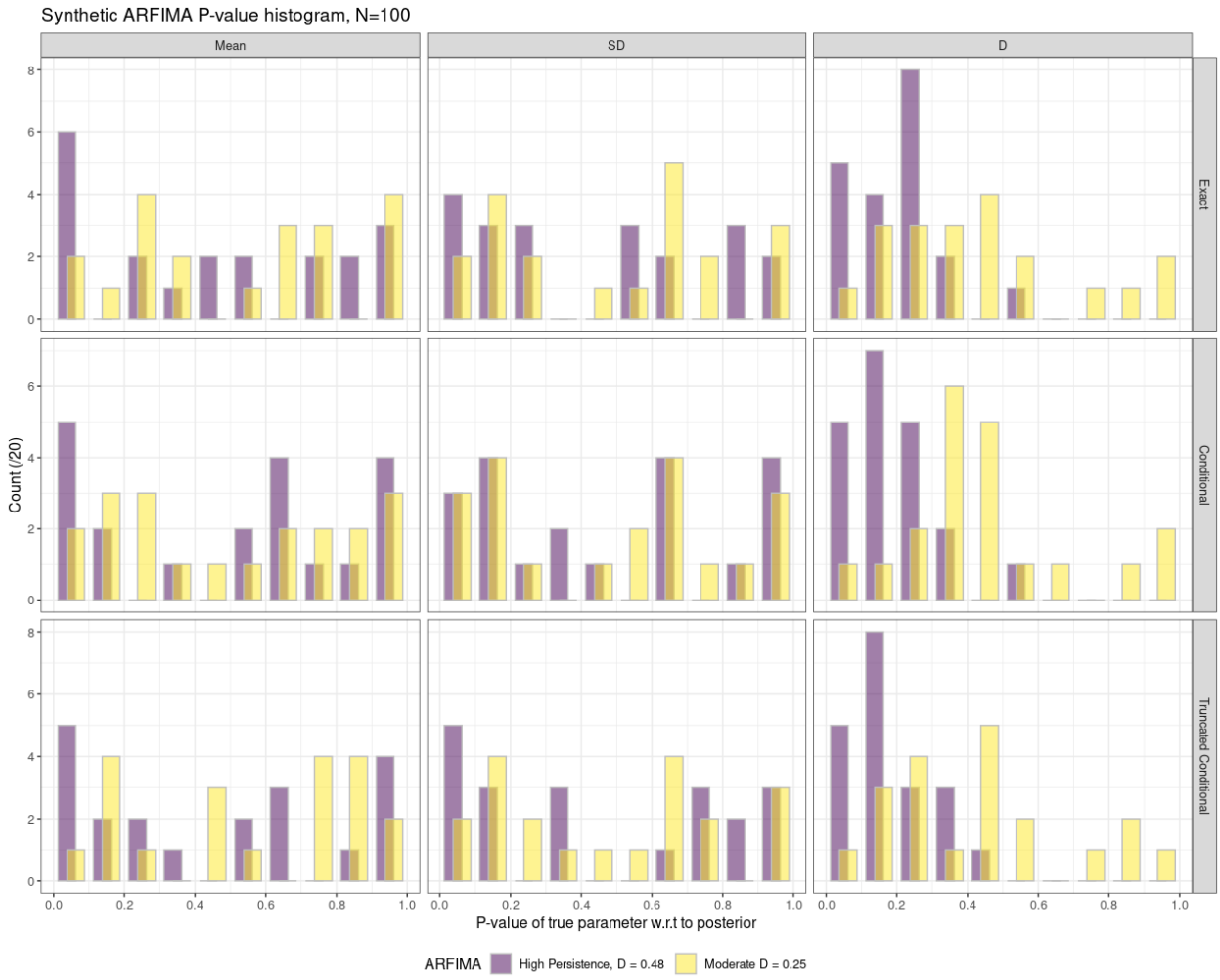

Figure 4-3: P-value distribution for 100-year ARFIMA analysis. The p-value of the underlying 'true' parameter was calculated with respect to the corresponding inferred posterior.

For the 100-year ARFIMA analysis, the proportion of p-values within or outside the 90% credible interval is shown in Figure 4-4. We can see that, aside from the truncated conditional likelihood inference on the mean parameter, the 90% CI consistently contained the true parameter.

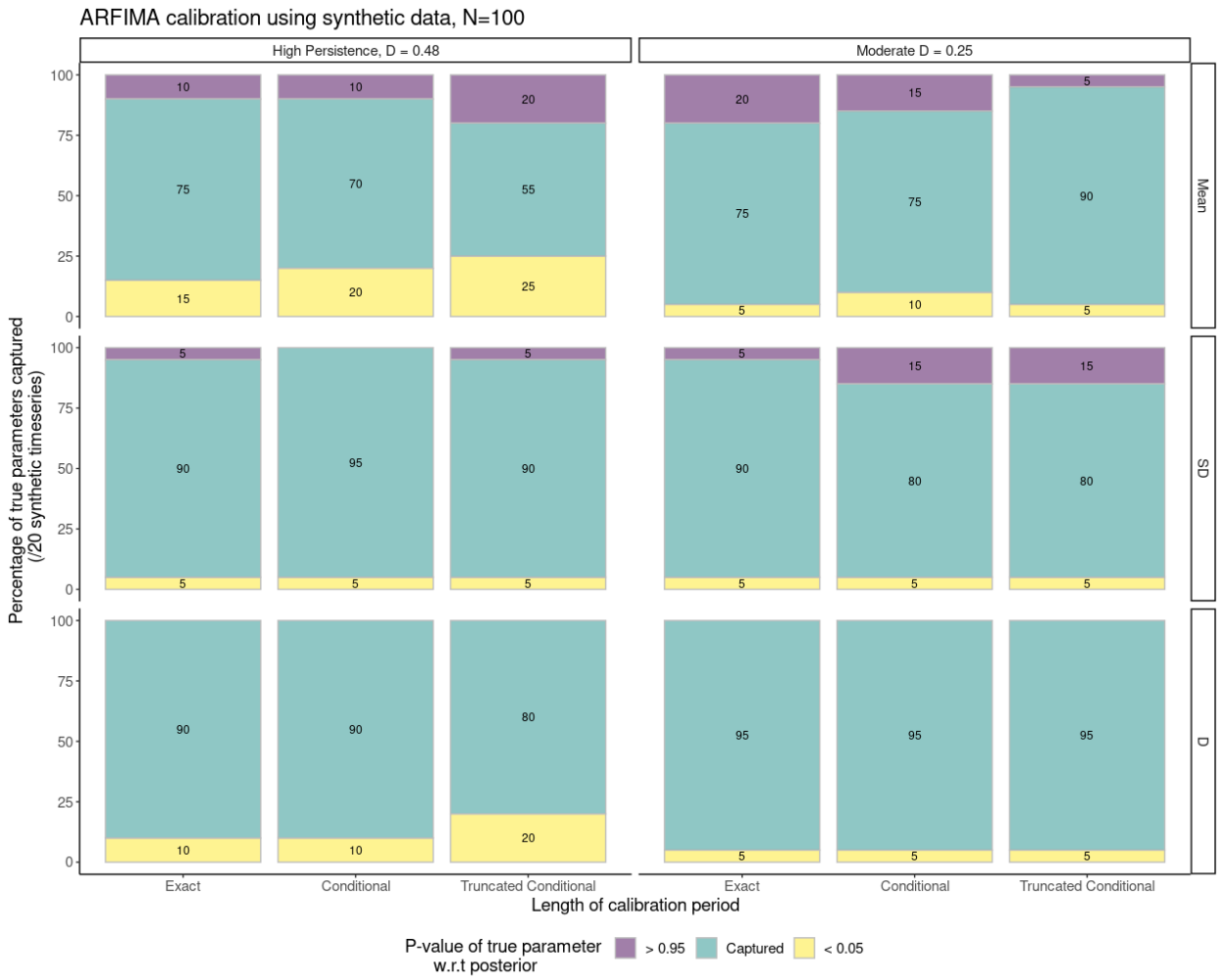

Figure 4-4: Summary of Figure 4-3, with p-values classified as either within the 90% credible interval ('captured') or outside the credible interval (i.e having a value < 0.05 or > 0.95).

Results for the 100-year ARMA analysis are shown in Figure 4-5. For the moderate persistence model, the mean and SD p-value distributions are approximately uniform for all likelihood functions. However, the $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ p-value distributions indicate a tendency to underestimate $\boldsymbol{\phi}$ and overestimate $\boldsymbol{\theta}$. For the high persistence model, the posteriors always underestimated $\boldsymbol{\phi}$ and overestimated $\boldsymbol{\theta}$, regardless of likelihood function.

Figure 4-5: Same as, but for the ARMA(1,1) model.

For the 100-year ARMA analysis, the proportion of p-values within or outside the 90% credible interval is shown in Figure 4-6. The different likelihood variants returned identical results for the high and moderate persistence models. For the moderate persistence model, the 90% CI always captured the true value (note that we would expect 90% CI to capture the true value 90% of the time, but this experiment was only repeated 20 times, meaning 100% capture is possible). In contrast, for the high persistence model, the true $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ parameters were never within the 90% CI.

Figure 4-6: Same as, but for the ARMA(1,1) model.

4.6.2 Examining the high persistence ARMA posterior

Figure 4-7 shows that even with an exact likelihood and a known 'true' model, the NUTS algorithm was unable to properly explore the persistence posterior space. To better understand the persistence posterior space, we enumerated over the $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ stationary range and inferred the joint posterior density for several synthetically generated timeseries (using an exact likelihood and setting the mean and residual standard deviation to 0 and 1 respectively).

Several example joint posterior densities for $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are shown in Figure 4-7. For the high persistence ARMA, joint posterior densities are typically bimodal, peak close to the non-stationary zone, and are not constrained around the true parameter value (spanning across the stationary parameter space). Spanning most of the stationary parameter space indicates that 100-year timeseries contain limited information about centennial-scale variability. In contrast,

for the moderate persistence ARMA, joint posterior densities are typically unimodal and somewhat constrained around the true parameter values.

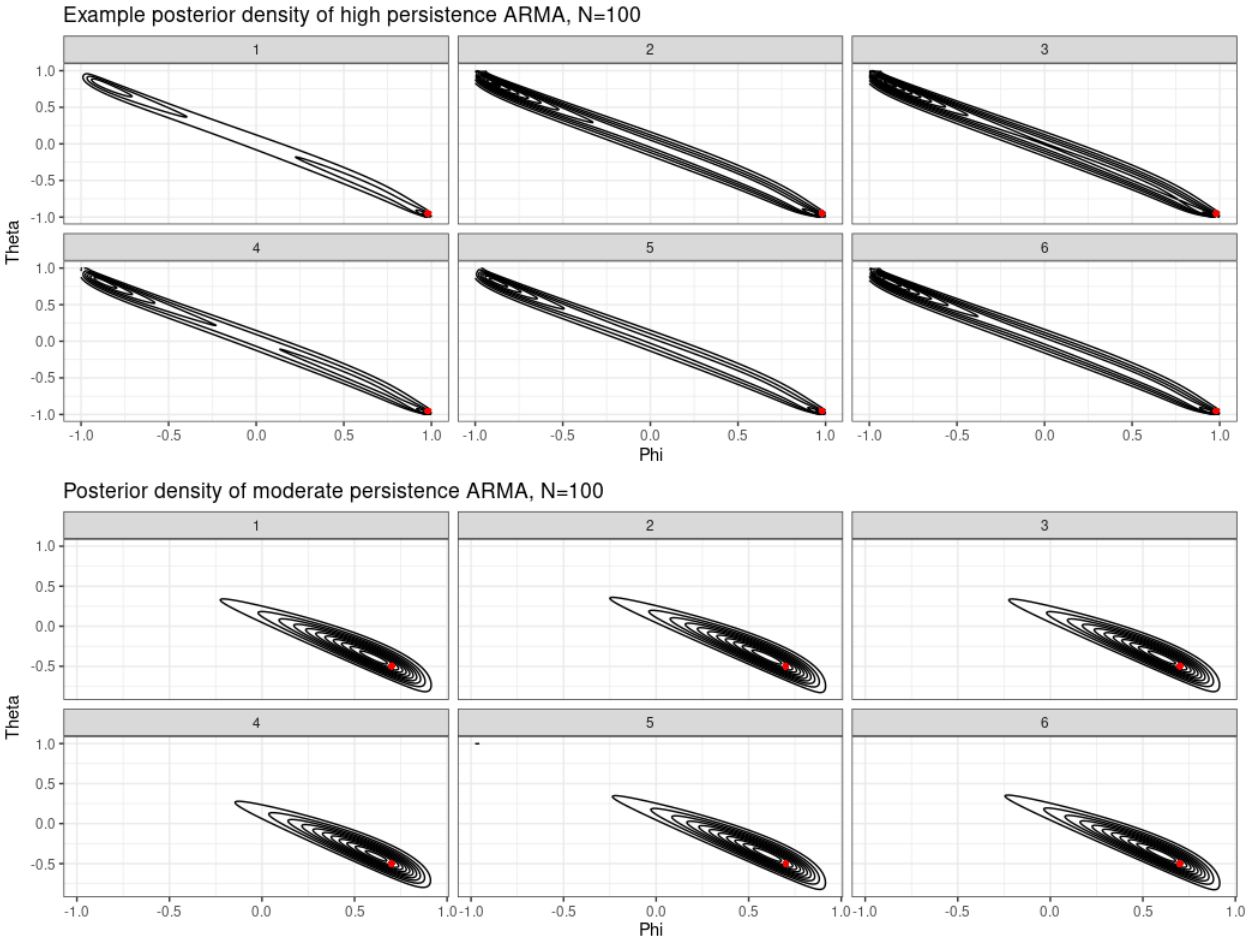

Figure 4-7: Joint posterior density of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ for 100-year synthetic timeseries. Exact likelihood values were used, with the mean and residual variance set to zero and one respectively. True parameter values are shown in red.

4.6.2.1 How much data is needed to identify high persistence ARMA parameters?

With respect to the high persistence ARMA model, bimodal, largely separated posteriors are particularly hard for MCMC samplers to correctly infer (Vrugt et al., 2009). To better explore these complicated posteriors, we repeated the Figure 4-5 analysis using exact likelihoods with longer timeseries (100; 500; 1,000; and 2,000 years). We also assessed if increasing the NUTS acceptance probability (from 0.8 to 0.99) and individual chain length (from 2,500 to 20,000) improved posterior inference. For this analysis, note that for longer timeseries we expect any likelihood biases caused by the initial conditioning error to be insignificant. Therefore, only the exact likelihood was considered.

Results from this analysis are shown in Figure 4-8. We can see that:

- For all parameters and timeseries lengths, the moderate persistence model produced approximately uniform p-value distributions.
- For the high persistence model, longer timeseries improved posterior inference, but the p-value distributions still indicate bias.
- For the high persistence model, running the NUTS algorithm with a higher acceptance probability and longer chains did improve posterior inference. However, the p-value distributions still indicate bias towards underestimating $\boldsymbol{\phi}$ and overestimating $\boldsymbol{\theta}$.

Figure 4-8: P-value distribution for ARMA(1,1) analysis using different timeseries lengths. Exact likelihoods were used for all posterior inference. For the 'High Persistence, Long Chains' analysis, 8 MCMC chains with 20,000 iterations were generated. For the 'High Persistence, Standard Chains' analysis, 8 MCMC chains with 2,500 iterations were generated. The 'Moderate Persistence' analysis used the same NUTS configuration as the 'High Persistence, Standard Chains' analysis.

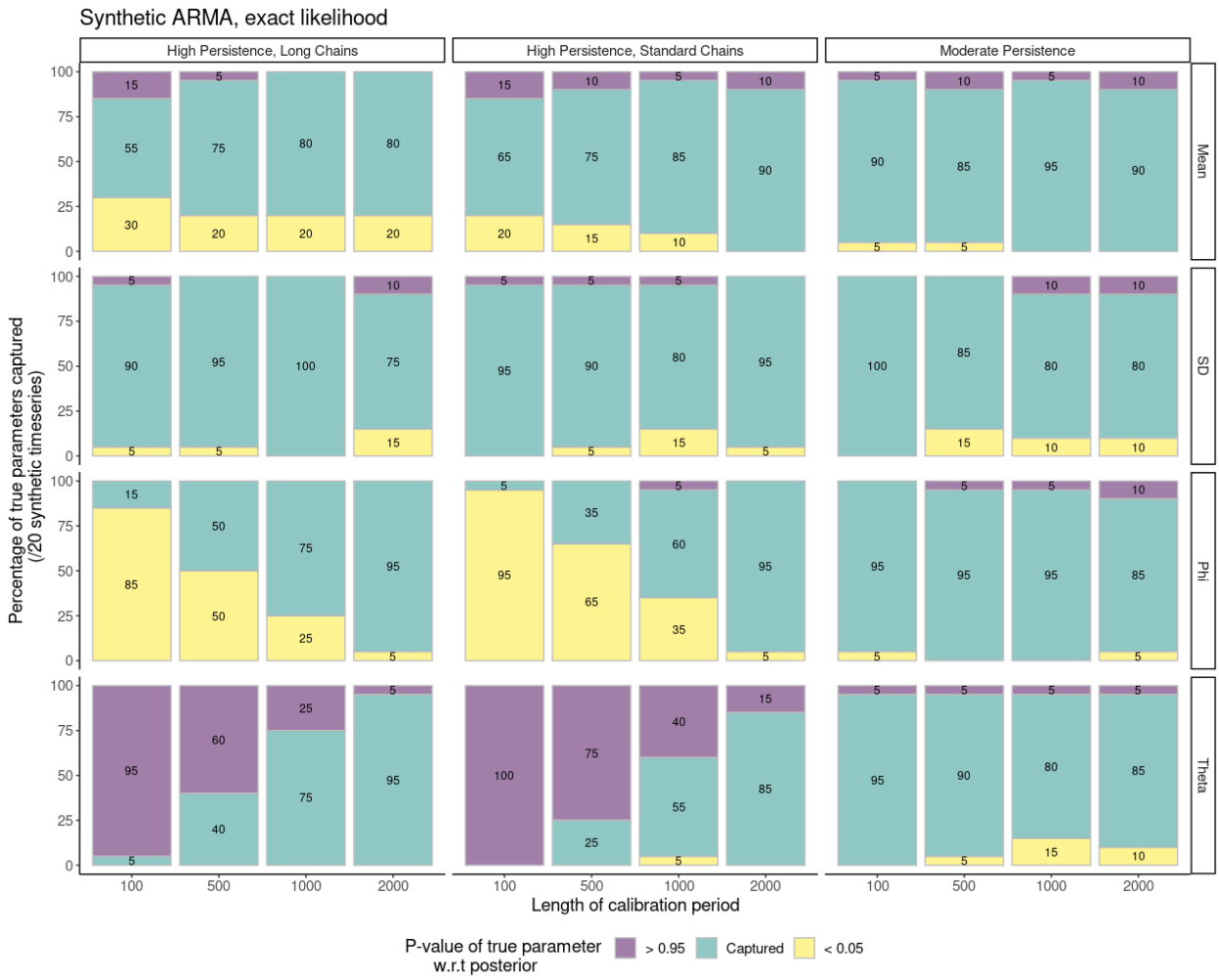

Figure 4-9: Summary of Figure 4-8 with p-values classified as either within the 90% credible interval ('captured') or outside the credible interval (i.e having a value < 0.05 or > 0.95).

4.6.3 500-year ARFIMA timeseries

Results from the 500-year, high persistence ARFIMA simulation are shown in Figure 4-10. The p-value distribution for the D parameter is slightly right skewed, indicating that the posteriors typically underestimated persistence. However, for the 500-year ARFIMA analysis, Figure 4-11 shows that the posterior 90% CI derived from a conditional likelihood consistently contained the true parameter.

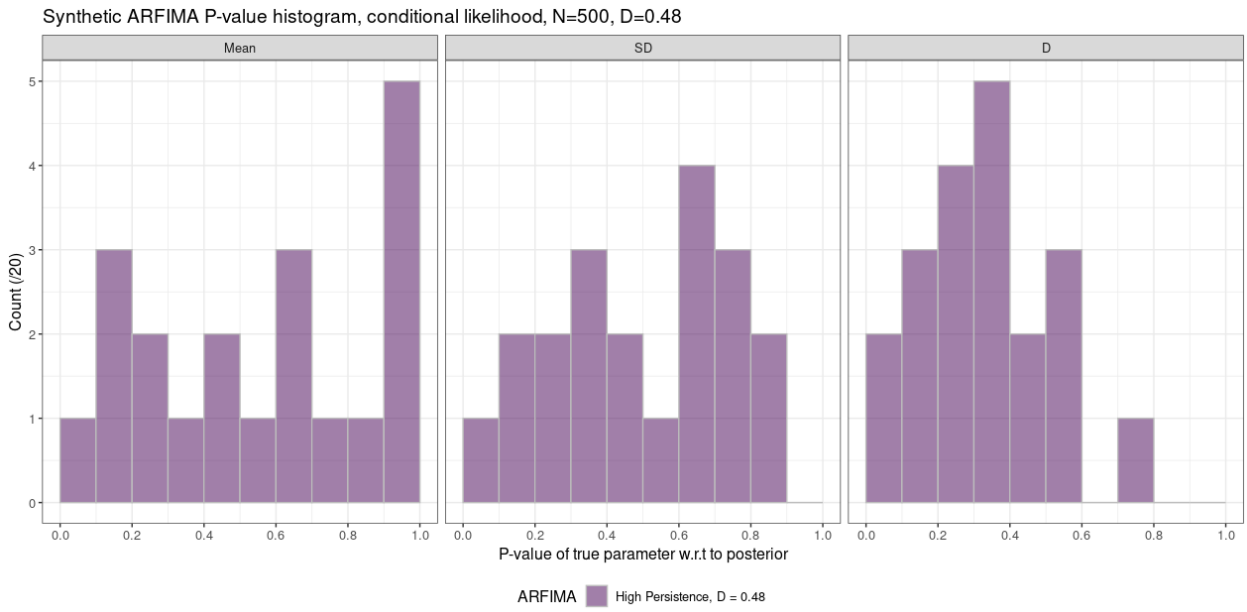

Figure 4-10: P-value distribution for ARFIMA(0,D,0) analysis for timeseries of length 500. A D parameter of 0.48 was used. Conditional likelihood functions were used for inference.

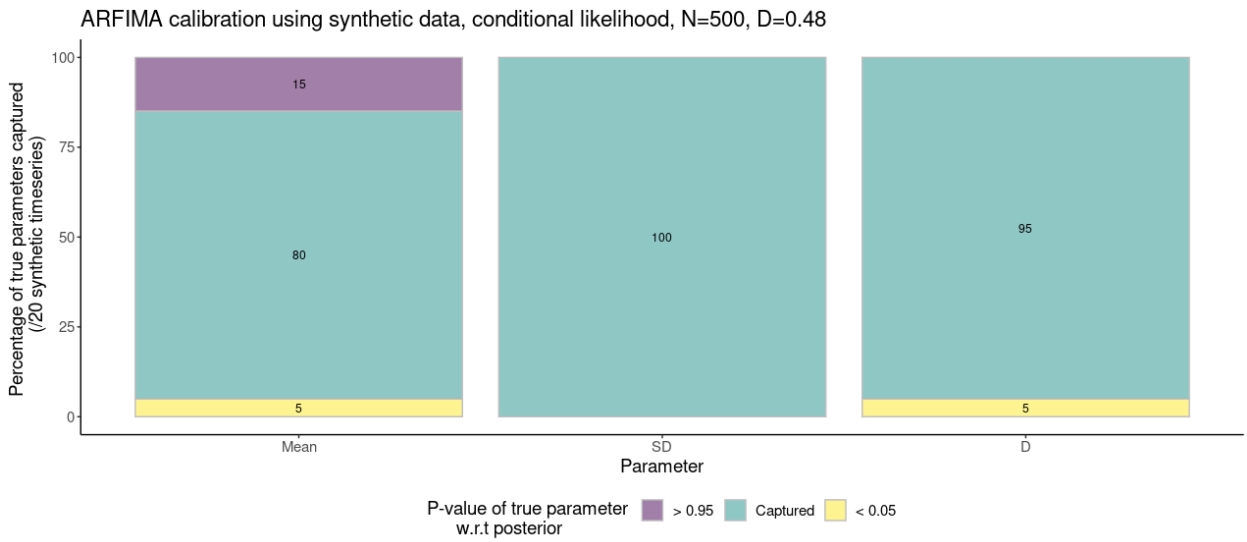

Figure 4-11: Summary of Figure 4-10 with p-values classified as either within the 90% credible interval ('captured') or outside the credible interval (i.e having a value < 0.05 or > 0.95).

4.6.4 Skewed timeseries

For the skewed timeseries, which required inference on an additional transformation parameter, results are broadly similar to the normally distributed timeseries. Aggregate results for different timeseries lengths and persistence/transformation parameters are shown in Table 4-3 (with four combinations of each timeseries length and persistence/transformation parameter), indicating that:

- For the high and moderate persistence ARFIMA timeseries, the posterior 90% credible interval typically contained the true parameter
- For the high persistence, 100-year ARMA timeseries, the posterior 90% credible interval for persistence parameters **ϕ** and **θ** did not contain the true value. However, for longer ARMA timeseries, the posterior credible interval did contain the true **ϕ** and **θ** values.
- For the moderate persistence, 100-year ARMA timeseries, the posterior 90% credible interval for **ϕ** and **θ** typically contained the true value.
- For all simulations, the posterior 90% credible interval typically contained the true Box-Cox **λ** parameter.

Table 4-3: Number of 'true' model parameters captured for the skewed, synthetic timeseries. 'HP' refers to 'high persistence' timeseries, 'MP' refers to 'moderate persistence' timeseries – see Table 4-1.

		ARMA(1,1)				**ARFIMA**			
		100	500	1,000	2,000	100	500	1,000	2,000
HP	**μ**	3/4	4/4	4/4	3/4	4/4	3/4	4/4	4/4
	σ	3/4	4/4	4/4	3/4	4/4	3/4	4/4	4/4
	λ	4/4	4/4	4/4	3/4	3/4	4/4	3/4	4/4
	ϕ	0/4	3/4	4/4	3/4	NA	NA	NA	NA
	θ	0/4	3/4	4/4	3/4	NA	NA	NA	NA
	D	NA	NA	NA	NA	4/4	4/4	4/4	4/4
MP	**μ**	4/4	3/4	3/4	4/4	4/4	4/4	4/4	4/4
	σ	4/4	3/4	4/4	3/4	3/4	4/4	4/4	4/4
	λ	3/4	4/4	4/4	4/4	4/4	3/4	3/4	4/4
	ϕ	4/4	4/4	4/4	3/4	NA	NA	NA	NA
	θ	4/4	4/4	3/4	3/4	NA	NA	NA	NA
	D	NA	NA	NA	NA	4/4	4/4	4/4	4/4

4.7 Discussion and Conclusion

This study explored how sampling bias, conditioning error, and likelihood approximation can influence posterior inference on synthetic timeseries exhibiting moderate and high persistence. It was found that, for 100-year timeseries, approximate conditional likelihoods returned similar results to an exact likelihood. It was also found that, for high persistence timeseries, posteriors inferred from 100-year periods tended to underestimate the true persistence. In contrast, posteriors inferred from moderate persistence timeseries more accurately reflected the underlying true persistence.

The key results differed between the ARFIMA and ARMA models. For the ARFIMA model, although in many cases the p-value distribution was not uniform, the 90% credible interval typically contained the true parameter value (regardless of likelihood function). Therefore, even though the NUTS sampler produced biased posteriors, at the very least the 90% credible interval is potentially a reliable measure of the 'true' parameter.

For the ARMA(1,1) model, long-term persistence combined with short record sampling bias produces posterior distributions that were hard for the NUTS algorithm to explore (irrespective of using an exact or conditional likelihood). However, this should not be viewed as a criticism of the NUTS algorithm (which, relative to other MCMC samplers, is quite sophisticated). Rather, it raises the questions: when inferring stochastic model parameters from a relatively short record, what kind of posteriors should we expect? Is it reasonable to expect parameters indicative of centennial-scale climate variability when the calibration data is too short to suggest otherwise? And, if the inferred stochastic model parameters accurately reproduce various climate statistics, why would we continue looking for alternative parameters? For the ARMA model, only longer timeseries provide sufficient information to properly identify centennial-scale climate variability.

Model specific results aside, this study highlights the difficulties with inferring climate risk from relatively short 100-130-year records, which may be a small sample of a longer-term process exhibiting centennial-scale variability. However, several key issues remain.

For these key issues, first consider that (a) a stochastic model calibrated to ~100 years of climate data can reproduce a variety of observed climate risk metrics (Markonis et al., 2018; Srikanthan and McMahon, 2001), (b) centennial-scale climate variability produces timeseries with extended periods above or below some longer-term mean, and (c) water infrastructure

planning horizons are typically around 50-100 years (Serinaldi, 2015). Therefore, from a practical perspective, the key issue is not whether the posterior contains the 'true' parameter. Rather, it is whether the posterior produces reasonable estimates of risk over the future planning horizon. This will be determined by the rate at which climate varies and the degree to which adjacent 100-year periods are similar. For timeseries that vary slowly, perhaps a biased posterior is still useful for inferring future risk over a 100-year planning horizon?

With respect to using a stochastic model to infer future climate risk, note that in this study we examined stationary models. This may not reflect the real-world behaviour of climate processes, particularly under climate change. However, from a climate risk perspective, understanding if and how a timeseries may be non-stationary requires an understanding of how stationary models behave as they approach non-stationarity. This study suggests that, for 'almost' non-stationary models, parameters will be subject to considerable sampling bias and require a very long timeseries to properly identify. For annual-scale ARMA and ARFIMA models, this will make identifying non-stationarity from observational records difficult.

Although this study has implications for how stochastic models are used and calibrated in climate risk assessment, this analysis was mainly exploratory. Twenty synthetic timeseries is an insufficient sample size to draw robust conclusions about posterior bias. For such a small sample, it is possible to observe a relatively high proportion of 'biased' posteriors (i.e., the true value lies on the tails of the inferred posterior). Consider a binomial distribution with a success probability of 0.9 (analogous to the chance of a 90% credible interval containing the true parameter when the 'true' model is known). Assuming a binomial distribution, from a sample size of 20 there is an approximately 10% chance that four credible intervals will not contain the true parameter value. Although larger samples are desirable, analysing a sufficiently large sample of, say, 100 synthetic timeseries was not computationally feasible. Therefore, the key conclusions drawn from this analysis are:

- For high and moderate persistence timeseries, exact likelihoods and conditional, approximate likelihoods return qualitatively similar posteriors.
- Moderate persistence can be inferred from short, 100-year timeseries using ARMA and ARFIMA models with conditional likelihoods.
- High persistence can be inferred from short, 100-year timeseries using the ARFIMA model D posterior. However, this posterior will be biased towards underestimating persistence.

- High persistence cannot be reliably inferred from short, 100-year timeseries using the ARMA model, due to the complex joint posterior of the $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ parameters. This issue can be alleviated using longer timeseries (that contain ~1,000-2,000 values).
- Skewed timeseries, which required inference on an additional transformation parameter, returned qualitatively similar results to normally distributed timeseries.

4.8 Links with following chapters

Chapter 4 highlights how, even under idealised conditions, stochastic model posteriors can be biased. This insight will guide the interpretation of results in Chapter 5 and Chapter 6. For Chapter 5, which evaluates stochastic model parameter stationarity using proxy records, Chapter 4 can provide a statistical explanation for any potential non-stationarity. For Chapter 6, which uses an ice core record to guide the calibration of stochastic model persistence, Chapter 4 highlights how centennial-scale variability cannot be reliably inferred from short, 100-year hydrological records.

Chapter 5. Assessing stochastic model parameter stationarity over centennial timescales using a global network of millennium-length hydroclimatic proxy records

5.1 Abstract

Drought risk estimates inform the design of water supply infrastructure and management plans. They are typically estimated using statistical timeseries models, called stochastic models, and assume model parameters are constant in time (i.e. the stationarity assumption). However, validating the stationarity assumption is difficult. This is because instrumental rainfall and streamflow records are short. Short records are subject to considerable statistical uncertainty (making it hard to identify clear statistical change, even under global warming) and may not capture long-term climate variability. Palaeoclimate proxy records, which span hundreds/thousands of years, can better assess stationarity because longer record lengths will (a) reduce statistical uncertainty; and (b) contextualise if any recent hydroclimatic changes are consistent with historic climate variability. Therefore, in this study, stochastic model parameter stationarity is assessed using 31 millennium-length hydroclimatic proxy records. Two models capable of reproducing low-frequency climate variability - the ARMA(1,1) and ARFIMA(0,D,0) models - were examined. It was found that the number of models with at least one non-stationary parameter was inconsistent with the stationarity assumption. Furthermore, when examining individual model parameters at multi-centennial and millennial timescales, it was found that (a) mean and standard deviation were typically non-stationary (b) persistence was typically stationary. However, when comparing adjacent 100-year periods, all parameters were either stationary or marginally non-stationary. This suggests that, under natural climate variability, only recent observations can be used to infer future mean and standard deviation, but long-term observations can be used to infer persistence. For the mean and standard deviation, this effectively limits the degree to which parameter uncertainty can be reduced by calibrating models to longer timeseries, resulting in irreducibly 'wide' parameter uncertainty.

5.2 Introduction

Historic drought risk estimates, which inform the design of water supply infrastructure and management plans, are typically derived using stochastic models calibrated to instrumental measurements (available from ~1900 onwards) (Loucks and Van Beek, 2017). These statistical models generate synthetic timeseries with similar statistics to the calibration data. But, by accounting for the inherent randomness and persistence of hydroclimatic timeseries, the synthetic timeseries contain more severe droughts and pluvials than those recorded via instrumental measurements (Matalas, 1967; Srikanthan and McMahon, 2001). This improves multi-year drought risk estimates, leading to better informed water management plans/infrastructure design (Koutsoyiannis, 2000; Stedinger and Taylor, 1982b; Vogel et al., 1999). However, is this approach – and the underlying assumptions – valid under climate variability and change?

A key assumption underpinning stochastic model calibration – and associated drought risk estimates – is one of parameter stationarity (Milly et al., 2008). Parameter stationarity assumes that stochastic model parameters are time invariant (Koutsoyiannis and Montanari, 2015; Montanari and Koutsoyiannis, 2014). This assumption means that, for the same timeseries, a stochastic model calibrated to different time periods should return similar parameters (subject to sampling uncertainty). For various reasons (explained below), this assumption has been difficult to validate. However, recently developed palaeoclimate proxy records can be used to better explore this assumption (Razavi et al., 2015). Therefore, in this study, we will use a global dataset of millennium-length hydroclimatic proxy records to assess if stochastic model parameters are historically stationary.

Assuming or validating stochastic model parameter stationarity has several issues. First, the range of acceptable parameters can be changed by climate shifts (Milly et al., 2008). These shifts could be caused by external forcings, both natural and anthropogenic. Natural forcings include changes in solar insolation, and anthropogenic forcings include increased radiative forcing due to increased $C0_2$ emissions (Ait Brahim et al., 2018; Grose et al., 2015, 2020; Raspopov et al., 2008). Second, proper/unambiguous identification of stochastic model parameters is not possible using short instrumental records because parameter uncertainty is large (Patskoski and Sankarasubramanian, 2015; Serinaldi and Kilsby, 2015; Thyer et al., 2006). This means small/moderate parameter changes cannot be detected (Matalas, 1997).

Third, identifying stochastic models that can reproduce long-term climate variability is difficult because short instrumental records have limited cycles of multi-decadal climate variability (Cook et al., 2022). This makes it hard to determine if a stochastic model calibrated to instrumental measurements is producing realistic low-frequency climate variability (Chapter 3). Any evaluation of stochastic model parameter stationarity should be performed using models capable of simulating natural climate variability. These three issues highlight a need for longer timeseries to validate (a) whether the instrumental record is representative of the longer-term past, and (b) whether stochastic model parameters are consistent across multi-centennial timescales.

Because short instrumental measurements cannot properly assess stochastic model parameter stationarity, in this study we use alternative sources of data: extended palaeoclimate proxy records. These records are derived from climatically sensitive, naturally forming 'layers' (e.g., tree-rings, ice cores). Because these records span hundreds or thousands of years, they better characterise multi-decadal and centennial variability. These records also give evidence for droughts and pluvials that cannot be explained by internal variability alone (Ault et al., 2018, 2014). This is a potential indicator of parameter non-stationarity. Moreover, previous work by Razavi et al. (2015) using Canadian tree-ring records also identified non-stationarity in the proxy mean and Lag-1 autocorrelation. However, this study was focussed on two statistics for one region. There is a need to broadly assess parameter stationarity in the stochastic models used by water managers and hydrologists to quantify drought risk.

The need to assess stochastic model parameter stationarity using millennium-length proxy records was also noted in Chapter 3, which validated the performance of various stochastic models using a global dataset of proxy records. This study found that (i) some models could reproduce observed climate variability when calibrated to the full period (i.e., both instrumental and pre-instrumental); and (ii) regardless of stochastic model used, 'instrumental-period' stochastic models were unable to reproduce pre-instrumental climate variability and drought risk.

Although Chapter 3 contained a preliminary assessment of statistical stationarity, a more rigorous assessment is needed. In this study, a more rigorous assessment was conducted to explore whether poor 'instrumental-period' stochastic model performance was due to (a) parameter non-stationarity between instrumental and pre-instrumental periods or (b) not

considering parameter uncertainty when calibrating to the instrumental-period. We will also explore whether stochastic model parameters are stationary across multi-centennial timescales.

Validating or invalidating the stationarity assumption has implications for estimating drought risk and, subsequently, determining appropriate water management decisions and infrastructure design. Decisions include trigger points for implementing water use restrictions, infrastructure design includes determining reservoir size or whether to include desalination. These decisions and infrastructure are designed to ensure supply under extreme droughts. Stationarity implies that (a) a stochastic model calibrated to the instrumental record can properly characterise this drought risk, and (b) historic risk is representative of future risk. Parameter non-stationarity in the palaeoclimate record means that future risk is, potentially, greater than historic risk (irrespective of global warming). Such non-stationarity would suggest that existing water supplies may be more vulnerable than currently assumed and that adaptation is required to ensure future water security.

Considering these issues, in this study two research questions will be answered using extended palaeoclimate proxy records:

1. Are instrumental/pre-instrumental stochastic model parameters stationary?
2. Are stochastic model parameters stationary across multi-centennial timescales?

Stationarity will be evaluated within a Bayesian calibration framework. For the same proxy record, this involves inferring and comparing stochastic model posteriors from different time periods.

5.3 Data

The proxy records used in this study are a subset from Chapter 3, where these records were used to evaluate different stochastic models. From the original study, which examined 45 tree-ring, snow accumulation, and ice core Na+ records, we selected 31 records – removing all ice core Na+, except for the Law Dome summer sea salt record. Na+ records were not considered in this study because (a) the links between hydroclimate and Na+ are poorly understood (except for the Law Dome record) and (b) the presence of non-stationary variance in some Na+ timeseries, the cause of which is unknown and not necessarily related to hydroclimate.

The resultant dataset, shown in Table 5-1 and Figure 5-1, comprises of 31 millennium-length proxy records (25 tree-ring records, 5 snow accumulation records, and 1 ice core Na+ record). Links between each proxy record and hydroclimate are described in the relevant Table 5-1 citations.

Table 5-1: Proxy records used in this study

Record	Continent	Period Analysed	Proxy Type	Reference	ITRDB Code
Dulan, China	Asia	159-1993	Tree Ring	Sheppard et al., 2004	chin006
Delingha, China	Asia	1000-2003	Tree Ring	Shao et al., 2005	chin050-chin054
Uurgat, Mongolia	Asia	488-2013	Tree Ring	Hessl et al., 2018	mong042
Khorgo, Mongolia	Asia	15-2014	Tree Ring	Hessl et al., 2018	mong041
Southern Finland	Europe	670-2012	Tree Ring	Helama, Meirläinen and Tuomenvirta, 2009	finl030-finl034
Mount Smolikas, Greece	Europe	730-2015	Tree Ring	Klippel et al., 2018	gree013-gree016
Flowerpot, Canada	North America	650-1989	Tree Ring	Buckley et al., 2004	NA
Whirlpool Point, Canada	North America	896-2008	Tree Ring	Case and MacDonald, 2003	cana220
Cedar Knob, USA	North America	950-1998	Tree Ring	Maxwell et al., 2011	wv005
Barranca de Amealco, Mexico	North America	880-2008	Tree Ring	Stahle et al., 2011	mexi047
Tavaputs Plateau, USA	North America	6-2005	Tree Ring	Knight, Meko and Baisan, 2010	ut530
Mount San Gorgonio, USA	North America	651-1998	Tree Ring	MacDonald, 2007	ca051
Southern Colorado Plateau, USA	North America	570-1990	Tree Ring	Salzer and Kipfmuller, 2005	az570
Jemez Mountains, USA	North America	824-2007	Tree Ring	Touchan et al., 2011	nm583
Upper Arkansas Basin, USA	North America	216-2007	Tree Ring	Woodhouse, Pederson and Gray, 2011	Multiple

Upper Klamath Basin, USA	North America	1000-2010	Tree Ring	Malevich, Woodhouse and Meko, 2013	or093
El Malpais,USA	North America	5-2004	Tree Ring	Stahle et al., 2009	nm580
Bear River, USA	North America	916-2013	Tree Ring	DeRose et al., 2015	ut541
Summitville, USA	North America	10-2009	Tree Ring	Routson, Woodhouse and Overpeck, 2011	co656
Atlas Mountains	Africa	985-1984	Tree Ring	Esper et al., 2007	morc014
Choctawhatchee River	North America	993-1992	Tree Ring	Stahle et al., 2012	fl001
Lee's Ferry	North America	760-2005	Tree Ring	Meko et al., 2007	ut529
Colorado River	North America	985-1984	Tree Ring	MacDonald, Kremenetski and Hidalgo, 2008	nv516
Sacramento River	North America	997-1996	Tree Ring	MacDonald, Kremenetski and Hidalgo, 2008	or062
Albermarle Sound	North America	934-1985	Tree Ring	Stahle, Burnette and Stahle, 2013	va021
Law Dome Snowfall	Antarctica	17-2016	Ice Core Accumulation	Jong et al., 2022	NA
Roosevelt Island	Antarctica	13-2012	Ice Core Accumulation	Winstrup et al., 2019	NA
West Antarctic Ice Sheet Divide	Antarctica	8-2007	Ice Core Accumulation	Sigl et al., 2016	NA
SPICE Snowfall	Antarctica	15-2014	Ice Core Accumulation	Winski et al., 2019	NA
Quelccaya Ice Core	South America	683-2009	Ice Core Accumulation	Thompson et al., 2013	NA
Law Dome Sea Salt	Antarctica	17-2016	Ice Core Na	Jong et al., 2022	NA

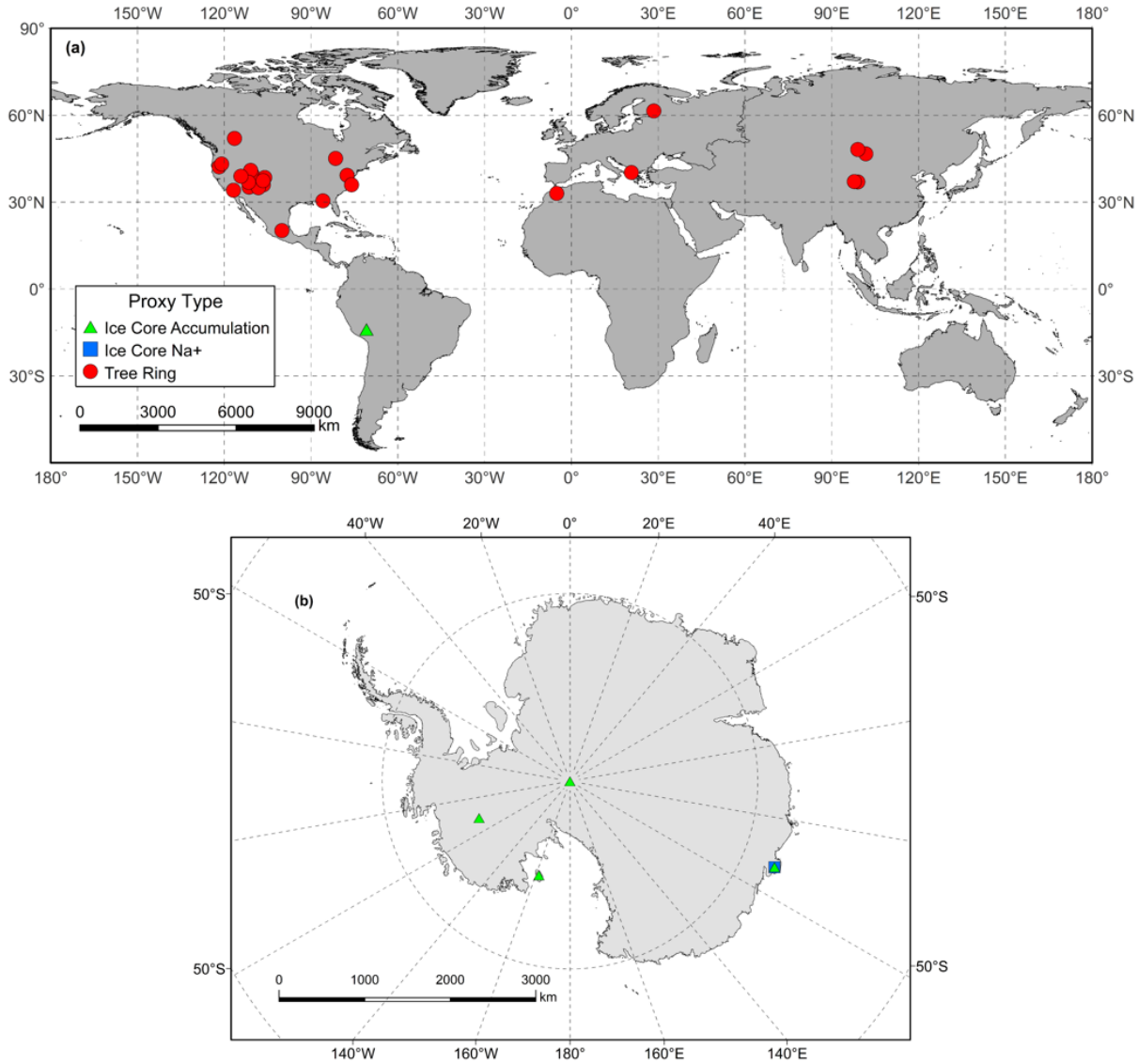

Figure 5-1: Location of proxy records used in this study.

5.4 Stochastic models

The Autoregressive Fractionally Integrated Moving Average (ARFIMA) (0,D,0) and Autoregressive Moving Average (ARMA) (1,1) models were used in this study. Chapter 3 demonstrated that these models are (a) able to reproduce low-frequency variability when calibrated to millennium-length hydroclimatic proxy records; and (b) unable to reproduce low-frequency variability when calibrated (evaluated) on the instrumental (pre-instrumental) period of the same proxy records. For more detailed explanations of respective models, refer to Chapter 3 and references therein – here we will only give a quick overview of both models.

For a timeseries y, both the ARMA(1,1) and ARFIMA(0,D,0) models represent individual observations at time t as a conditional mean $\bar{y}_t$ plus a normally distributed residual ε_t with zero mean and variance σ^2_ε

$$y_t = \bar{y}_t + \epsilon_t \qquad \text{Equation 5-1}$$

$$\epsilon_t \sim N(0, \sigma_\epsilon{}^2) \qquad \text{Equation 5-2}$$

For the ARMA(1,1) model, the conditional mean is calculated as:

$$\bar{y}_t | y_{t-1}, \epsilon_{t-1} = \mu + \phi(y_{t-1} - \mu) + \theta\epsilon_{t-1} \qquad \text{Equation 5-3}$$

Where $\boldsymbol{\mu}$ is the timeseries mean, $\boldsymbol{\phi}$ is the lag-1 autoregressive parameter, and $\boldsymbol{\theta}$ is the lag-1 moving average parameter. In stationary mean models, both $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are restricted to values between -1 and 1.

For the ARFIMA(0,D,0) model, the conditional mean can be calculated as an infinite moving average of prior residuals:

$$\bar{y}_t | \epsilon_{t-1}, \dots, \epsilon_{t-k} = \mu + \sum_{k=1}^{\infty} a(k)\epsilon_{t-k} \qquad \text{Equation 5-4}$$

Where

$$a(k) = \frac{\Gamma(k + D)}{\Gamma(k + 1)\Gamma(D)} \qquad \text{Equation 5-5}$$

However, as k →∞, weights can be approximated as a(k) as

$$a(k) \sim \frac{1}{\Gamma(D)} k^{D-1} \qquad \text{Equation 5-6}$$

Considering all previous timeseries values when estimating the conditional mean can significantly increase calibration time. To reduce calibration time, the number of previous

weights considered in the conditional mean estimate can be reduced to 100 (following Haslett and Raftery (1989) and Chapter 4), resulting in a conditional mean approximated as:

$$\bar{y}_t | \epsilon_{t-1}, \dots, \epsilon_{t-k} = \mu + \sum_{k=1}^{\min(100,t)} a(k)\epsilon_{t-k} \qquad \text{Equation 5-7}$$

For stationary models, D is restricted to values between -0.5 and 0.5.

For both models, we were interested in analysing the Autocorrelation Function (ACF), which describes the relationship between lagged timeseries values. For the ARMA(1,1) model, the ACF is a function of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, and for lag *k* can be written as:

$$\rho(1) = \frac{(1-\phi\theta)(\phi-\theta)}{1+\phi^2-2\phi\theta}$$
$$\rho(k) = \rho(1)\phi^{k-1} \; for \; k > 1 \qquad \text{Equation 5-8}$$

For the ARFIMA(0,D,0) model, the ACF for lag *k* can be written as:

$$\rho(k) = \frac{\Gamma(1-D)\Gamma(k+D)}{\Gamma(D)\Gamma(k+1-D)} \qquad \text{Equation 5-9}$$

5.5 Methods

5.5.1 Model calibration

To assess stochastic model parameter non-stationarity, parameter uncertainty must be quantified. There are various approaches for doing so. In this study, a Bayesian approach was used. Given some observed data, Bayesian methods use a likelihood function to infer a distribution of suitable model parameters (called the posterior distribution) (Gelman et al., 2013). The posterior is explored and quantified using Monte Carlo 'chains', which iteratively assess the relative likelihood of randomly generated proposed parameters (Brooks, 1998). As different proposed parameters are accepted or discarded, the posterior distribution is inferred and parameter uncertainty is quantified.

5.5.1.1 Inferring stochastic model parameter posteriors

There are various Bayesian algorithms that can be used to infer posterior distributions – we used the No U-Turn sampling (NUTS) algorithm (Homan and Gelman, 2014). NUTS is an efficient algorithm that can sample from complex posterior distributions. It is a variant of the Hamiltonian Monte Carlo (HMC) algorithm, which automatically adapts to the geometry of the target posterior distribution during sampling (Upadhyay et al., 2015). These algorithms work by simulating the motion of a particle through a high-dimensional space (each dimension corresponds to a parameter of the probability distribution being sampled) (Betancourt, 2018). A combination of random and deterministic steps are used to guide the particle towards regions of high probability – this 'guidance' makes these algorithms more efficient than standard random walk/Metropolis sampling algorithms used in Bayesian inference (Betancourt, 2018).

For each model calibration, 20,000 posterior samples were taken from eight chains (2,500 samples per chain, plus a prior 500 samples as a burn-in period). After simulation, the 20th sample from each chain was extracted and combined into the final posterior sample. This maximised independence between posterior samples (Jones and Qin, 2022). Chains were initialised at the calibration data's sample mean; sample standard deviation; a $\boldsymbol{\lambda}$ value of one (corresponding to no transformation); and persistence parameters of zero (corresponding to no persistence).

5.5.2 Comparing stochastic model parameter posteriors

5.5.2.1 Overview of analyses

Four different analyses were performed to assess stochastic model parameter stationarity across different timescales. To simplify comparisons across each proxy timeseries, Analyses 1-3 were performed on the most recent 1,000 year period of each record. For the nine records at least 2,000 years in length, Analysis 4 was performed on the most recent 2,000-year period.

- Analysis 1: Comparison of instrumental and pre-instrumental stochastic model posteriors. For this analysis, the final 100-years of each proxy record was taken as the instrumental-period. The 900 years prior were considered the pre-instrumental period.
- Analysis 2: Split-sample comparison of stochastic model model posteriors. For this analysis, each 1,000 record was split into two 500-year halves and compared.
- Analysis 3: Comparison of 100-year subsets. For this analysis, each record was separated into 10 non-overlapping 100-year periods. Adjacent periods were then compared.

- Analysis 4: For the 9 records 2,000-years in length, performing a split-sample comparison on the most recent and previous 1,000-year periods. Out of the initial sample of 31 proxy records, 9 were at least 2,000-years in length.

With respect to Analysis 3, for some records, not all 100-year periods returned a satisfactory model calibration using standard priors (i.e. MCMC chains did not mix - Gelman and Rubin (1992)). Updating model calibrations for each poorly performing period was not feasible. Instead, adjacent 100-year periods where both model calibrations were adequate were identified. Out of these periods, 5 from each record were randomly selected for further analysis.

To determine if parameters were stationary, we derived the 'difference posterior'. The 'difference posterior' was calculated by randomly sampling from corresponding posteriors, then subtracting one sample from the other (e.g. for the same record and model, randomly sampling from corresponding posteriors derived from different time periods, then subtracting one value from the other). When analysing this 'difference posterior', the parameter was considered stationary if the 90% credible interval contained zero.

For each analysis, record, and model parameter, these general steps were followed to evaluate parameter stationarity:

1. Partition the proxy timeseries into separate calibration periods.
2. Infer parameter posteriors for each calibration period.
3. Calculate the difference posterior and its 90% credible interval. If the credible interval contained zero, parameters were considered stationary.

Figure 5-2 shows an example method schematic for Analysis 1, using the Law Dome Summer Sea Salt proxy record and the ARFIMA(0,D,0) model.

Note that both the ARMA(1,1) and ARFIMA(0,D,0) models have a mean, standard deviation, and Box-Cox parameter. However, both models represent persistence using different parameters – ARMA(1,1) with ϕ and θ, ARFIMA(0,D,0) with **D**. Furthermore, the ARMA(1,1) ϕ and θ posteriors will be highly correlated. To better compare persistence stationarity across both models, the posterior ACF was calculated for both models (using EQUATIONS). Persistence non-stationarity was evaluated by comparing the Lag-1, Lag-30, and Lag-100 posteriors. Results were the same for each ACF lag, so only Lag-1 is presented here.

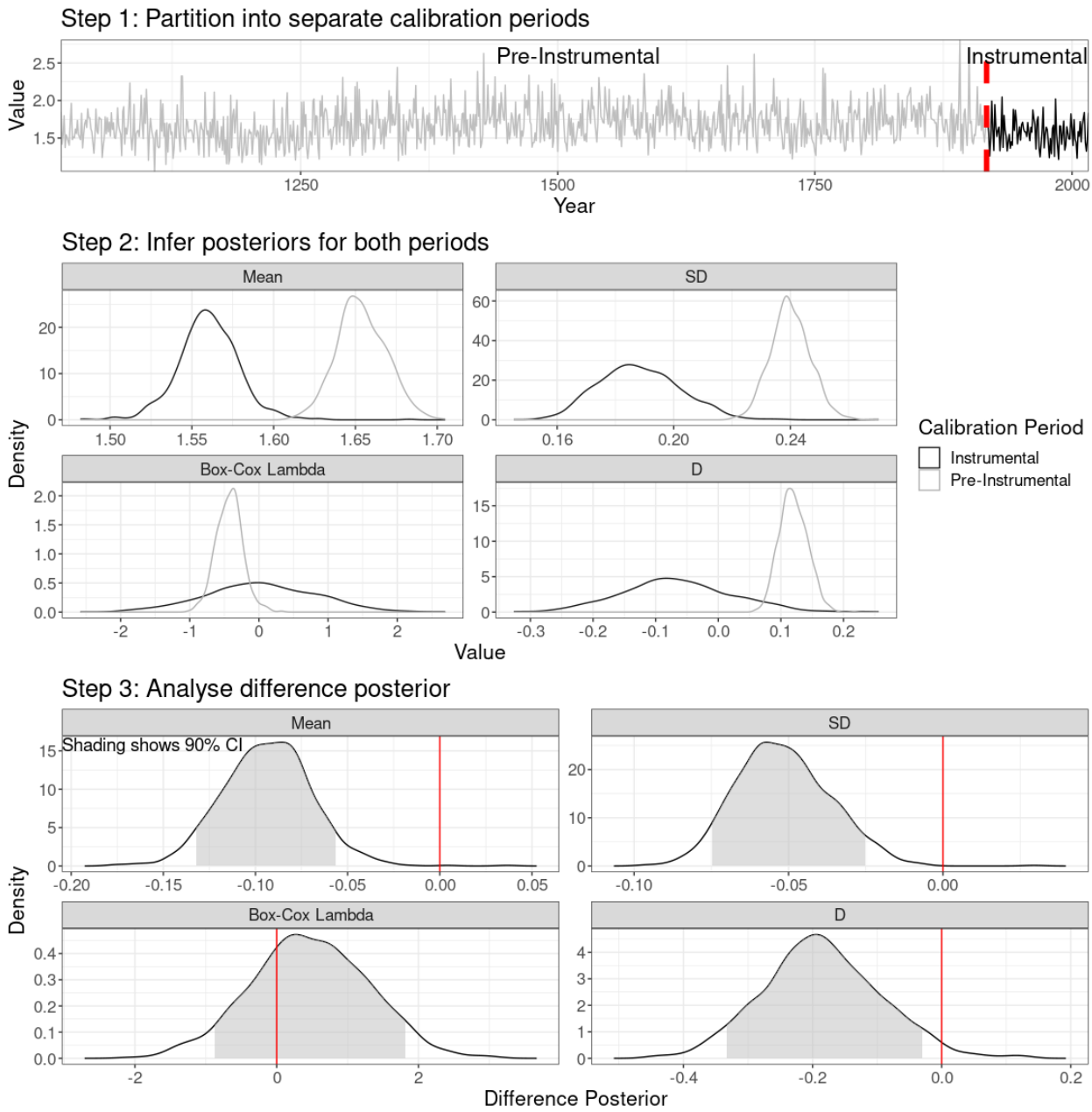

Figure 5-2: Schematic of method. Experiment 1 is given as an example.

For each analysis, we then aggregated the total number of stationary and non-stationary parameters across the entire proxy sample. By aggregating results, we can make general inferences about stochastic model parameter stationary using the entire proxy sample. For example, assuming a null hypothesis of parameter stationarity, for each analysis we can infer:

1. The expected number of stationary parameters, and whether the proxy sample is consistent with this expectation.

2. The expected number of stationary models (i.e. all parameters are stationary), and whether the proxy sample is consistent with this expectation. We performed this analysis to evaluate if there were many proxy records where at least one parameter was non-stationary or if there were a small number of proxy records where all parameters were non-stationary.

For inference (1), consider two methodological choices. First, we assumed a null hypothesis of parameter stationarity. Second, we tested this null hypothesis using the 90% credible interval of the posterior difference. Under the null hypothesis, the 'true' posterior difference is zero, which will lie within the 90% credible interval 90% of the time. Therefore, when aggregating results across the proxy sample of size 'N', which returned 'Y' stationary parameters, we can use a binomial distribution (with a 'success' probability of 0.9) to estimate the probability of observing these 'Y' stationary parameters. If that probability is low, we can make more general inferences that stochastic model parameters are non-stationary, irrespective of calibration data.

For inference (2), we can also use the binomial distribution to infer the probability of observing 'Y' stationary models (i.e. all parameters are stationary). To do so, we can calculate the probability of all parameters being stationary for a single model. This calculation uses a binomial distribution with a 'success' probability of 0.9 and a sample size of 4; the probability that all parameters are stationary is 0.6561. We can use this value of 0.6561 as the 'success' probability in another binomial distribution, which, for a proxy sample size of 'N', calculates the probability of observing 'Y' stationary models.

Figure 5-3 provides an example for how the sample-scale results were tested for significance using a binomial distribution with 0.9 success probability. Note that the proxy sample size of 31 is only applicable to Experiment 1 and Experiment 2. Experiment 3 contained 155 samples, and Experiment 4 contained 9 samples.

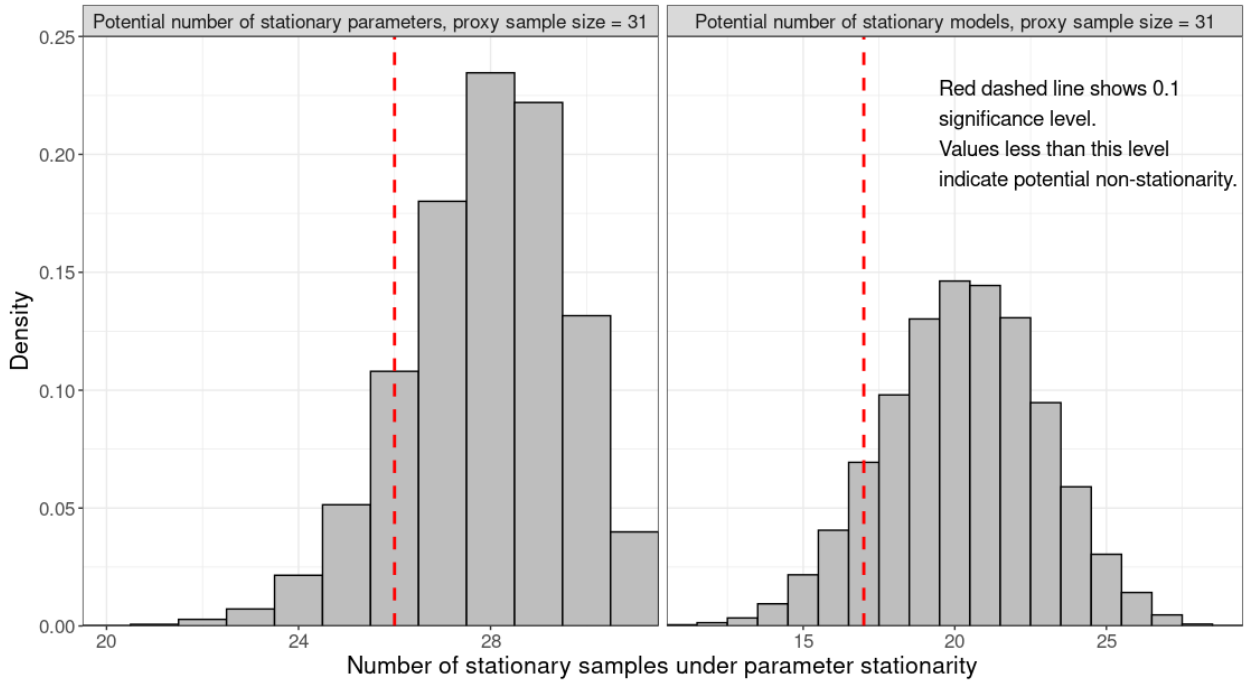

Figure 5-3: Example of how individual results are aggregated and tested for significance using a binomial distribution.

5.5.3 Evaluating model residuals

Reasonable comparison of model posteriors requires that both ARMA(1,1) and ARFIMA(0,D,0) assumptions are met. These models assume that residuals are independent (i.e., have no serial correlation or periodicity), have a mean of zero, and follow a normal distribution. If these assumptions are violated, estimates of posterior uncertainty may be biased (Kavetski et al., 2006), which in turn biases any posterior comparison.

Evaluating model residuals for normality and serial dependence is further complicated by posterior uncertainty. Instead of a single residual set to evaluate (as is the case for maximum likelihood optimisation), there are several (one residual set for every posterior sample). To account for this uncertainty, the following residual diagnostic checks were performed for each model calibration:

Residual mean was assessed by calculating the residual mean for each residual set, then checking if the 90% credible interval contained zero.

Normality was assessed via a Quantile-Quantile plot (QQplot). These compare the observed residual quantiles against the theoretical quantiles for a normally distributed variable. For each

observed quantile, 90% credible intervals were calculated. Residuals were considered normal if (a) 90% of the theoretical quantiles were within the corresponding observed credible interval or (b) the median residual set returned a Shapiro-Wilks test p-value > 0.1.

As a further check on residual normality, we also calculated the L-Skew and L-Kurtosis of each residual set. The percentile rank of the theoretical Normal L-skew and L-Kurtosis were then calculated. This served as a check on the performance of the Box-Cox transformation, which aims to remove skew and does not ensure normality.

To evaluate serial independence, we examined both the residual Lag-1 autocorrelation and residual cumulative periodogram. For the Lag-1 autocorrelation, residuals were considered independent if (a) the corresponding 90% credible interval contained zero; or (b) if the median Lag-1 autocorrelation returned a p-value >= 0.1 for the corresponding Pearson correlation test statistic. Check (b) is less rigorous than check (a) but was necessary because, in many cases, this posterior contained values close to zero (but not zero).

As a further check on serial independence, we also evaluated periodicity using the residual cumulative periodogram. This compares the standardised frequency of the residuals against the cumulative spectral power. For white noise (i.e., no periodicity), the cumulative periodogram increases by a factor of 0.5 times the standardised frequency. Deviations from this line are indicative of periodicity in the residuals. To evaluate periodicity, the 90% credible interval of the cumulative periodogram was calculated at each standardised frequency. Residuals were considered independent (i.e. no periodicity) if either (a) 90% of the credible intervals contained the theoretical white noise line or (b) the median cumulative periodogram returned Kolmogorov-Smirnov test p-value >0.1 when compared against the theoretical white noise line.

Figure 5-4 shows an example residual diagnostic plot for the Law Dome summer sea salt record (Analysis 1). For both periods, the residuals have mean zero, have no periodicity, and are approximately normally distributed. For the instrumental period, the Lag-1 autocorrelation is zero. For the pre-instrumental period, the Lag-1 autocorrelation is slightly negative, however the Lag-1 median is insignificant.

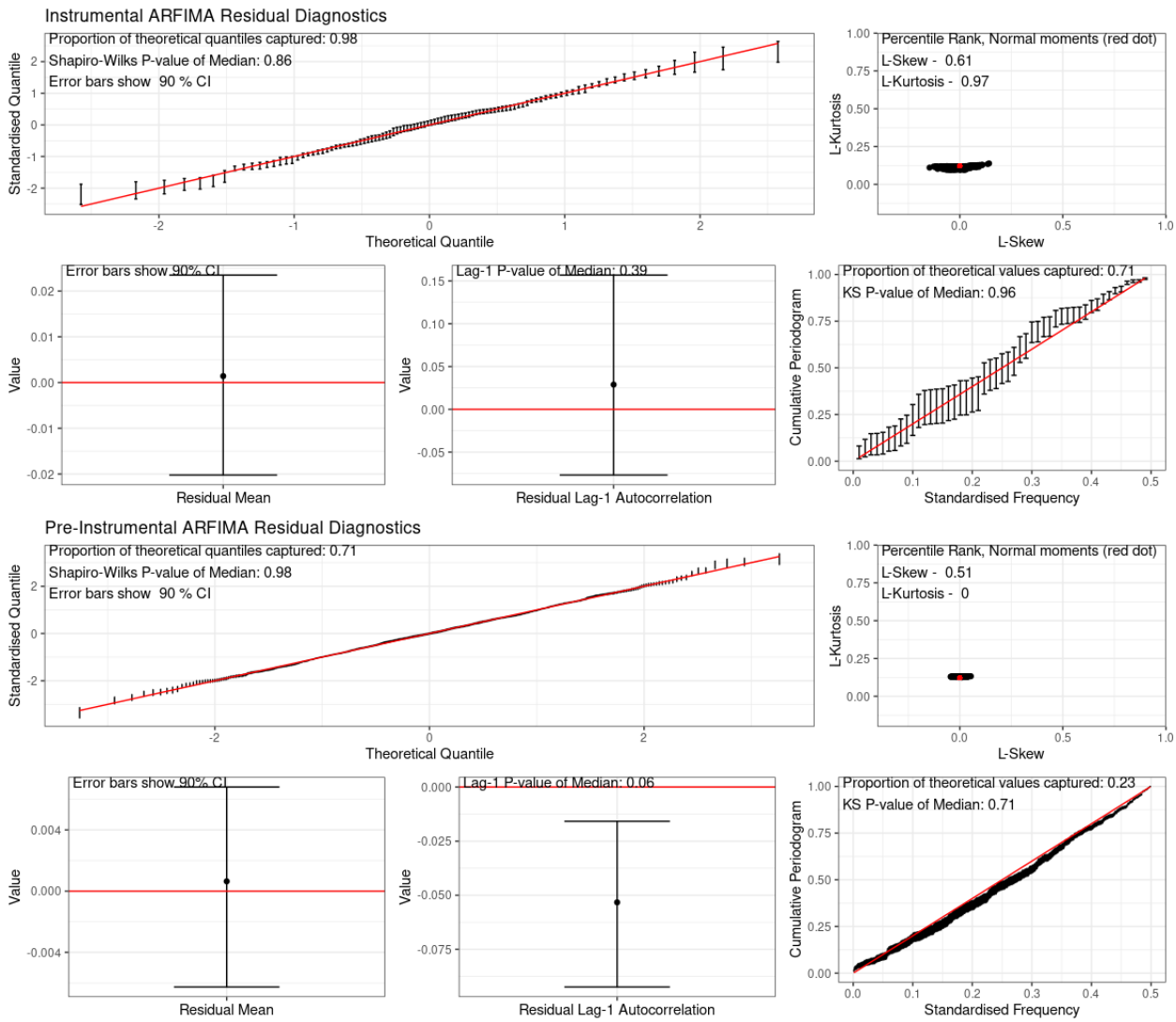

Figure 5-4: Example of how residual diagnostics were evaluated in this study.

5.5.4 Study assumptions and potential methodological limitations

Before presenting and interpreting results, there are some statistical assumptions underpinning our analysis that must be discussed first. When evaluating parameter stationarity, the use of proxy records, ARMA and ARFIMA models combined with Bayesian calibration methods come with three implicit assumptions:

- Assumption 1: That the proxy records contain accurate and relatively unbiased hydroclimatic climate information.

This is an assumption that, to varying extents, must be made for any study using palaeoclimate proxy records. To minimise potential biases in the proxy record sample, we selected records from peer reviewed articles with demonstrated links to hydroclimate (Chapter 3). However, even with this selection approach, there remains a potential for these proxy records to have

confounding, non-hydroclimatic signals (e.g. a temperature signal, or an artificial signal introduced by statistical processing of raw proxy measurements – see Chapter 3).

- Assumption 2: That the ARMA and ARFIMA models are 'correct' models for describing annual-scale variability.

This is an assumption made for any Bayesian inference that does not consider multiple model 'hypotheses' simultaneously (Gelman et al., 2013). Although this is a potentially limited assumption, we emphasise that the ARMA(1,1) and ARFIMA(0,D,0) models are capable of reproducing realistic annual-scale variability, can reproduce a wide variety of autocovariance functions, and are parsimonious (which avoids overfitting). Hence, they were a reasonable choice. However, results may differ for other stochastic models. The proxy dataset will be made publicly available for those who wish to test other models.

- Assumption 3: For the specified likelihood function, that the Bayesian MCMC calibration algorithm, in this case the NUTS algorithm, can adequately explore and define the parameter posterior.

For 100-year samples of a process exhibiting centennial-scale variability, this may not be the case. Chapter 4 demonstrated that, for ARMA(1,1) ϕ and θ parameters commensurate with centennial-scale variability, the 100-year joint posterior is highly correlated, spans the -1 to 1 stationary zone, and bimodal. Such posteriors are difficult to infer, even for the sophisticated NUTS algorithm, and the 100-year posterior will probably not contain the underlying 'true' parameter. Only longer timeseries provide the NUTS algorithm with sufficient information to properly explore the ARMA(1,1) 'centennial-scale variability' parameter space. In contrast to the ARMA(1,1) model, Chapter 4 demonstrated that, for the ARFIMA(0,D,0) model, a 100-year timeseries was sufficient to infer persistence parameters close to the non-stationary zone, but with a tendency to underestimate the underlying true persistence.

Although the ARMA and ARFIMA posteriors may be biased for short samples of a timeseries exhibiting long-term persistence, these posteriors are still able to reproduce various climate statistics from the calibration period. Therefore, identifying parameter, even for adjacent 100-year periods, is still a useful heuristic for indicating whether historic observations are representative of future risk, particularly water infrastructure planning horizons (which are around 30-50 years).

In short, even a biased posterior can be useful for inferring climate risk, but this bias means we must interpret parameter stationarity or non-stationarity with caution. Any evidence against stationarity may not be indicative of some underlying physical changes in hydroclimatic processes represented in the proxy records. Centennial-scale variability, ARMA and ARFIMA model deficiencies, complex posterior shapes, and limitations with MCMC samplers mean that non-stationarity could be inferred from some longer-term stationary process. However, even with these limitations, evidence against stationarity will indicate that climate risk - inferred from seemingly 'good' stochastic models calibrated to historic observations with advanced Bayesian methods - may not represent future climate risk.

5.6 Results

5.6.1 Parameter and model stationarity

Figure 5-5 summarises the proportion of stationary and non-stationary parameters (top) and nonstationary models (bottom) for each analysis and model. The red dashed line corresponds to the proportion of stationary parameters or models expected under a null hypothesis of parameter stationarity (significance level of 0.1). P-values for each parameter or model are also shown.

For Analysis 1, which compared instrumental and pre-instrumental periods, we can see that:

- For both models, evidence against stationarity in the mean parameter (a p-value of 0.03 for ARFIMA and 0.01 for ARMA).
- For both models, marginal evidence supporting stationarity in the standard deviation (a p-value of 0.08 for both models) and evidence supporting stationarity in the Box-Cox parameter (a p-value of 0.38 for both models).
- Evidence against stationarity for ARFIMA persistence (a p-value of 0.03), and evidence supporting stationarity for ARMA persistence (a p-value of 0.19).
- For both models, marginal evidence against stochastic models having only stationary parameters (a p-value 0.08 for ARFIMA and 0.04 for ARMA).

For Analysis 2, which compared 500-year periods, we can see:

- For both models, evidence supporting stationarity in the mean parameter (a p-value of 0.08 for ARFIMA and 0.19 for ARMA).

- Marginal evidence against stationarity in the standard deviation for the ARFIMA model (a p-value of 0.01) and marginal evidence against stationarity in the standard deviation for the ARMA model (a p-value of 0.08)
- For both models, evidence supporting stationarity in persistence (a p-value of 0.19 for ARFIMA and 0.83 for the ARMA).
- For both models, evidence supporting stationarity in the Box-Cox parameter (a p-value of 0.19 for both models).
- Marginal evidence against all ARFIMA models having only stationary parameters (a p-value of 0.04) and evidence supporting ARMA models having only stationary parameters (a p-value of 0.24).

For Analysis 3, which compared 100-year periods, we can see:

- For both models, evidence in favour of stationarity for the mean parameter (a p-value of 0.98 for ARFIMA and 0.14 for ARMA) and persistence (a p-value of 1 for ARFIMA and 0.99 for ARMA).
- Evidence support stationarity in the standard deviation for the ARFIMA model (a p-value of 0.14) and evidence against stationarity for the ARMA model (a p-value of 0.01).
- Marginal evidence against stationarity in the ARFIMA Box-Cox parameter (a p-value of 0.09) and evidence supporting stationarity for the ARMA Box-Cox parameter (a p-value of 0.79).
- For both models, evidence that all models are comprised of only stationary parameters (a p-value of 0.97 for ARFIMA and 0.35 for ARMA).

For Analysis 4, which compared 1,000-year periods, we can see:

- For both models, evidence against stationarity in the mean parameter (a p-value of 0.05 for ARFIMA and <0.01 for ARMA) and standard deviation (a p-value <=0.01 for both models).
- For both models, evidence supporting persistence stationarity (a p-value of 0.23 for ARFIMA and 0.61 for ARMA).
- For both models, evidence supporting stationarity in the Box-Cox parameter (a p-value of 0.23 for both models).
- For both models, evidence against all models having only stationary parameters (a p-value of 0.05 for ARFIMA and <0.01 for ARMA).

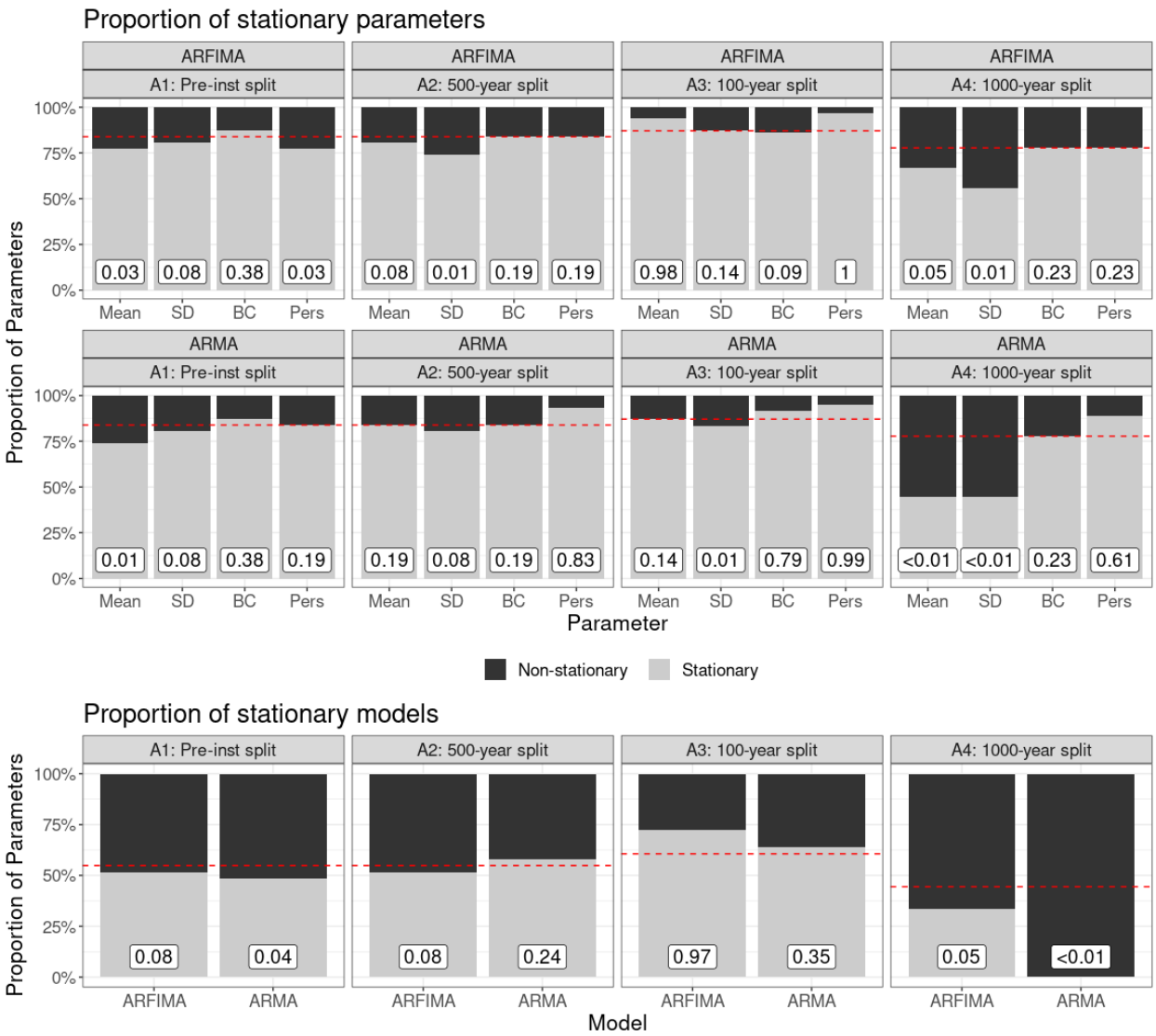

Figure 5-5: Aggregated results for each experiment. 'Pers' refers to the theoretical ACF Lag-1 posterior (results were the same for other lags). Numbers on each bar show the corresponding P-value. Top: Proportion of non-stationary parameters across each experiment. Bottom: Proportion of model comparisons with at least one non-stationary parameter for each experiment. P-values show the probability of observing the number of stationary parameters or models under a null hypothesis of stationarity. The red dashed line shows the 10% significance level.

5.6.2 Residual diagnostics

A summary of residual diagnostics for each model is shown in Figure 5-6. For all analyses, residuals typically had a mean of zero and were consistent with white noise (either the Lag-1 autocorrelation was insignificant or the cumulative periodogram was consistent with white noise). However, residuals consistently violated the normality assumption.

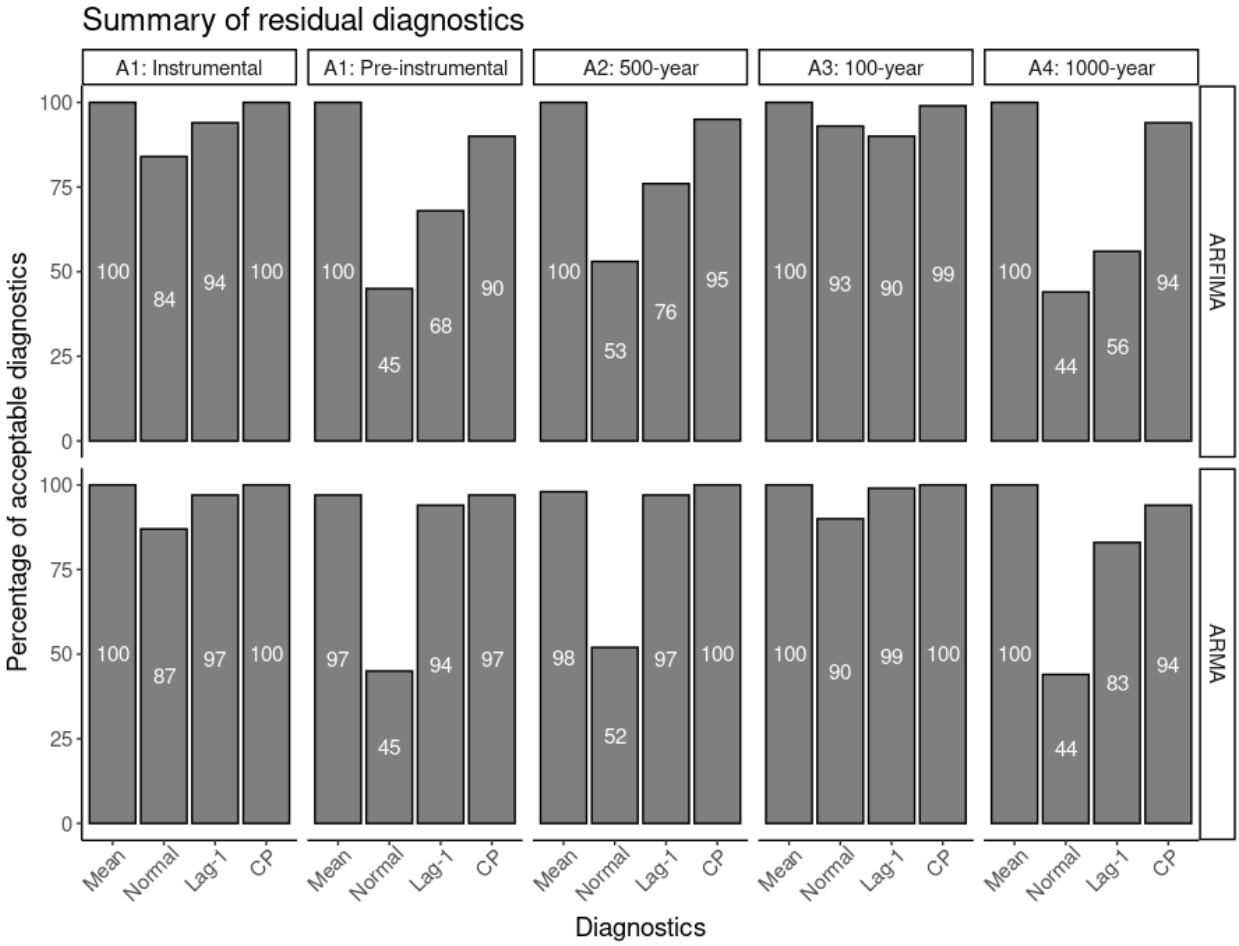

Figure 5-6: Summary of residual diagnostics for each calibration scenario.

As an additional check of the residual marginal distribution, we examined if residuals were normally distributed, skewed, kurtotic, or skewed and kurtotic. This was to evaluate the ability of the Box-Cox transformation to remove marginal distribution skew (i.e., it's intended purpose).

A summary of the residual marginal distributions is shown in Figure 5-7. Across all analyses, for the most part, the Box-Cox transformation successfully removed marginal distribution skew (most residuals were either normally distributed or kurtotic). However, for some analyses (e.g. the Analysis 1 pre-instrumental models), many residuals were still skewed.

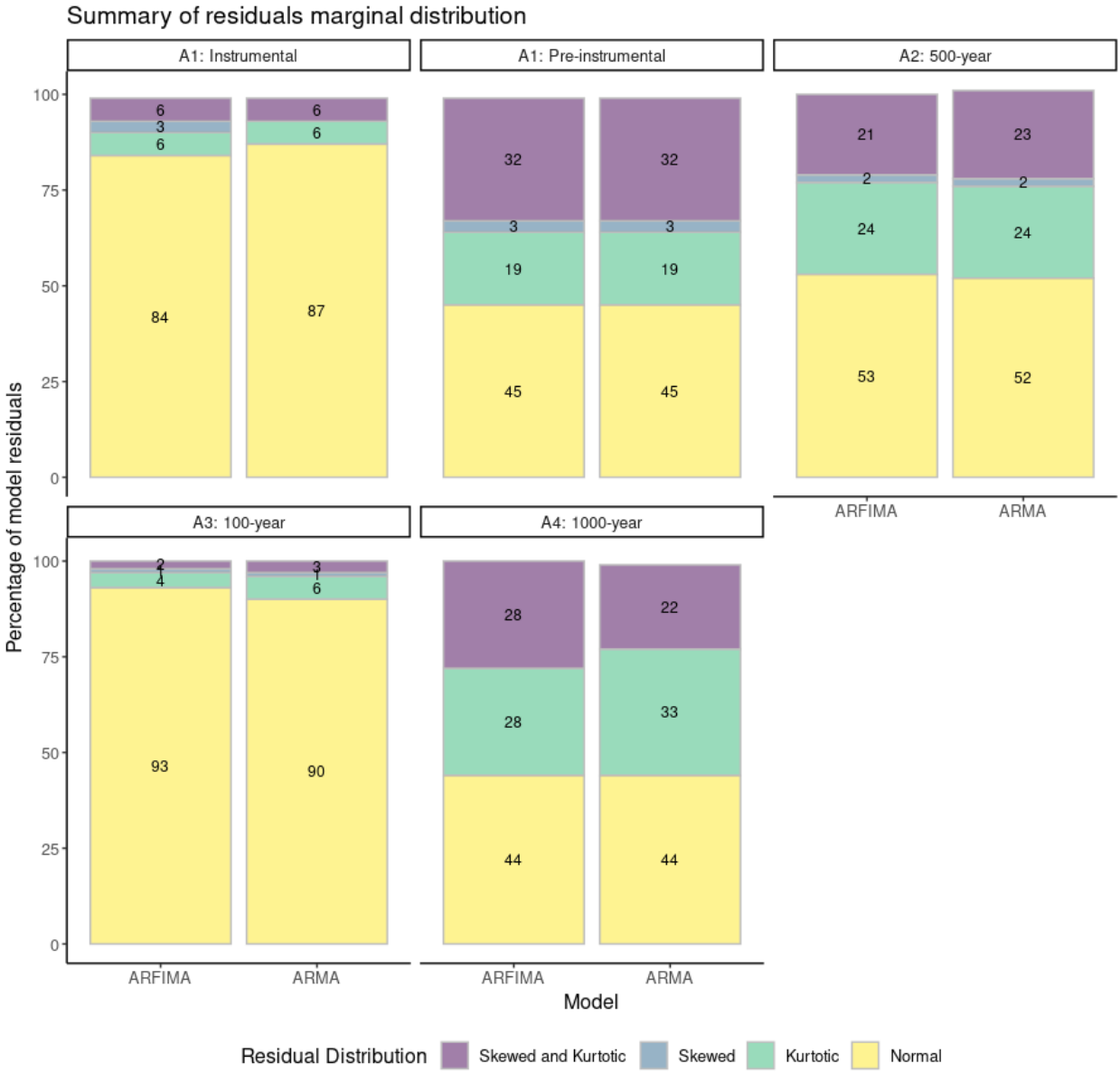

Figure 5-7: Summary of residual marginal distributions, categorised as either Normal, Kurtotic, Skewed, and Skewed and Kurtotic.

Even though residual assumptions were not consistently met, linear model parameters (e.g. ARMA and ARFIMA) are typically robust to non-normal residuals, especially for the large sample sizes in each proxy record (Knief and Forstmeier, 2021). Moreover, that the model residuals were typically consistent with white noise is reassuring (residuals with positive serial dependence is indicative of underestimated posterior variance).

5.7 Discussion

In water management, discussions on the stationarity assumption have been ongoing for decades (Klemeš, 1989). More recently, Milly et al. (2008) controversially declared that

‘stationarity is dead’. In this study, we wanted to explore if, for annual scale stochastic models, stationarity was ever really ‘alive’. The answer is dependent on the parameter, time horizon, and model. Of particular interest is that, across all analyses, the persistence parameters in both models were more likely to be stationary. In contrast, the mean, standard deviation, and Box-Cox parameters were more likely to be non-stationary. For some time horizons and parameters, there was strong evidence for non-stationarity (i.e. a p-value <0.01). However, for others, evidence of non-stationarity was marginal. So, is non-stationarity the norm for annual-scale stochastic models? Our study suggests ‘probably’.

Marginal evidence against stationarity raises questions as to if and how historic non-stationarity should be considered in long-term climate risk modelling. From our perspective, the key issue facing climate risk modelling is not removing (or hiding) all stationarity assumptions. Rather, when assuming a particular parameter or relationship is stationary, how wrong are we prepared to be? This is a somewhat subjective judgement, which depends on (a) the modelling task, (b) whether non-stationary models are demonstrably better than a stationary model, and (c) the consequences of wrongly assuming stationarity.

From a climate risk perspective, the consequences of wrongly assuming stationarity are, in our opinion, most important. This study highlights that, under historic climate variability, (a) stochastic model mean and standard deviation posteriors are similar at centennial timescales, but not multi-centennial and millennial timescales and (b) stochastic model persistence posteriors are similar across centennial, multi-centennial, and millennial timescales. This means that under historic climate variability and over a 100-year planning horizon, assuming stationarity can still produce reasonable climate risk estimates.

If only ~100 observations can be used to infer future mean and standard deviation over a 100-year planning horizon, the corresponding stochastic model parameter uncertainty will be large (Thyer et al., 2006). This uncertainty, which is unavoidable, should be considered to ensure water security under climate variability (Berghout et al., 2017). However, further research is needed to better understand the trade-offs between reducing parameter uncertainty and accurately estimating future risk. At what point exactly are previous observations, and the mean and standard deviation inferred from those observations, no longer representative of future observations?

In contrast to the mean and standard deviation, results indicate that persistence is likely stationary, irrespective of timescale. Further research is needed to explore the source of this stationarity. We recommend such research be conducted within the context of how proxies respond to and record various scales of hydroclimatic variability (as opposed to immediately proposing and evaluating physical climate mechanisms of stationary persistence).

Potentially stationary persistence also suggests that longer hydroclimatic timeseries, such as palaeoclimate proxy records, can be used to constrain persistence parameter uncertainty without introducing a parameter bias due to non-stationarity (provided the longer timeseries accurately records the 'true' persistence). Future work constraining persistence parameter uncertainty is a key recommendation from this study.

Although the results give marginal evidence against stationarity, stationarity remains an important statistical concept that is, perhaps, better served without reference to being 'dead' or 'alive'. Stationarity is neither alive nor dead. It is an ubiquitous assumption made whenever climate risk is modelled statistically.

To some degree, even supposedly non-stationary models, non-parametric models, and physically based models will, at some point, assume stationarity (Montanari and Koutsoyiannis, 2014). For non-stationary parametric models, even if a model parameter can vary in time, the way in which it varies is typically governed by a conditionally stationary statistical relationship (unless physical equations are used). For example, stochastic streamflow models can incorporate non-stationarity in the mean by adding temperature as a covariate (Kiem et al., 2021). However, the relationship between temperature and streamflow is conditionally stationary (any non-stationarity is due to non-stationarity in the covariate). Non-parametric models (e.g., k-Nearest Neighbour resampling - Lall and Sharma (1996)) also implicitly assume stationarity. This is because non-parametric models assume that the calibration data accurately describes unobserved data. If the unobserved data is different to the calibration data, then the implicit assumption of stationarity is violated. Many complex physically based models also assume stationarity. Such assumptions are made via parameterised sub-routines (Davini et al., 2017). For example, currently, global and regional climate models do not directly simulate cloud convection and rainfall (Evans et al., 2012; Huang et al., 2020). These processes are instead parameterised, with parameters held constant during simulation (Hong et al., 2006). Although these models are physically based (in the sense

that energy and mass is conserved between variables), there are some stationarity assumptions within!

Considerations of stationarity aside, this study also highlights some technical limitations of using a Box-Cox transformation prior to stochastic model calibration. For many long timeseries, model assumptions of normality were often violated (even with a Box-Cox transformation). To ensure residual assumptions are met, alternative probability distributions that account for skew and kurtosis could be considered (e.g Stedinger (1980)). However, because non-normality violations in this study have limited influence on the model parameters (Knief and Forstmeier, 2021), alternative distributions should be considered in light of why we use the normal distribution. More specifically, in stochastic modelling, using a normal distribution for annual data is often justified in terms of (a) the Central Limit Theorem (CLT) and (b) statistical convenience.

The CLT states that the sum of random variables converges to a normal distribution as the number of random variables being summed increases. Annual data used in this study can be viewed as the sum of sub-annual random events (e.g., snowfall or tree-ring growth induced by rain). Is the CLT valid for these processes? Issues of considering physical processes such as tree-growth and ice formation the outcome of a random process aside, there may not be enough sub-annual events for the annual total to be normally distributed. Furthermore, the CLT assumes that (a) the random variables are independent (or weakly dependent) and (b) the probability of observing different random variables is constant through time. In terms of hydroclimatic timeseries (e.g., rainfall, proxy records), neither assumption is likely. These timeseries typically exhibit long-term dependence and are influenced by various physical processes, such as the El Nino Southern Oscillation, which alter flood and drought likelihood (Kiem et al., 2003; Kiem and Franks, 2004). Furthermore, skewed probability distributions can be derived by applying the principle of maximum entropy to positively bounded random variables (Papalexiou and Koutsoyiannis, 2012). Along with the previously mentioned limitations of the CLT, the availability of theoretically rigorous alternative distributions suggests that, for annual scale stochastic modelling, the CLT is a poor justification for the use of a normal distribution.

This leaves statistical convenience as the primary justification for using a normal distribution. Unlike many other distributions, normally distributed variables remain normal when summed

or multiplied. For timeseries models, this makes it simple to calculate marginal and conditional distributions and likelihoods (Box et al., 1970). Therefore, if considering an alternative probability distribution, a trade-off must be considered between the convenience of the normal distribution and the added benefits of the alternative distribution.

5.8 Conclusion

In this study, we evaluated stochastic model parameter stationarity using 31 millennium-length, hydroclimatic proxy records. By examining millennium-length records, we could better account for parameter uncertainty when evaluating if these parameters are constant in time (i.e. are stationary) or if these parameters vary (i.e. are non-stationary). We found that:

- At multi-centennial and millennial timescales, marginal evidence that the mean and standard deviation parameters are non-stationary
- At centennial scales evidence that the mean and standard deviation parameters are stationary.
- At centennial, multi-centennial, and millennial timescales, evidence that stochastic model persistence is stationary.

5.9 Links with following chapters

Chapter 5 indicates that stochastic model persistence parameters are likely stationary, irrespective of timescale. In Chapter 6, this finding is used to inform the development of a palaeoclimate-informed stochastic modelling framework. The proposed framework uses ice core information to calibrate stochastic model persistence parameters for an annual scale stochastic rainfall model in mid-latitude Australia. More specifically, Chapter 5 is used to justify using an *entire* proxy record to define the prior distribution for stochastic model persistence parameters.

Chapter 6. Using ice core data in drought risk assessment and water resource management

6.1 Abstract

Palaeoclimate proxy data, such as ice core records, contain more severe droughts and pluvials than those in short instrumental rainfall and streamflow records. These proxy records can provide better drought risk estimates, which can better inform water management. In this study, we present a Bayesian modelling approach for using proxy data in water management. We use proxy data from an Antarctic ice core and instrumental measurements from southeast Australia to calibrate a catchment-scale stochastic rainfall model (i.e., a type of climate risk model used in water management). The proxy data is used to define a Bayesian prior for instrumental persistence. This extracts the proxy persistence signal, which is representative of broader regional persistence, without using the proxy to predict catchment-scale rainfall. When validated, the proposed model reproduces the observed drought risk. However, compared with the 'standard' model calibrated using a non-informative prior, our 'palaeoclimate-informed' model simulates much longer and more severe droughts/pluvials. From a water management perspective, these extended droughts mean that more water storage is required to meet demand. Furthermore, in comparison to existing methods of palaeoclimate-informed stochastic modelling, the proposed model also simulates more severe droughts. This study (a) highlights significant limitations with using instrumental records to characterise climate risk; and (b) presents a flexible framework that incorporates palaeoclimate persistence signals in catchment-scale drought risk assessment, enabling direct applications of palaeoclimate data in water resource management.

6.2 Introduction

In water management, stochastic models calibrated to instrumental rainfall and/or streamflow measurements (typically available from ~1900 onwards in Australia) are used to infer drought risk (Loucks and Van Beek, 2017). These stochastic models generate synthetic timeseries with similar statistics to the calibration data but with different, and potentially more severe, longer droughts (Fiering, 2013). Once generated, the synthetic timeseries are used as inputs into a water system model (Kuczera, 1992). This simulates water system behaviour under the various droughts generated by the stochastic model, which then informs water system operation, design, and adaptation (Vogel, 2017). This makes stochastic model calibration a crucial task for the design and operation of water supply systems. In this study, a novel calibration framework is presented that uses palaeoclimate information to inform and constrain the calibration of stochastic model persistence, which subsequently improves the drought risk estimates used in water management.

Stochastic models are used because they simulate two key features of hydroclimatic timeseries: randomness and persistence. Randomness means that the observed record represents just one plausible realisation of a hydroclimatic process (Deser et al., 2020; Koutsoyiannis, 2010; Sivakumar, 2000). Alternative, and more severe, realisations are possible. Hydroclimatic persistence is the tendency for wet or dry years to cluster in sequence (Graves et al., 2017; Hurst, 1951). This means that multi-year droughts are possible. Combined, randomness and persistence mean that droughts measured in a ~100-year rainfall or streamflow record are often not the worst droughts possible (Cook et al., 2022). By simulating randomness and persistence, stochastic models generate time series that contain droughts of greater severity/duration than those in the instrumental record (Matalas, 1967; Srikanthan and McMahon, 2001). This allows water managers to better characterise, and plan for, a wider range of plausible extreme events than that available in a single observational record.

Although stochastic models are useful to water managers, the use of relatively short calibration data (i.e., instrumental measurements) gives rise to two limitations. First, short records may not capture the full extent of multi-decadal climate variability (Hessl et al., 2018; Mundo et al., 2012; Vance et al., 2015; Verdon-Kidd et al., 2017). This means an 'instrumental-period' stochastic model will not reproduce realistic low-frequency climate variability and, by extension, underestimate long-term historic drought risk (Chapter 3). Second, short records,

such as annual-scale hydrologic records, are subject to considerable parameter uncertainty (Serinaldi and Kilsby, 2015). This means a large range of stochastic model parameters produce similarly good calibrations (Thyer et al., 2006). Large parameter uncertainty will propagate through water system models, resulting in highly uncertain estimates of sustainable yield (i.e., the amount of water a system can sustainably provide) (Berghout et al., 2017; Stedinger and Taylor, 1982a). Under this uncertainty it is difficult to identify optimal operational rules and/or system adaptations.

Potentially, palaeoclimate proxy records can be used to address these limitations. These records, which span hundreds of years, are measurements of climate-sensitive physical 'layers' (e.g., tree-rings and ice cores). Because they are much longer than instrumental measurements, proxy records can (a) reduce stochastic model parameter uncertainty (Patskoski and Sankarasubramanian, 2015) and (b) contain more severe droughts than those produced by a stochastic model calibrated to instrumental measurements. With respect to (b) this suggests that existing water supply systems, which were designed to mitigate instrumental-period drought risk, may not ensure water supply under historic climate variability (Gober et al., 2016). This means that to ensure future water supply, adaptation may be required (Armstrong et al., 2020; Cahill et al., 2023; Flack et al., 2020).

Given the advantages of proxy records, they should inform stochastic model calibration. In this study, a Bayesian method for incorporating palaeoclimate information in stochastic modelling is proposed. Proxy records are used to define Bayesian priors, which then inform the calibration of a stochastic rainfall model. These 'palaeoclimate-informed' stochastic models can better characterise long-term drought risk, leading to better-informed water management adaptation decisions.

Although potentially useful, proxy records have limitations that complicates their use in stochastic model calibration, drought risk assessment, and water management. These limitations are:

1. Most proxy-based climate reconstructions, such as those produced using linear regression, underestimate instrumental-period variance (Meko et al., 2022). By extension, these reconstructions also underestimate the magnitudes of instrumental-period extremes (Patskoski et al., 2015). These statistics are crucial for any drought risk estimate.

2. Limited in-situ (i.e., local) proxy data for catchments of interest (Galelli et al., 2021; Tingstad et al., 2014). This is a particularly prevalent issue across the mid-latitude Southern Hemisphere (Croke et al., 2021; Goodwin et al., 2022; O'Connor et al., 2022).

The proposed paleoclimate-informed stochastic modelling framework addresses these limitations. Limitation (1) is addressed by explicitly preserving instrumental-period variance. Limitation (2) is addressed by using information from remote hydroclimatic proxies. In this study remote ice core proxies are used, which are a major source of palaeoclimate information in the Southern Hemisphere.

Ice cores contain particularly useful information to water managers (Kiem et al., 2020). They are long and contain relatively unbiased signals of regional persistence (unlike some tree-ring records) (Chapter 2). These features are useful because regional persistence, which significantly influences water system design and behaviour (Vogel et al., 1999), is poorly characterised in instrumental measurements (Thyer et al., 2006). Various studies also demonstrate that there is limited spatial variability in regional-scale persistence (Fatichi et al., 2012; Tyralis et al., 2018, 2018), meaning that ice core persistence can be used as a proxy for mid-latitude persistence (Chapter 2). Assuming that regional persistence remains coherent over time, ice cores can be used to better infer catchment-scale persistence (irrespective of locally available proxies).

Regardless of proxy data used, any palaeoclimate-risk framework must preserve instrumental-period variances and extremes during calibration. Existing frameworks have done this by using proxy data to inform resampling of the instrumental record (Erkyihun et al., 2016; Gangopadhyay et al., 2009). Given that different palaeoclimate reconstructions of the same region typically agree on relative wet/dry state but disagree on the magnitudes of these wet/dry states, these approaches will use palaeoclimate reconstructions (or the proxy record directly) to infer wet/dry state (Prairie et al., 2008). Then, a corresponding wet or dry value is sampled from the instrumental record. Alternatively, values can be sampled from statistical models calibrated to wet or dry instrumental-periods (Henley et al., 2011).

Although these methods preserve variance, the sampled wet or dry values are limited to either (a) those contained in instrumental data; or (b) those derived from stochastic models calibrated to wet or dry instrumental periods. Regarding (a), this means extrapolation to larger,

unrecorded extremes is not possible (such extremes are possible and should be accounted for when quantifying drought risk). Regarding (b), given that these wet/dry periods are a subset of an already short instrumental record, parameter uncertainty will be substantial. As stated previously, large parameter uncertainty will propagate through a water system model, which makes it hard to identify optimal management rules and infrastructure (Berghout et al., 2017; Stedinger and Taylor, 1982a). Considering these limitations, a palaeoclimate-informed stochastic modelling approach that extrapolates to unobserved values while minimising parameter uncertainty is desirable.

Given (a) the need to consider palaeoclimate data when calculating drought risk; (b) that certain palaeoclimate proxies, such as ice cores, contain realistic and extended samples of regional hydroclimatic persistence; and (c) limitations with existing methods of palaeoclimate-informed stochastic modelling, in this study we propose a new stochastic modelling framework for using palaeoclimate data in drought risk assessment and water management.

In this framework, we initially calibrate a stochastic model to a proxy timeseries. From the proxy calibration, we then extract persistence parameter posteriors. Posterior uncertainty is small because the proxy record is long. These posteriors are then used to define a Bayesian prior distribution for the target catchment rainfall model. This prior is 'informative' - it limits rainfall persistence parameters to those consistent with proxy low-frequency climate variability. In contrast, catchment rainfall mean and standard deviation priors are non-informative. This approach leverages the robust characterisation of low-frequency climate variability contained in palaeoclimate records and explicitly preserves the target catchment coefficient of variation. This preservation removes issues of variance loss common in standard palaeoclimate reconstruction methods. However, unlike existing paleoclimate-informed stochastic models, the proposed framework does not subset mean/standard deviation parameters into wet/dry states. This will reduce parameter uncertainty.

6.3 General modelling framework

The proposed modelling framework uses a proxy record to inform catchment-scale stochastic model calibration. This requires understanding (a) how different stochastic model parameters are sensitive to different features of the calibration data marginal distribution (e.g., mean, variance, skew etc.) and persistence/autocorrelation structure; and (b) what information, if any,

available proxy records have that can better characterise these features/parameters. This involves:

1. Selection of an appropriate stochastic model (i.e., one capable of reproducing realistic climate variability across multi-decadal/centennial scales). We selected the ARMA(1,1) model.
2. Identifying which parameters are sensitive to what features of the calibration data:
    a. The mean ($\boldsymbol{\mu}$); residual variance ($\boldsymbol{\sigma}$); and Box-Cox exponent ($\boldsymbol{\lambda}$) parameters are sensitive to the sample mean, variance, and skew respectively (and any corresponding non-stationarities). For the ARMA(1,1) model, $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are sensitive to the persistence, or autocorrelation, structure of the calibration data ($\boldsymbol{\lambda}$ may also have a minor influence (Montanari et al., 1997)).
3. Identification of potential palaeoclimate proxy records for catchment and climate variable of interest.
    a. For this case study, the catchment of interest was the Williams River catchment (Figure 6-1). This is an important water supply catchment for Newcastle - a city in southeast Australia. Annual rainfall was the target climate variable.
    b. The summer sea salt record from Law Dome, East Antarctica, was identified as a suitable proxy (Figure 6-1). This proxy region is linked to southeast Australia by Southern Ocean synoptic systems. Variations in the north and south position of Southern Ocean synoptic systems are conducive to increased (decreased) sea salt deposition (rainfall) at Law Dome (Williams River). Both locations and variables are also influenced by the El Niño Southern Oscillation and the Interdecadal Pacific Oscillation, two leading drivers interannual and multidecadal global climate variability.
4. Assessing what proxy record features, if any, contain useful information about the target catchment marginal distribution or persistence structure (including potential non-stationarities in either).
    a. Ice core records, such as Law Dome, provide relatively unbiased estimates of regional (i.e., mid and high-latitude Southern Hemisphere) persistence. For mid-latitude catchments, this makes ice core proxies ideal for inferring stochastic model persistence.
    b. Due to the physical links between Law Dome and east Australia, this proxy does have some skill in predicting east Australian rainfall (statistically significant

correlation ~0.30). Such skill could also be accounted for within the general modelling framework. For example, the proxy could be used as a stochastic model covariate (which would incorporate any proxy non-stationarities in the marginal distribution). However, such considerations - which increase the complexity of the model - are left for future work.

5. Formulate Bayesian model whereby 'useful' proxy features inform Bayesian priors for the target catchment model.
    a. Given that (i) ice cores provide relatively unbiased estimates of regional hydroclimatic persistence; and (ii) there is common ENSO signal between Law Dome and the Williams River catchment, east Australia, we use ice core data to inform the calibration of an annual stochastic rainfall model. The proxy record is used to define the Bayesian priors for annual rainfall persistence. This involves inferring proxy record persistence posteriors. This posterior is then used to select hyperparameters of the rainfall persistence prior. Because the ice core record is long, uncertainty in rainfall persistence parameters will also be reduced considerably.

Although this study uses a remote ice core record to calibrate stochastic model persistence parameters, the general concept can be extended to other proxies/parameters. However, we focussed on stochastic model persistence because (a) hydroclimatic persistence has a significant influence on the design and operation of water supply systems, (b) there is considerable uncertainty in persistence statistics and parameters inferred from short instrumental records (Chapter 4), and (c) stochastic model persistence parameters are dimensionless (meaning that proxy persistence can be transferred to rainfall persistence without a unit conversion model).

There are two key assumptions of the proposed method. First, that the persistence parameters are stationary. Chapter 5 examined this assumption using a global dataset of extended proxy records, finding that persistence parameters were typically stationary at millennial and centennial scales. Second, that the proxy persistence signal is representative of broader regional persistence. We address this by using Antarctic ice core records. These records contain relatively unbiased persistence signals. For some sea salt records, there is a tendency to overestimate persistence.

For other proxy types, additional care must be taken when assuming persistence signals are unbiased. For example, if used as a proxy for rainfall, tree-ring records typically overestimate low-frequency climate variability (Ludescher et al., 2020; Yuan et al., 2021). Statistical methods that address this bias (such as pre-whitening) will remove the low-frequency persistence signal from the proxy timeseries (Razavi and Vogel, 2018). These biases will propagate through the proposed model and give biased drought risk estimates.

6.4 The ARMA(1,1) Model

In this study, we demonstrate the Bayesian framework using the Autoregressive Moving Average (1,1) (ARMA(1,1)) model (Box et al., 1970). This model was selected because it is parsimonious and able to simulate a wide variety of different climate timeseries (Boes and Salas, 1978). Furthermore, Chapter 3 demonstrated that this model can reproduce low-frequency climate variability when calibrated to extended proxy records.

The ARMA(1,1) model represents a timeseries y of length t as weighted sum of the prior observation (y_{t-1}) and prior residual error (ϵ_{t-1}) plus an independent, normally distributed residual (ϵ_t). The Φ and θ parameters are the respective weights of the prior observation and prior error, meaning that these parameters are sensitive to timeseries persistence.

$$y_t = \mu + \phi(y_{t-1} - \mu) + \theta\epsilon_{t-1} + \epsilon_t \qquad \text{Equation 6-1}$$

Where:

$$\epsilon_t \sim N(0, \sigma) \qquad \text{Equation 6-2}$$

The ARMA(1,1) model assumes that the data follows a normal distribution. However, hydroclimate timeseries are typically skewed. To ensure normality, a Box-Cox transformation was used prior to model calibration.

$$\begin{aligned} if\ \lambda \neq 0: z_t &= \frac{y^{\lambda}{}_t - 1}{\lambda} \\ if\ \lambda = 0: z_t &= log(y_t) \end{aligned} \qquad \text{Equation 6-3}$$

6.5 Bayesian modelling framework

The proposed model uses proxy records to inform rainfall persistence calibration. This involves calibrating a stochastic model to the proxy. The proxy persistence posterior is then used to select hyperparameters of the rainfall persistence prior. Although we present this framework using the ARMA(1,1) model, the same concepts can be applied to different models.

Prior to calibration, both proxy and rainfall timeseries are transformed via a Box-Cox transformation. The transformed timeseries is represented as an ARMA(1,1) process, giving:

$$z_{proxy,t} = \mu_{proxy} + \phi\left(z_{proxy,t-1} - z_{proxy}\right) + \theta\epsilon_{proxy,t-1} + \epsilon_{proxy,t} \qquad \text{Equation 6-4}$$

$$\epsilon_{proxy,t} \sim TN(0, \sigma_{proxy}, lb, ub) \qquad \text{Equation 6-5}$$

$$z_{rain,t} = \mu_{rain} + \phi(z_{rain,t-1} - \mu_{rain}) + \theta\epsilon_{rain,t-1} + \epsilon_{rain,t} \qquad \text{Equation 6-6}$$

$$\epsilon_{rain,t} \sim TN(0, \sigma_{rain}, lb, ub) \qquad \text{Equation 6-7}$$

Where the lower and upper bounds of the Truncated Normal distribution are based on the Box-Cox transformation.

$$\begin{aligned} if\ \lambda > 0: lb &= \frac{-1}{\lambda}\ ;\ ub = \infty \\ if\ \lambda < 0: lb &= -\infty\ ;\ ub = \frac{-1}{\lambda} \end{aligned} \qquad \text{Equation 6-8}$$

These lower/upper bounds are set to satisfy the positivity constraint introduced by the Box-Cox transformation

$$z_t\lambda + 1 > 0 \qquad \text{Equation 6-9}$$

The proxy record posteriors are then inferred using the likelihood function in the Appendix. Note that this likelihood function infers the posteriors of the mean and standard deviation of the untransformed timeseries *y* (based on first order approximations of the mean and standard deviation of the transformed timeseries *z*).

Following Frost *et al.*, 2007, non-informative priors were used when calibrating the proxy ARMA(1,1) model.

$$\begin{gathered}\phi_{proxy}, \theta_{proxy} \sim Uniform(-1, 1) \\ \lambda_{proxy} \sim Uniform(-2, 2) \\ \mu_{proxy} \sim Normal(\overline{\mu}_{proxy}, \overline{\sigma}^2_{proxy}) \\ \sigma_{proxy} \sim Inverse\ Gamma(1, \overline{\sigma}_{proxy})\end{gathered} \qquad \text{Equation 6-10}$$

From the proxy Φ and θ posteriors, the mean and covariance matrix are then used as hyperparameters to the rainfall Φ and θ priors. Because Φ and θ parameters are typically correlated, priors are modelled as a truncated multivariate normal distribution.

$$\phi_{rain}, \theta_{rain} \sim MVTN(\mu_{proxy\ \phi,\theta}, \Sigma_{proxy\ \phi,\theta}) \qquad \text{Equation 6-11}$$

Both Φ and θ have lower and upper bounds of -1 and 1. For the other rainfall parameters, non-informative priors are used:

$$\begin{gathered}\lambda_{rain} \sim Uniform(-2, 2) \\ \mu_{rain} \sim Normal(\overline{\mu}_{rain}, \overline{\sigma}^2_{rain}) \\ \sigma_{rain} \sim Inverse\ Gamma(1, \overline{\sigma}_{rain})\end{gathered} \qquad \text{Equation 6-12}$$

6.5.1.1 Inference of model posteriors

Using the derived likelihood function, Bayesian inference is possible. Bayesian inference is useful because it can quantify stochastic model parameter uncertainty (i.e., parameter posteriors). This uncertainty can significantly influence drought risk estimates and water system performance.

Parameter posteriors were inferred using the No U-Turn Sampling (NUTS) algorithm (Homan and Gelman, 2014). This is an efficient implementation of the Hamiltonian Monte Carlo (HMC) algorithm. HMC algorithms automatically adapt to the geometry of the target posterior distribution being sampled, making them suitable for inferring a variety of complex posteriors (Betancourt, 2018). Posteriors are sampled simultaneously from independent 'chains'. In this study, we generated 2500 samples from 8 chains (each with a prior burn-in of 500 samples).

Every 20th sample from each chain was then extracted and combined to produce a final posterior with 1000 samples. Chain 'mixing' (i.e., the tendency for chains to sample from the same posterior space) was achieved, with an Rhat value of 1 (Brooks and Gelman, 1998). Algorithms were implemented using the Stan programming language (Carpenter et al., 2017).

6.6 Study site and data

The proposed modelling framework was demonstrated using a case study in the Williams River catchment, located in coastal southeast Australia (Figure 6-1). This catchment provides inflow to Grahamstown Dam, an important water source for the city of Newcastle (providing ~40% of the water supply for a population of ~600,000).

Annual rainfall for the Williams River catchment was taken from the Australian Water Availability Project (AWAP). AWAP provides gridded daily rainfall data (on an ~5km x 5km grid) Australia wide, interpolated from a network of quality controlled rain gauges (Raupach et al., 2009). Grids within the catchment boundary were extracted, averaged, then aggregated to annual rainfall totals to produce the final rainfall timeseries used in stochastic modelling.

The recently extended Law Dome Summer Sea Salt proxy record of Jong *et al.*, 2022 was used in this case study. This proxy has been used to reconstruct southeast Australian hydroclimate (Tozer et al., 2016); ENSO (Vance et al., 2013); and the IPO (Vance et al., 2022). The Law Dome site is linked to east Australia by Southern Ocean synoptic systems, which are sufficiently large to influence Australian/Antarctic climate (Udy et al., 2021). More specifically, the meridional position of Southern Ocean high pressure systems are conducive to coincident increased/reduced sea salt concentration in Law Dome snowfall and east Australian rainfall (Udy et al., 2022). Consistent with the Law Dome based IPO reconstruction of Vance *et al.*, 2022, proxy measurements from CE 1-2011 were used in this study.

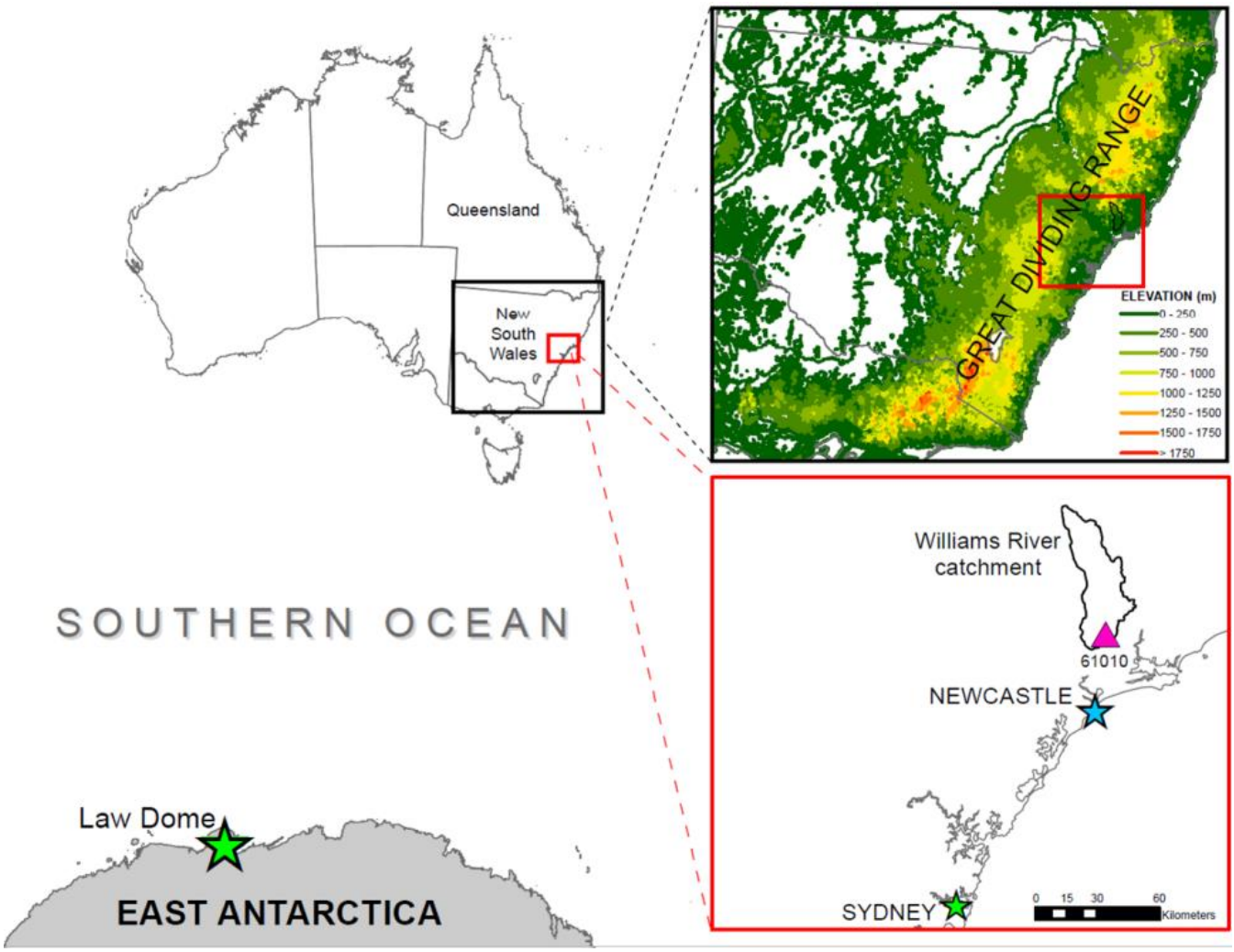

Figure 6-1: From Tozer et al., 2016. Site map of Williams River Catchment (located in eastern Australia) and Law Dome (located in east Antarctica).

Although we have only selected one study site/proxy record, the proposed framework is flexible and can be used for any target catchment/proxy, provided the proxy contains reliable information about regional low-frequency climate variability.

6.7 Methods

After calibration, the proposed model was validated on the proxy/rainfall timeseries (described in Section 5.1), then compared against other stochastic models (described in Section 5.2).

6.7.1 Model validation

The proposed model simultaneously calibrates two stochastic models - one for the proxy timeseries, another for rainfall. For both proxy/rainfall models, posteriors contained 1,000 parameter samples. For each sample, a single synthetic timeseries of equal length to the rainfall

or proxy timeseries was generated. Statistics for each synthetic timeseries were calculated. The corresponding proxy/rainfall statistic was then compared against the stochastic sampling distribution. The statistics evaluated were:

- Mean
- Lag-1 Autocorrelation
- Standard Deviation
- Hurst Coefficient, calculated using the Whittle estimator.
- Skew
- Minimum
- Maximum
- Minimum and maximum cumulative sums for overlapping 2, 5, 10, 30, 50, and 100-year windows.
    - Due to the limited rainfall record length, minimum/maximum 50 and 100-year cumulative sums were only validated for the Law Dome model.

We emphasise that validating both rainfall and proxy models is important. The purpose of rainfall model validation is to demonstrate that proxy-based priors can still generate realistic rainfall statistics. The purpose of proxy model validation is to demonstrate that the calibrated model can reproduce long-term climate variability that is not present in instrumental rainfall records.

A stringent model validation also involves checking if residual assumptions are met. These assumptions are that the model residuals have mean zero, are normally distributed, and are independent (i.e., no serial dependence). Residual diagnostics were calculated following Chapter 4. This involved calculating the model residuals from each posterior parameter set. These residuals were then used to evaluate model assumptions:

- Residuals having a mean of zero.

This was assessed by calculating the 90% credible interval of the residual mean posterior. Residuals were considered acceptable if the 90% credible interval contained zero.

- Residuals are normally distributed.

Normality was assessed via two checks. First, we used a Quantile-Quantile plot (QQplot) comparing the posterior of standardised residuals against theoretical Normal quantiles. A 90% credible interval was calculated for each standardised residuals. Residuals were considered

normal if 90% of the theoretical quantiles fell within the corresponding credible interval or if a Shapiro-Wilks test using the median residual distribution returned a p-vale greater than 0.1).

- Residuals are serially independent.

To evaluate independence, two different checks were performed. First, we evaluated the significance of residual Lag-1 autocorrelations (we consider this a proxy for dependence at greater time lags). Lag-1 autocorrelations were computed for each residual set. Residuals were considered independent if (a) the corresponding 90% credible interval contained zero; or (b) if the median Lag-1 autocorrelation returned a p-value >= 0.1 for the corresponding Pearson correlation test statistic. Second, the residual cumulative periodogram was calculated. For white noise, the cumulative periodogram should increase by 0.5 times the standardised frequency. Residuals were considered white noise if (a) the corresponding 90% credible interval for each cumulative frequency contained the theoretical white noise line 90% of the time or (b) the median cumulative periodogram returned an insignificant Kolmogorov-Smirnov p-value when compared with the theoretical white noise line.

6.7.2 Comparison with existing stochastic models

After model validation, we then compared our proposed model with (a) the same model calibrated using only rainfall data (i.e., the standard calibration approach); and (b) two other methods of palaeoclimate-informed stochastic modelling. These methods of palaeoclimate-informed stochastic modelling were the Climate informed multi-time scale stochastic (CIMSS) framework of Henley *et al.*, 2011 (explained in Section 5.2.1) and the proxy-based non-parametric K-Nearest Neighbour (proxy-KNN) resampling of Gangopadhyay *et al.*, 2009 (explained in Section 5.2.2). Both methods used palaeoclimate data to infer hydroclimatic state, then derive corresponding state data using instrumental measurements.

Respective models were calibrated and used to generate 1,000 synthetic rainfall timeseries. For each model, timeseries 2,011 years long were generated (the same length as the corresponding Law Dome record). For each timeseries, minimum/maximum cumulative sums for overlapping 2, 5, 10, 30, 50, and 100-year windows were calculated. These sampling distributions from each model were then compared.

6.7.2.1 Climate informed multi-time scale stochastic framework

The CIMSS framework uses a two-level hierarchical model to first sample the length of wet and dry periods using palaeoclimate information, then generate wet and dry data using a stochastic model calibrated to instrumental measurements.

In the original CIMSS framework, the length of wet/dry periods were inferred based on various reconstructions of the Interdecadal Pacific Oscillation (IPO). The IPO is a leading driver of global multidecadal variability - different phases of the IPO (i.e., positive, negative, and neutral) are associated with increased/reduced rainfall. For example, in eastern Australia, a negative IPO is associated with increased rainfall and flooding, whereas a negative/neutral IPO is associated with reduced rainfall and drought. In order to generate annual rainfall totals during different phases of the IPO, an AR(1) stochastic model is calibrated to respective IPO phases following the method of Frost et al., 2007.

The original model combined various reconstructions of the Interdecadal Pacific Oscillation (IPO) into a single composite index. The length of consecutive periods above/below the composite index median (i.e., the 'run lengths') is then used to calibrate a Gamma distribution. In subsequent stochastic simulations, this distribution is used to generate IPO phase run lengths.

Instead of using a composite index, in this study we used the recent IPO reconstruction of Vance et al., 2022 to infer wet and dry run lengths. This reconstruction was selected over other IPO reconstructions (and the composite approach used in the original CIMSS framework) because it is based upon the same Law Dome Summer Sea salt proxy used in our proposed model. This removed proxy record selection as a potential confounding factor when comparing subsequent climate risk estimates. Furthermore, this reconstruction is also longer than previous IPO reconstructions (spanning ~2,000 years), with the early reconstructed period showing extended IPO neutral/positive states. Such states are associated with reduced rainfall across eastern Australia.

The tendency for longer IPO neutral/positive run-lengths in the Vance *et al.* (2022) reconstruction required some updates of the original CIMSS framework, which used the same Gamma distribution to generate run-lengths for different IPO phases. In this study, separate Gamma distributions were calibrated to IPO negative (i.e., IPO < -0.5 and IPO neutral-positive phases (i.e., IPO > -0.5).

For further details on the CIMSS likelihood function, refer to the Appendix.

6.7.2.2 Proxy-based k-Nearest Neighbour resampling

Instead of using proxy records to infer the state of a large-scale climate driver, then sample from instrumental measurements (i.e., the CIMSS framework); the proxy-based k-Nearest Neighbour (kNN) method of Gangopadhyay *et al.*, 2009 uses the proxy record directly to infer catchment wet/dry state. Based on the kNN resampling method of Lall and Sharma (1996), for each proxy measurement, this approach (a) identifies similar proxy measurements in the overlapping proxy/instrumental period; then (b) randomly samples one of these similar proxy measurements; and (c) samples the instrumental measurement corresponding to this proxy measurement. Assuming there is a significant climate/proxy relationship, this approach preserves the relative length of proxy wet/dry periods and, by resampling instrumental measurements directly, preserves climate variance.

The original method proposed by Gangopadhyay et al. (2009) was developed using multiple tree-ring chronologies - these chronologies differed in length and had data missing for different years. Accounting for these features required additional methodological steps not required for this single proxy analysis (e.g., subsetting of chronologies, dimension reduction via Principal Components Analysis). Therefore, we implemented the following, simplified method:

- For each proxy measurement, identify the K closest instrumental-period proxy measurements (using Euclidean distance as a measure of 'closeness'). Following Gangopadhyay et al. (2009), K was set to $\sqrt{N}$, with N being the length of the instrumental record.
- Assign sampling weights to each of the K nearest observations. Sampling weights were assigned following Lall and Sharma (1996). These weights increase the likelihood of resampling the closest observations (out of the K selected).
- Using these sampling weights, for each proxy measurement randomly sample an 'instrumental-period' proxy measurement.
- Sample the climate variable corresponding to the instrumental-period proxy.
- Repeat until 1,000 timeseries replicates are generated.

6.7.3 Comparing required storages for hypothetical reservoirs

6.7.3.1 Estimating flow and required storage

After comparing different paleo-stochastic models, we further compared the standard and proxy-informed ARMA model by (1) estimating annual streamflow using rainfall outputs and the Budyko water balance model then, using these flow outputs, (2) designing hypothetical reservoirs that will always meet a specific demand (i.e., will never fail). The storage size of this 'no-fail' reservoir was estimated using the Sequent Peak Algorithm (SPA). Reservoir size distributions from each model were then compared.

The Budyko model is a parameter-free model that can estimate runoff as a function of rainfall and potential evapotranspiration (PET). The model is skilful at predicting annual runoff in temperate climates (e.g., the Williams River catchment). Following Tozer et al. (2018), PET in each year was assumed to be the annual mean from the instrumental record. Although PET variability is an important influence on the annual water balance, this approach is justified because (a) in the Williams River, annual rainfall variability is much greater than annual PET variability (meaning rainfall variability is a much stronger driver of runoff variability); and (b) there are very few palaeoclimate PET reconstructions to use/model as an input. As such, for annual rainfall P and catchment average PET, annual runoff Q was estimated as:

$$Q = P - P * \sqrt{\frac{PET}{P}\,(1 - exp(-\frac{PET}{P})\,tanh(\frac{P}{PET})} \qquad \text{Equation 6-13}$$

Once the runoff timeseries were generated, hypothetical reservoirs were designed using the SPA. This is a preliminary screening method that can approximate 'no-fail' reservoir size for given inflow/demand sequences. For this analysis, demand was constant each year and set as a fixed proportion of the instrumental-period mean annual flow (MAF). The instrumental-period MAF was calculated from streamflow outputs produced using the Williams River AWAP rainfall timeseries described in Section 4. Five different demand scenarios were used, 0.1, 0.3, 0.5, 0.7, and 0.9 x MAF. These demand scenarios are consistent with a variety of different water supply systems (McMahon et al., 2007b).

For a given sequence, the SPA estimates 'no-fail' reservoir size by calculating the cumulative difference between inflow and demand. The highest cumulative difference is the storage

required to meet demand. A visual representation of the SPA is Figure 6-2.For inflow sequences of length N, the algorithm can be described as:

$$S = Maximum(S_t)\ for\ t = 1, \ldots, KN \qquad \text{Equation 6-14}$$

Where

$$S_t = S_{t-1} + Demand_t - Inflow_t\ ;\ if\ positive \qquad \text{Equation 6-15}$$
$$S_t = 0;\ if\ <= 0$$

And

$$Inflow_{t+KN} = Inflow_t\ ;\ Demand_{t+KN} = Demand_t \qquad \text{Equation 6-16}$$

In this formulation, the inflow/demand sequences are repeated (i.e., K = 2). This accounts for low inflow periods that may occur at the end of the sequence.

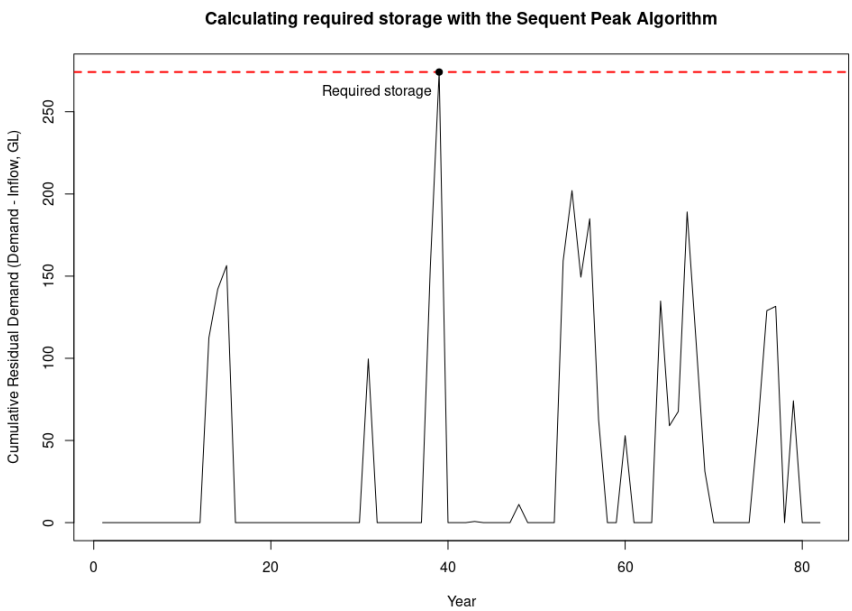

Figure 6-2: Example of the Sequent Peak Algorithm. The required storage is the maximum cumulative difference between demand and inflow (red dashed line).

6.7.3.2 Comparing storage estimates from unconditional and conditional rainfall simulations

When estimating required storage, typical infrastructure planning horizons are ~50 years. Therefore, when estimating required storages, we generated 50-year stochastic replicates from

the standard and proxy-informed models. We also compared the impact of conditioning the stochastic replicate on the most recent calibration period observation versus an 'unconditioned' replicate, which was initialised by randomly sampling from the calibration period marginal distribution. Note that the conditional simulation is essentially a 50-year ARMA(1,1) forecast.

The main reason for comparing conditional and unconditional simulations is that for timeseries exhibiting strong serial dependence (e.g. centennial-scale variability), the conditional variance over a 50-year lead time can be much lower than the unconditional variance. Figure 6-3 highlights how, for a theoretical ARMA(1,1) process with an unconditional variance of 1, the conditional variance for an ARMA(1,1) forecast will depend on the Φ and θ parameters. Crucially, for persistence parameters commensurate with centennial-scale variability (Φ = 0.98 and θ = -0.95, Chapter 3), it takes a forecast lead time of ~100-years before the conditional variance is equal to the unconditional variance. The difference between conditional and unconditional simulations may be important when estimating required storages over a 50-year planning horizon.

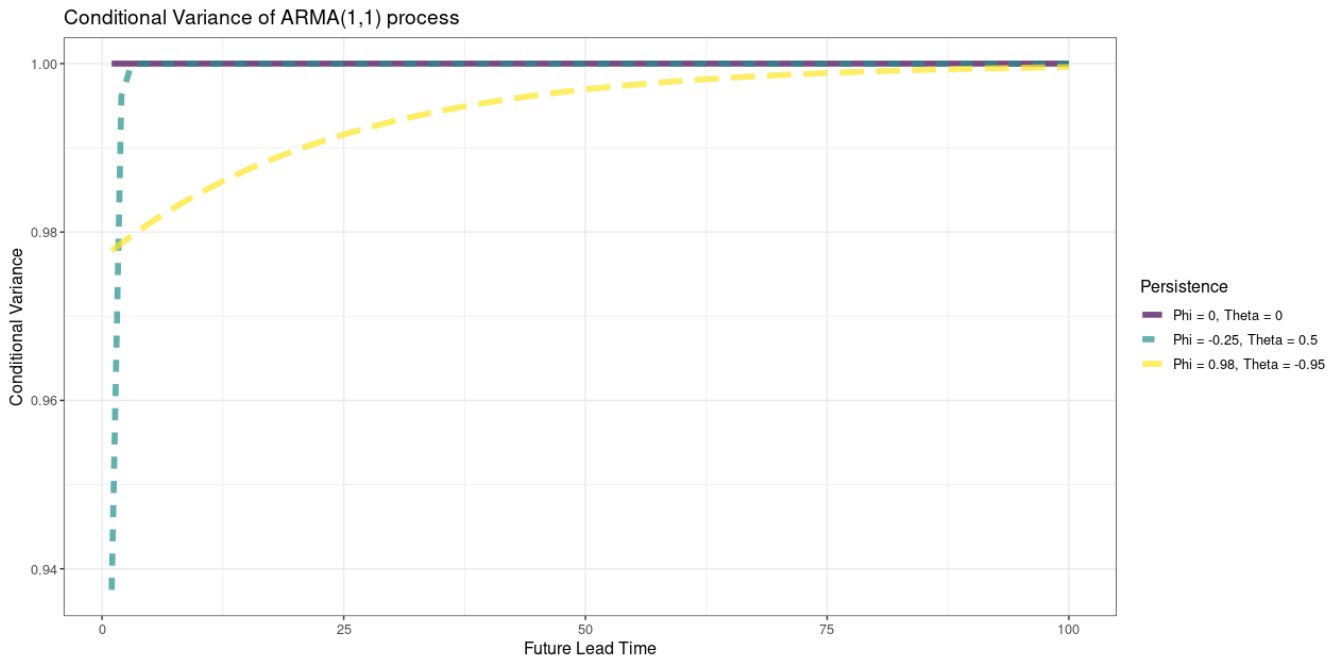

Figure 6-3: Conditional variance of an ARMA(1,1) process for different forecast lead times. All models have an unconditional variance of 1. Note that Phi = 0, Theta = 0 model is equivalent to white noise.

6.8 Results

6.8.1 Bayesian Hierarchical Model calibration

Figure 6-4 shows ARMA(1,1) posteriors for the 'standard' calibration (i.e. calibrated using only instrumental data) and the 'proxy-informed' calibration (i.e. the Empirical Bayes framework). We can see:

- A considerable reduction in persistence parameter uncertainty (i.e. Φ, θ and Φ + θ) for the proxy-informed model.
- For the standard model, the Φ + θ posterior contained zero (i.e. the instrumental record is consistent with white noise). For the proxy-informed model, the Φ + θ posterior was greater than zero, indicating positive persistence.
- However, the proxy-informed model had a wider Mean posterior than the standard model.

For the mean parameter, the increase in posterior uncertainty can be explained by the proxy-informed model consistently sampling from regions of higher persistence (Figure 6-5). Higher persistence reduces the number of independent observations, which in turn reduces the effective sample size (Hu et al., 2017; Macias-Fauria et al., 2012). Reduced effective sample size due to higher persistence in turn leads to greater uncertainty in the mean (Koutsoyiannis and Montanari, 2007).

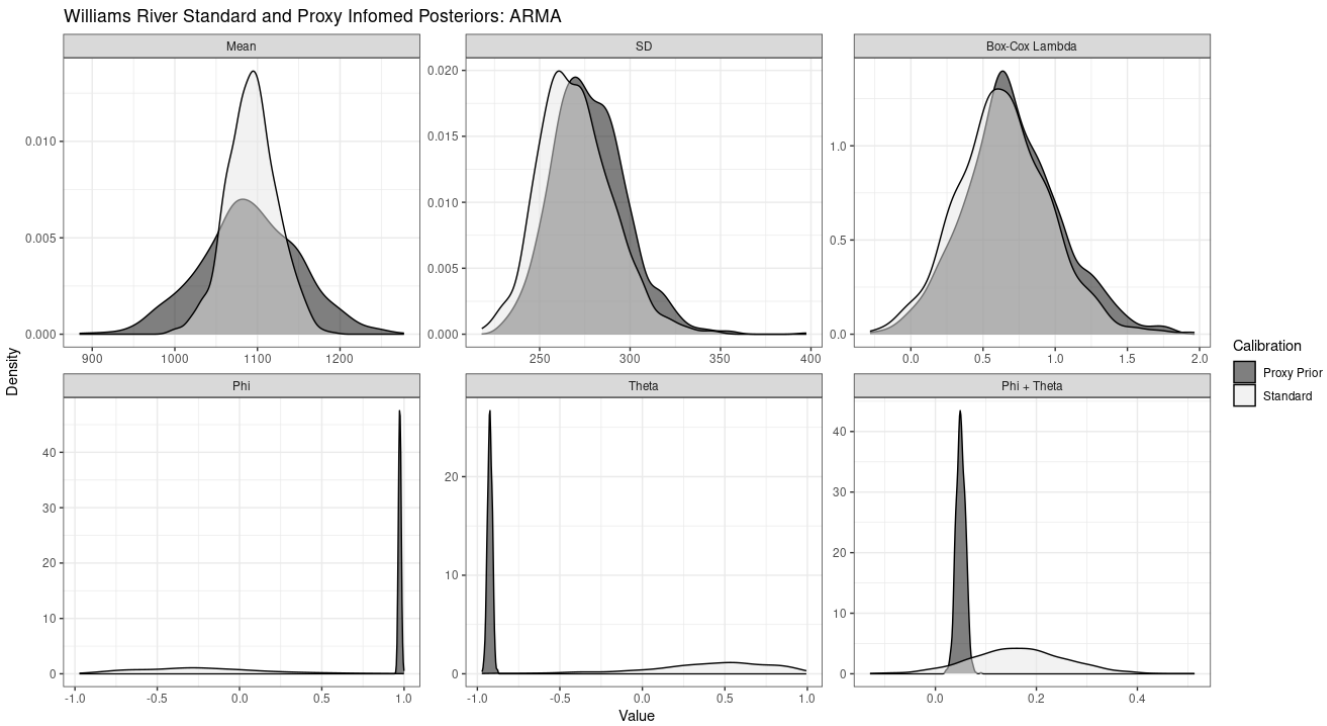

Figure 6-4: Comparison of 'Standard' ARMA(1,1) posteriors and 'Proxy-Prior' ARMA(1,1) model posteriors. 'Standard' refers to a model calibrated using only instrumental data. 'Proxy-Prior' refers to a model calibrated where Phi and Theta prior distributions were defined based on proxy information.

The higher persistence in the proxy-informed model is shown in Figure 6-5. From the top panel, we can see that the proxy-informed Autocorrelation Function (ACF) decays slowly. This slow decay is typical of timeseries exhibiting long-term persistence (i.e. low-frequency climate variability) (Dimitriadis and Koutsoyiannis, 2015; Koutsoyiannis, 2002; Tyralis and Koutsoyiannis, 2011). In contrast, the standard model ACF quickly reduces to zero, which is indicative of short-term persistence.

From the bottom panel of Figure 6-5, we can see large differences between the standard and proxy-informed Φ and θ joint posterior. That implies an inconsistency between the instrumental and proxy persistence signal. However, Chapter 4 highlighted that when calibrating an ARMA(1,1) model to a ~100-year record, MCMC samplers will not explore posterior regions associated with centennial-scale variability. This highlights a key strength of the proposed method – using long proxy records provides sufficient information for the MCMC sampler to explore posterior regions associated with centennial-scale variability.

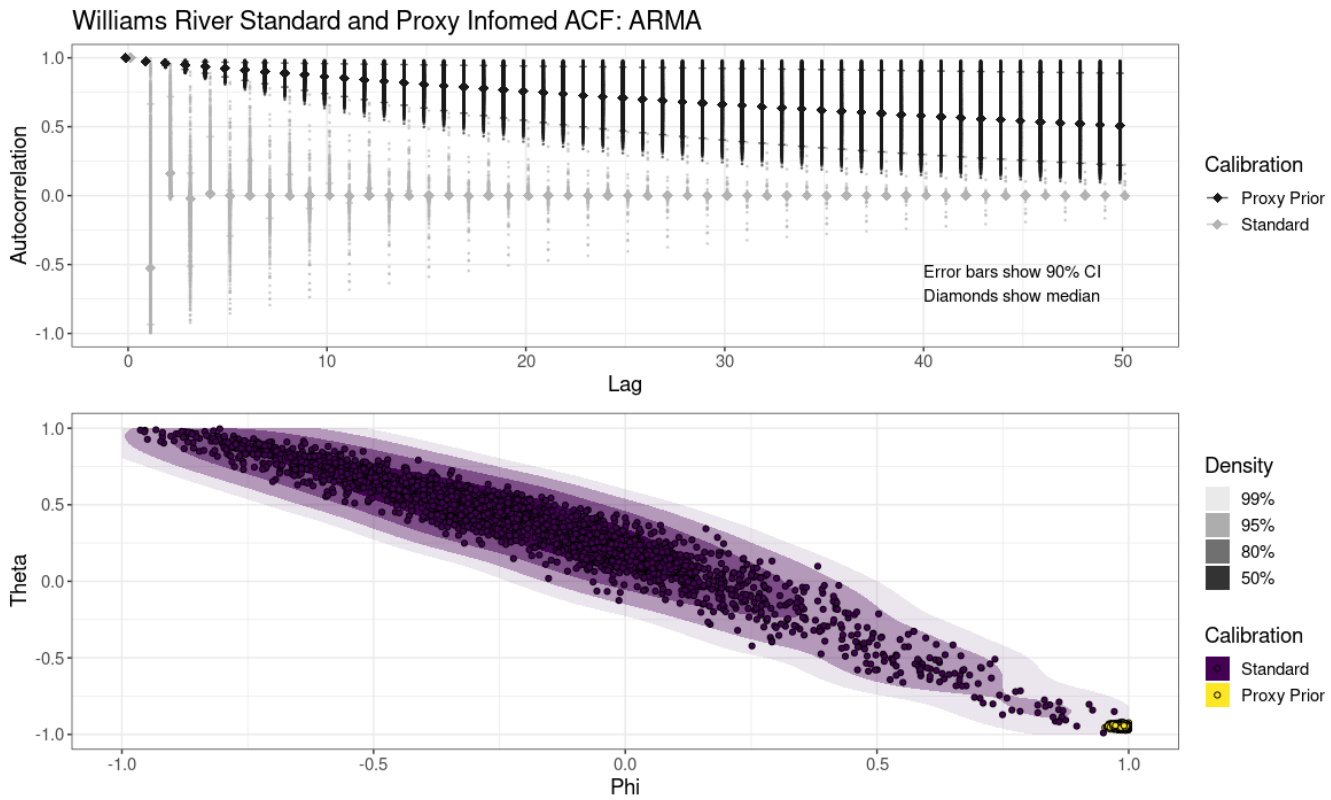

Figure 6-5: Top – comparison of theoretical Autocorrelation functions for the 'standard' and 'proxy prior' Williams River rainfall models. Bottom – comparison of bivariate Phi and Theta posteriors.

Residual diagnostics for the Law Dome and proxy-informed Williams River models are shown in Figure 6-6. Both models had posterior Lag-1 autocorrelations that did not contain zero, however, the median of respective posteriors was not significant when evaluated using a Pearson correlation. The Williams River residuals were approximately normal, whereas the Law Dome residuals were slightly kurtotic (note that slight kurtosis will not bias model parameters - Knief and Forstmeier, 2021).

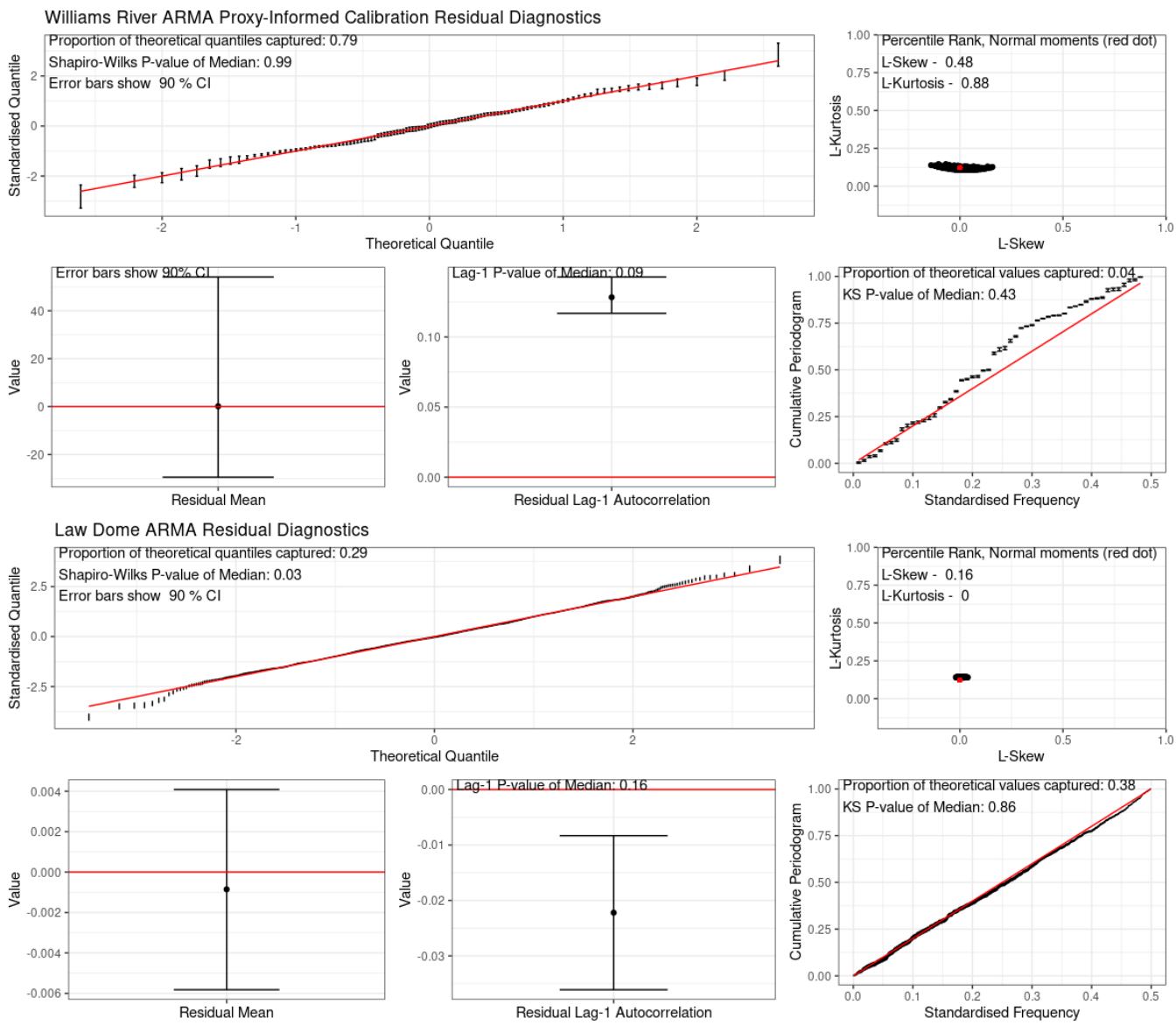

Figure 6-6: Residual diagnostics of the ARMA(1,1) model calibrated to (top): Williams River rainfall with proxy-informed priors; (bottom): the Law Dome summer sea salt record.

6.8.2 Bayesian model validation

Figure 6-7 shows validation results for the proxy-informed Williams River ARMA(1,1) and the Law Dome models. Figure 6-7 shows that both models were able to reproduce key statistics.

For the Williams River model, the key statistics were not biased by proxy-informed priors. For the Law Dome model, the key statistics related to low-frequency variability, the 100-year minimum and maximum, were captured. Capturing the 100-year minimum and maximum increases confidence that the Williams River model will simulate realistic low-frequency variability.

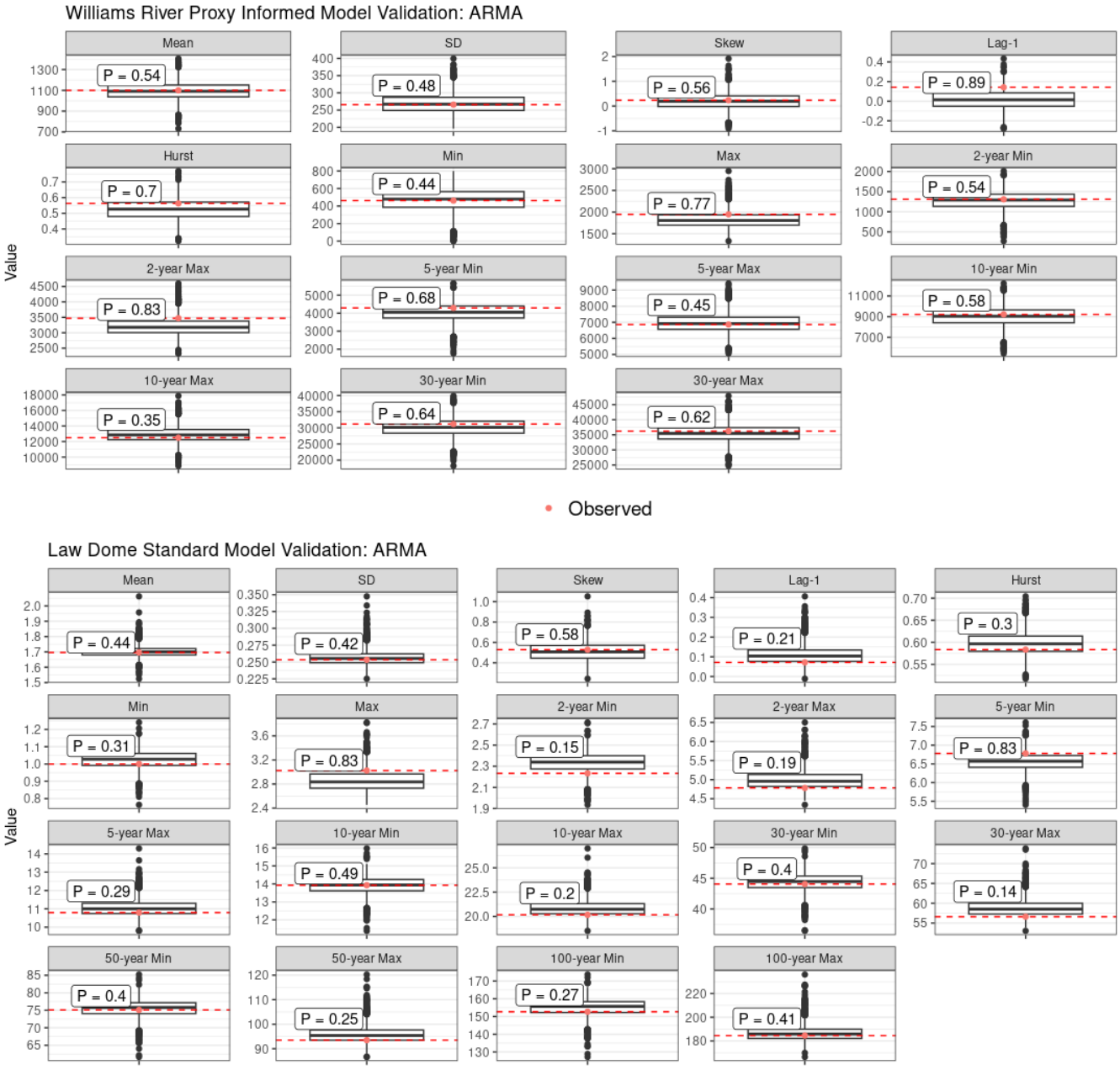

Figure 6-7: ARMA(1,1) validation results for (top): Williams River rainfall with proxy-informed priors; (bottom): the Law Dome summer sea salt record.

6.8.3 CIMSS calibration

CIMSS validation results are shown in Appendix 2 and indicate that (a) IPO run-lengths are consistent with a Gamma distribution and (b) the calibrated model can reproduce key hydrological statistics in the Williams River catchment.

CIMSS model posteriors are displayed in Figure 6-8. We can see that IPO positive/neutral phases had a much longer and more variable run-lengths than IPO negative. We can also see that IPO positive/neutral phases had lower and less variable rainfall that IPO positive phases (consistent with Verdon et al., 2004). However, compared with the parameter uncertainty, these differences in rainfall mean and variability are relatively small.

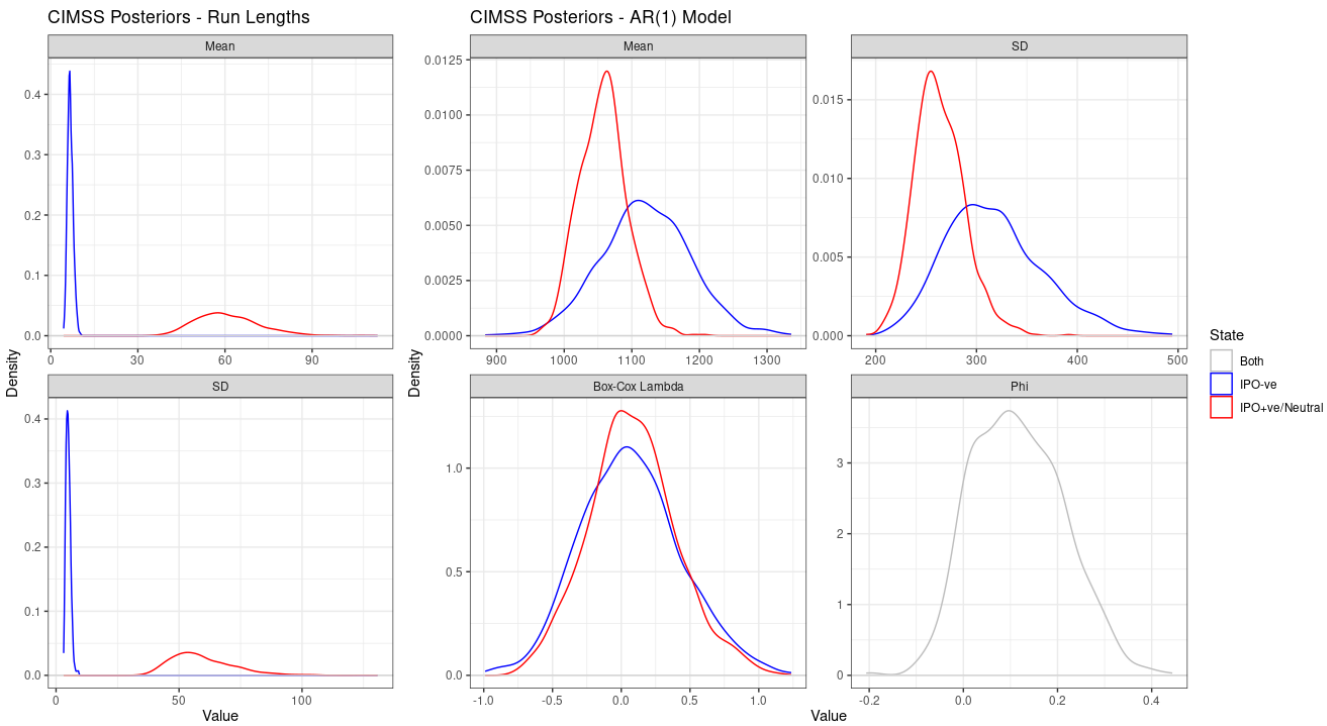

Figure 6-8: CIMSS posteriors for palaeoclimate IPO run-lengths (left) and AR(1) models calibrated to respective IPO phases (right).

Compared to the proxy-ARMA model, the CIMSS standard deviation posterior had higher variance. Posteriors for the mean and Box-Cox parameters had slightly higher variance. Higher variance in CIMSS posteriors was expected because these parameters are inferred from instrumental record subsets. In contrast, the proxy-informed model uses the entire instrumental record.

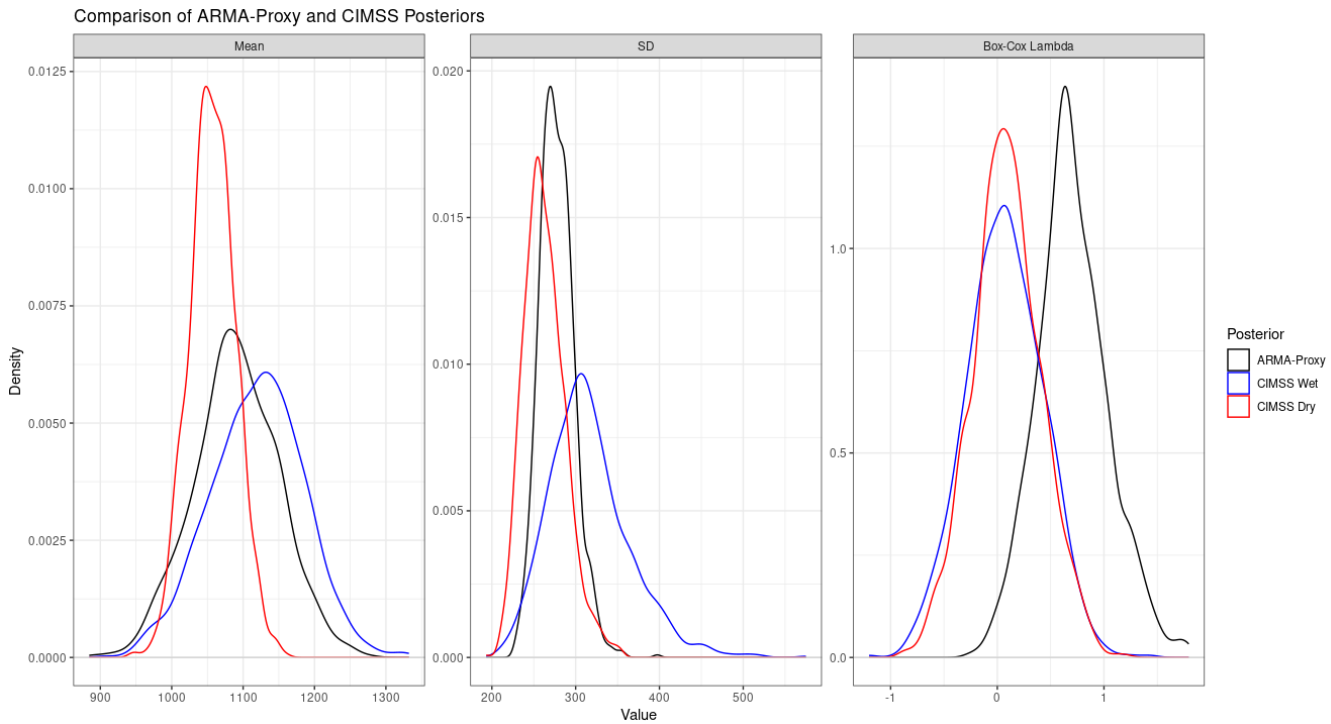

Figure 6-9: Comparison of CIMSS and proxy-informed ARMA(1,1) posteriors for the Williams River catchment.

6.8.4 Comparison of hydrological statistics

Figure 6-10 compares various hydrological statistics derived from different Williams River stochastic models. We can see that:

- The proxy-ARMA model simulated more severe droughts than all other models.
- The proxy-ARMA model simulated more variable statistics than all other models. The only exception was for the CIMSS model, which simulated more variable standard deviation and larger short-term maximums.
- The kNN model produced less severe droughts than all other models. Surprisingly, the standard ARMA model (i.e., the model calibrated using only instrumental measurements) produced more severe droughts than the paleoclimate-informed kNN model.

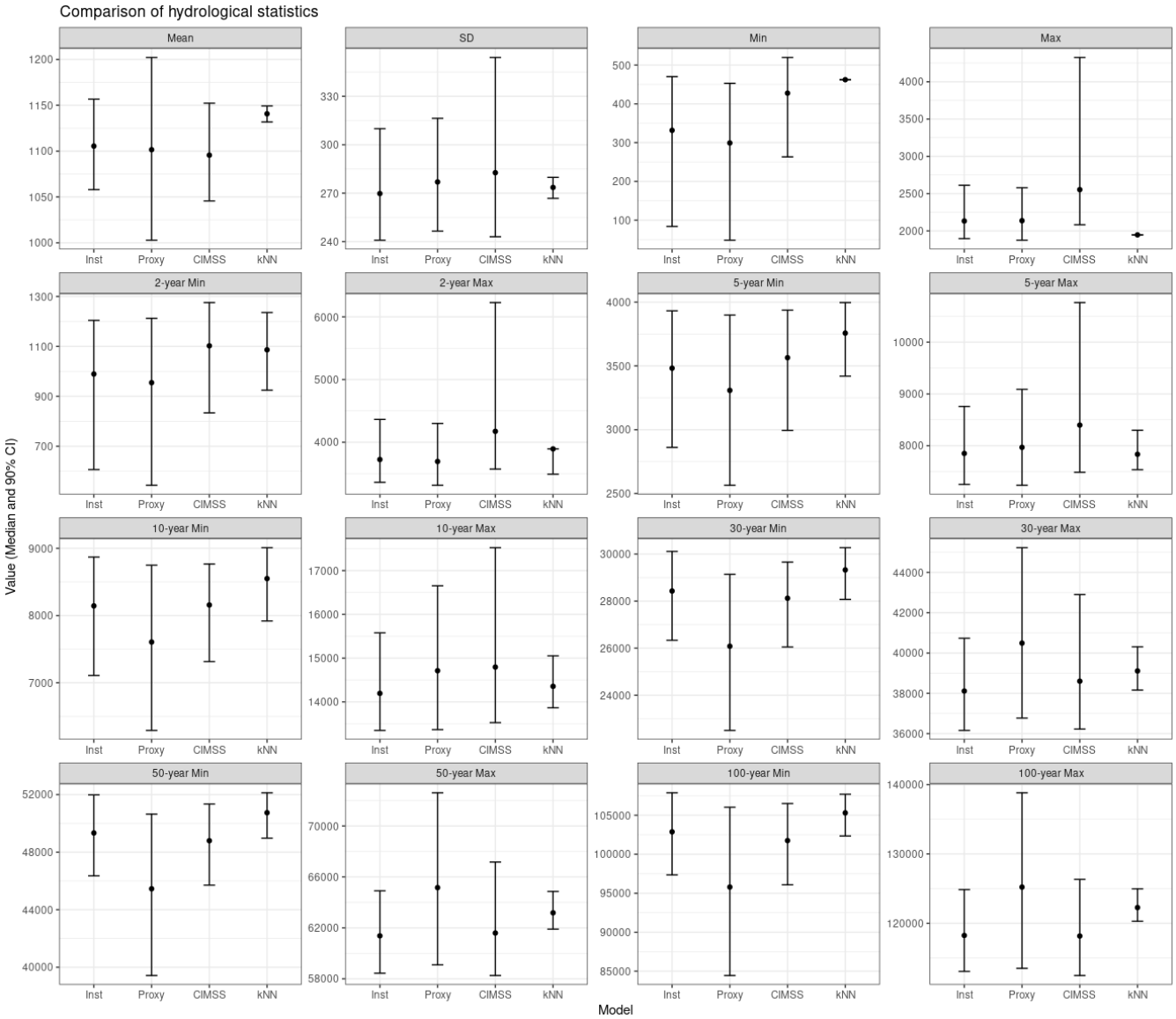

Figure 6-10: Comparison of hydrological statistics generated from the respective stochastic models. Dots show median, bars show 90% CI.

6.8.5 Exploring the limitations of the kNN method

Figure 6-10 indicates that the kNN method produced less severe low-frequency statistics than the standard ARMA(1,1) model. This is a surprising result. We expected the kNN method, which uses proxy data, to produce more severe low-frequency statistics than the standard approach (which did not use proxy data). However, note that the Law Dome proxy has relatively low skill in predicting Williams River rainfall (R^2 of ~25%). This means that resampled wet or dry proxy values may not correspond with similarly wet or dry rainfall values.

To assess if these results are due to limited proxy skill, we performed a synthetic experiment using a 'perfect-skill' proxy. This involved:

1. Scaling the Law Dome proxy record to have the same mean and standard deviation as Williams River rainfall. This method is commonly used to reconstruct rainfall and streamflow timeseries.
2. Extracting the most recent 110-years of the scaled proxy record. This is the 'synthetic rainfall', which has a perfect correlation with the 'instrumental proxy'.
3. Applying the kNN reconstruction method using the 'synthetic instrumental'. This is the 'perfect skill' reconstruction; pre-instrumental proxy values should now sample similarly wet and dry 'rainfall' values.
4. Calibrating the standard ARMA(1,1) model to the 'synthetic instrumental' then generating 1,000 rainfall replicates (2,011 years long, as per previous analyses).
5. Estimating and comparing 50 and 100-year minimums and maximums from ARMA(1,1) replicates and the 'perfect-skill' kNN reconstruction.

Results are shown in Figure 6-11. We can see that compared with the standard ARMA model, the 'perfect skill' kNN produced higher 50 and 100-year maximums. However, both models produced similar 50 and 100-year minimums. For both minimum and maximum statistics, the standard ARMA produced sampling distributions with higher variability.

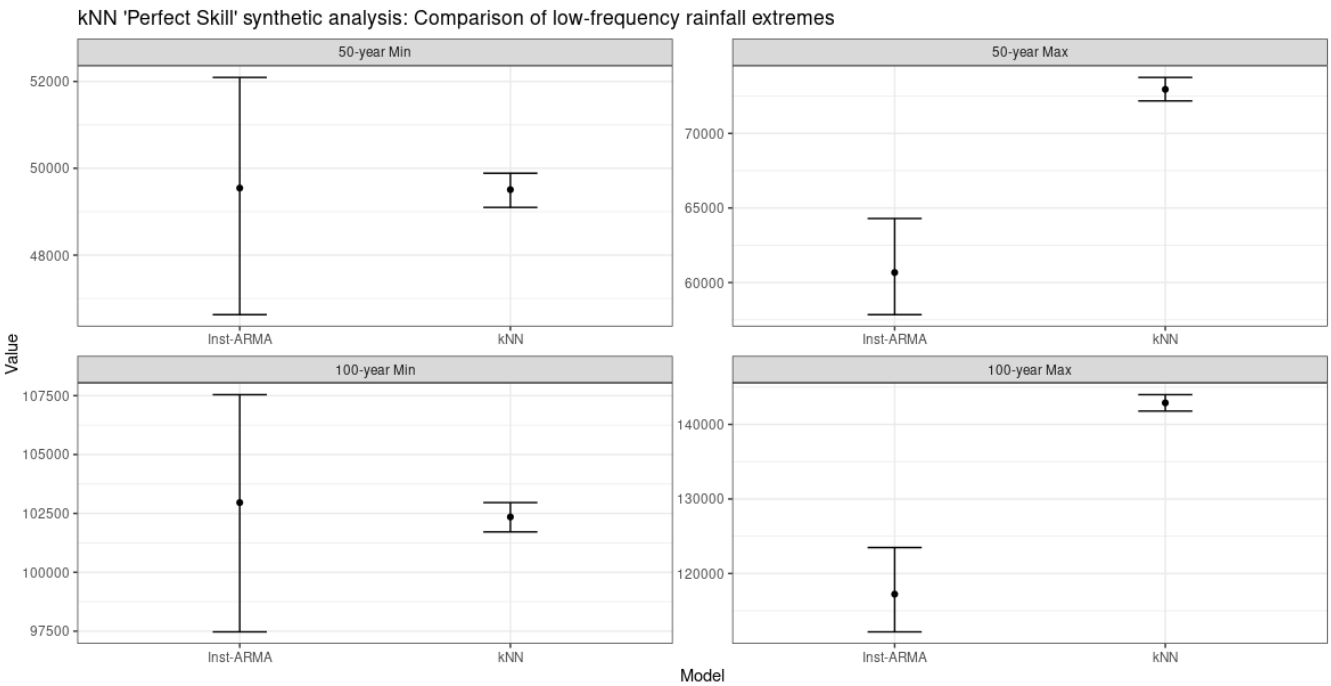

Figure 6-11: Comparison of required storages for the standard ARMA model (i.e., a model calibrated to instrumental data only) and the 'perfect-skill' kNN model. Synthetic climate data with a perfect correlation to the Law Dome proxy was used to for both models.

To further explore why the standard ARMA(1,1) model produced more severe droughts than the kNN model, Figure 6-12 compares three randomly selected rainfall timeseries generated from both models. There are three features of interest. First, the ARMA(1,1) model can extrapolate to more severe high and low flow rainfall totals. The kNN model is restricted to resampling rainfall totals from the instrumental-period. Second, the ARMA(1,1) model can generate novel sequences of wet and dry years. In contrast, the kNN method only resamples the proxy sequence. Changes in the sequencing of wet and dry years, and an ability to extrapolate to more extreme wet and dry years, will impact rainfall statistics. Third, the instrumental-period of the proxy timeseries is quite dry when compared with the pre-instrumental period. This dry instrumental-period was used to calibrate the ARMA(1,1) model, which meant that the ARMA(1,1) timeseries typically had a lower mean than the kNN method (this is also evident in Figure 6-10). However, which of these three features is driving the unexpected difference between the ARMA(1,1) and kNN models?

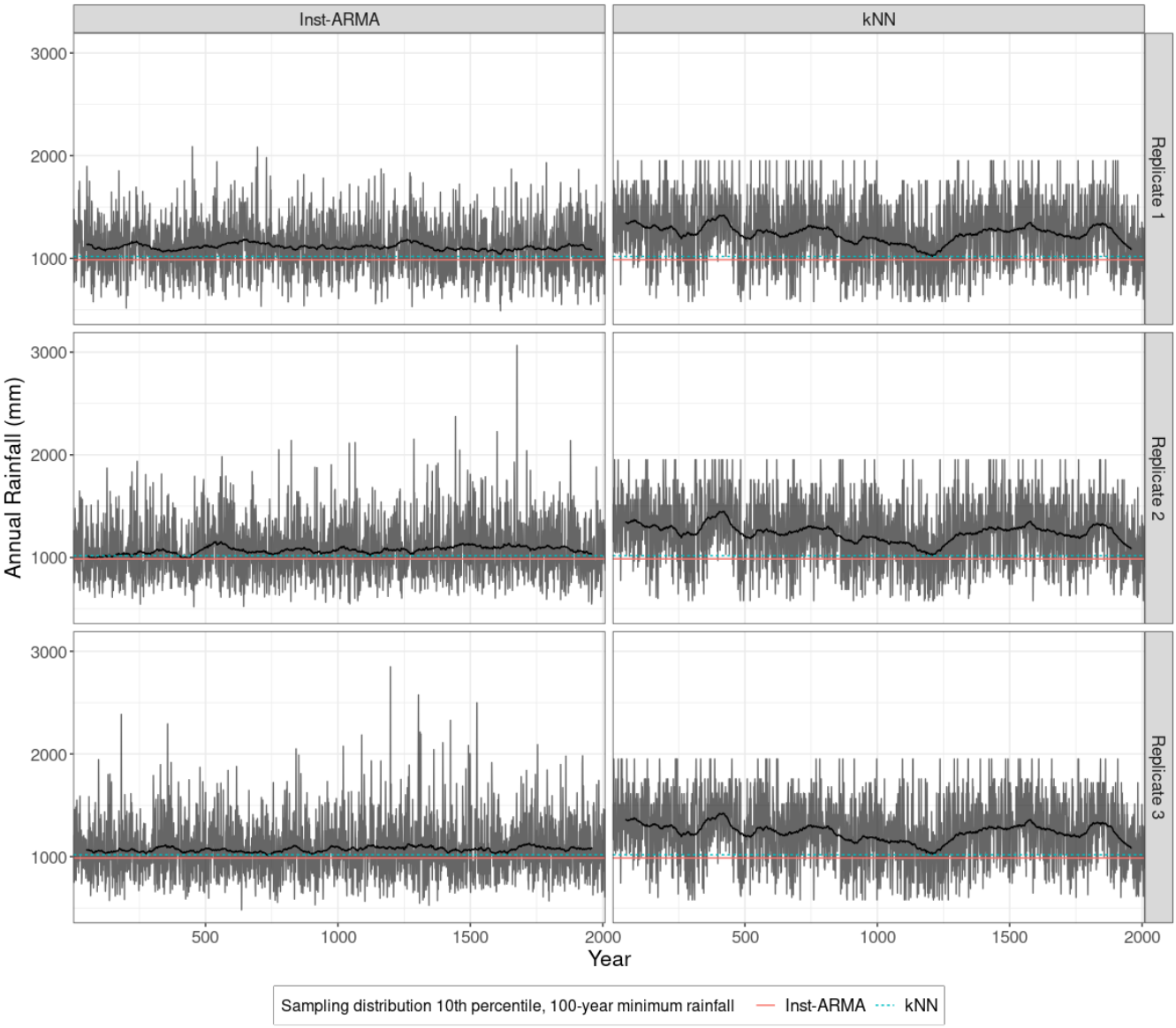

Figure 6-12: Comparison of three randomly selected Instrumental-ARMA and kNN flow timeseries for the 'perfect proxy' experiment. Solid black line shows the 100-year moving average. Coloured straight lines show the respective 10th percentile of the 100-year minimum flow sampling distribution for each model.

As a final diagnostic check, the perfect skill analysis was repeated, but with the proxy timeseries reversed. This meant that the synthetic rainfall used to calibrate the ARMA(1,1) model had a higher mean relative to some pre-instrumental periods. Rainfall statistics from this ‘perfect skill, reversed timeseries’ analysis are shown in Figure 6-13. In this instance, the kNN method did produce more lower 50 and 100-year minimums.

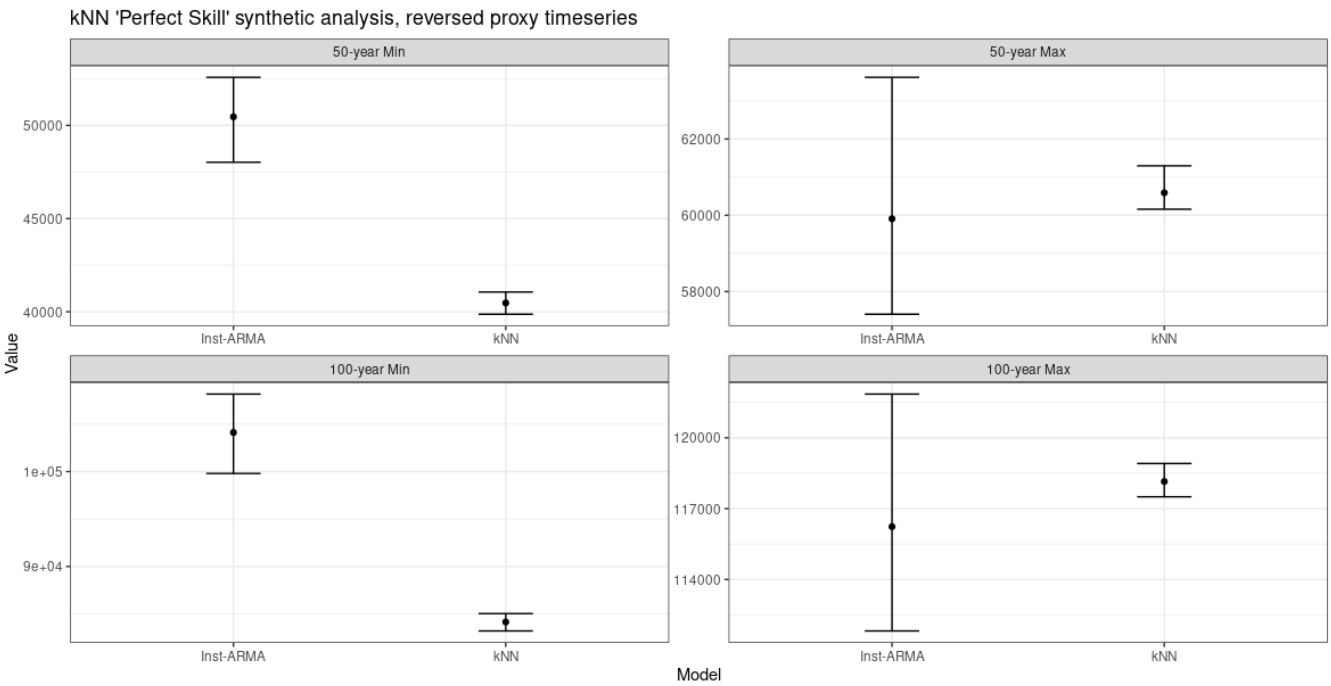

Figure 6-13: Same as, Figure 6-11 but with the Law Dome timeseries reversed prior to model calibration.

From Figure 6-11, Figure 6-12, and Figure 6-13, we can infer that, for the kNN model, repeatedly resampling the same proxy sequence can limit sampling variability in various hydrological statistics. Furthermore, the relative mean of the proxy instrumental-period will influence the magnitude of hydrological statistics. Finally, for proxies with limited skill in predicting the target variable, resultant hydrological statistics can have similar magnitudes to a standard ARMA model calibrated to instrumental measurements.

6.8.6 Comparison of required storages

Figure 6-5 and Figure 6-7 highlight that the proxy-informed ARMA(1,1) model is capable of simulating centennial-scale variability. What might this mean for the design of water infrastructure?

When using stochastic models to estimate required storages, water managers will either (a) generate stochastic replicates of equal length to the planning horizon or (b) generate extended stochastic replicates which are then considered representative of 'baseline' risk over a planning horizon.

Using a proxy-informed model, a 'baseline' risk approach will generate timeseries with much greater centennial-scale variability. In Figure 6-14, the proxy-informed model simulates much longer periods above or below the long-term mean (red dashed line). This translates to much higher required storages (because the SPA is much more sensitive to long-term change in mean, as opposed to short-term extremes). For a high demand scenario (e.g. 0.9 x MAF), the SPA indicates that centuries of water will have to be stored to ensure demand is consistently met.

This 'baseline' scenario neglects that typical reservoir planning horizons are ~50-years. Furthermore, the top row of Figure 6-14 indicates that hydroclimatic changes occur at rates slower than ~50-years, suggesting that a 'baseline' approach may overestimate risk over a planning horizon.

Instead of a 'baseline' approach, it is perhaps more appropriate to estimate the storage required to meet demand over a planning horizon (e.g. 50-years). This can be estimated using either stochastic replicates conditioned on the most recent observation (i.e., a timeseries forecast) or unconditioned stochastic replicates of equal length to the planning horizon.

Figure 6-15 compares hydrological statistics and required storages from conditional and unconditional simulations of the standard and proxy-informed ARMA models. By limiting stochastic replicate length to 50-years, the proxy-informed model produces much smaller required storages than those in Figure 6-14. The unconditional proxy-informed replicates typically produced storage distributions with greater variance and slightly higher medians. However, for low and moderate demand scenarios (0.1-0.5 x MAF) conditional simulations for the standard and proxy-informed models produced similar storage distributions.

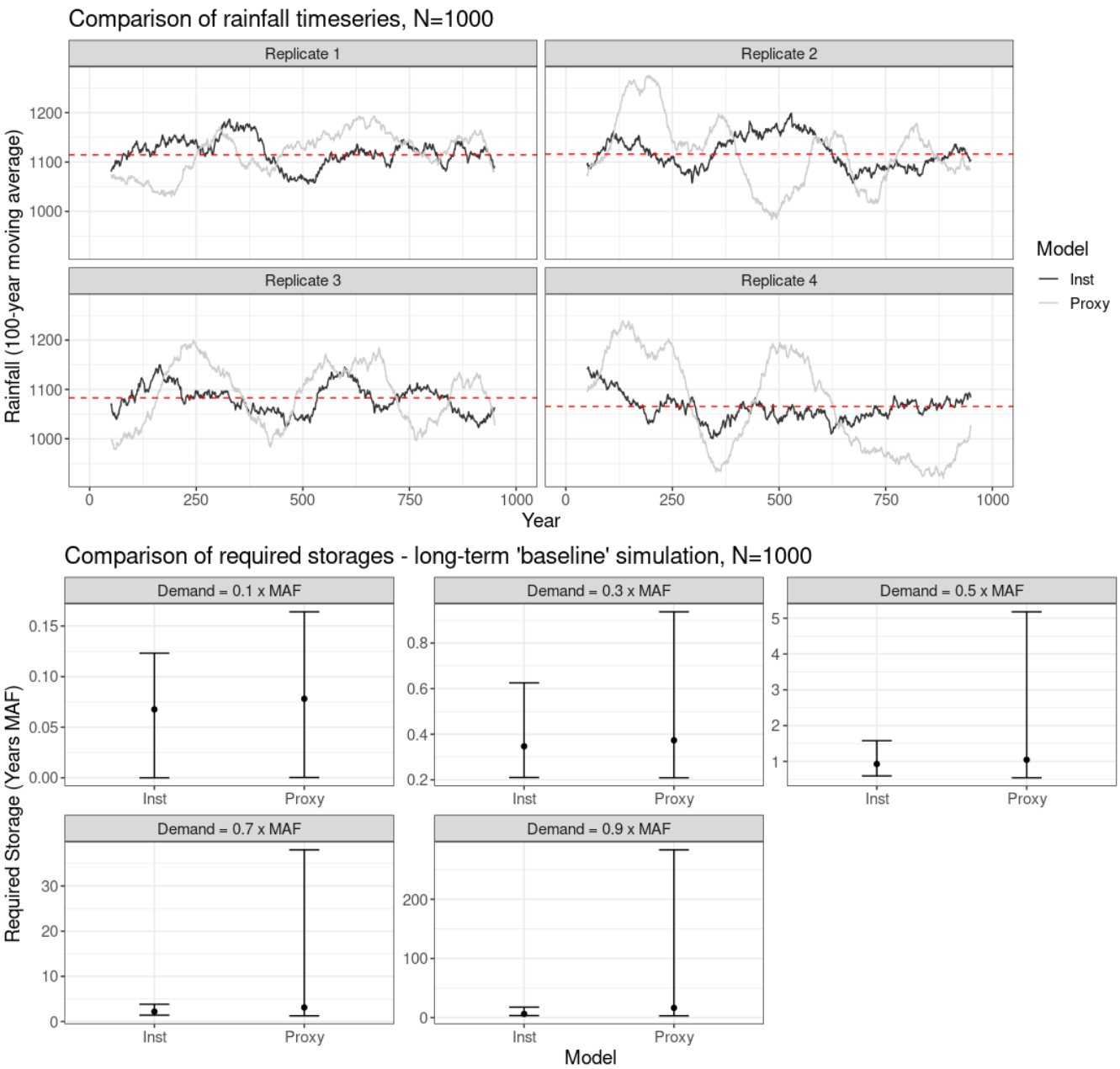

Figure 6-14: Top - comparison of four standard and proxy prior rainfall replicates. 100-year moving averages from 2,011-year replicates are shown. For each replicate, the standard and proxy informed timeseries have the same long-term mean. Bottom – distribution of required storages for different demand scenarios. Comparison of storages

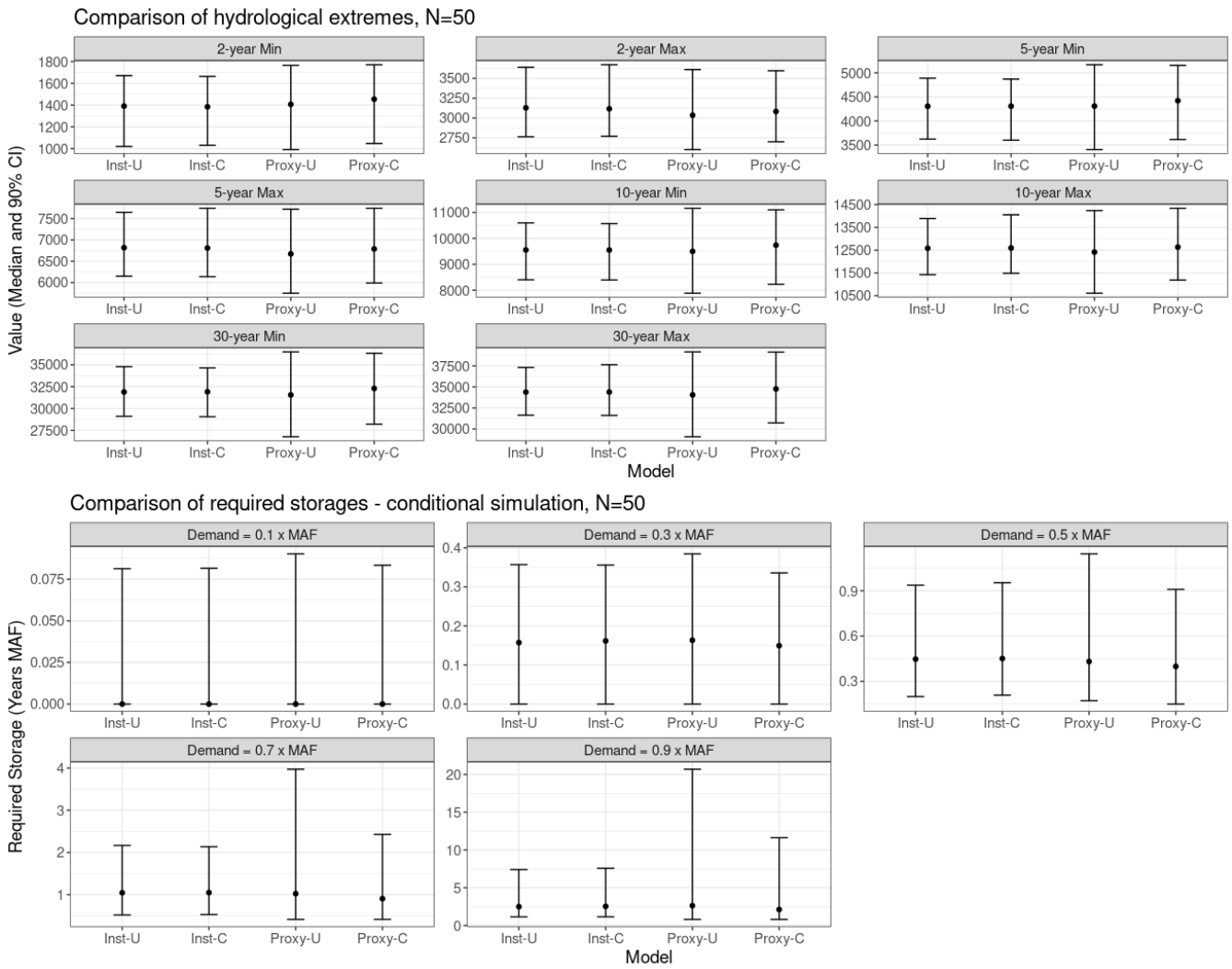

Figure 6-15: Comparison of hydrological extremes from 50-year stochastic replicates. The 'U' stands for unconditional simulation, 'C' stands for conditional simulation.

6.8.7 A cautionary note on the need to validate the proxy record model

Although not described in Section 3, we also calibrated an Autoregressive Fractionally Integrated Moving Average (ARFIMA) model using the Empirical Bayes approach. Like the ARMA(1,1) model, this model can potentially simulate realistic climate variability. However, when validated on the Law Dome proxy, the Bayesian ARFIMA model did not reproduce 100-year extremes (Figure 6-16). Potential reasons for this failure are outside the scope of this paper. Rather, we emphasise the importance of validating the stochastic model on the proxy record. This ensures that the proxy persistence is being accurately simulated when applied to catchment rainfall.

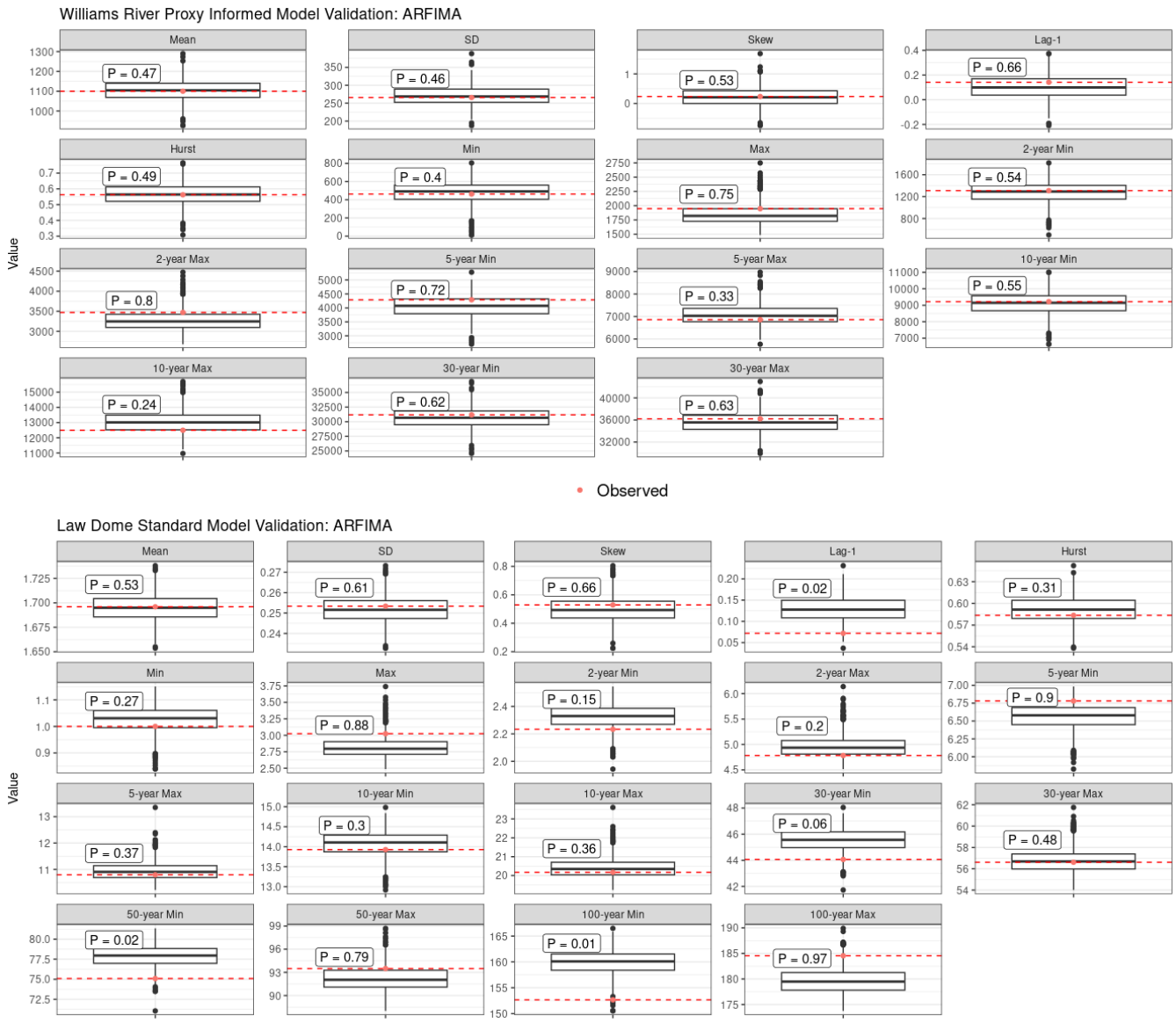

Figure 6-16: Same as Figure 6-7, but for the ARFIMA model.

6.9 Discussion

This study demonstrates several frameworks for incorporating palaeoclimate proxy into stochastic models. However, the framework matters. The proposed Bayesian framework resulted in well-defined stochastic model persistence, which in turn produced stochastic simulations of centennial-scale variability. This was a clear improvement over 'standard' stochastic model calibrated using only rainfall data. This 'standard' model had poorly defined stochastic model persistence and did not exhibit the same degree of centennial-scale variability. Surprisingly, the CIMMS and kNN proxy-informed frameworks did not demonstrate the expected benefits over the 'standard' model (i.e. simulation of more severe droughts under centennial-scale variability). Potential reasons for this result are discussed later.

The key concept proposed in this study is using palaeoclimate proxy data to inform the Bayesian calibration of stochastic model persistence. This is necessary because inferring persistence from short instrumental records is challenging, shown by the high variance posteriors in Figure 6-4. This large posterior variance means that instrumental rainfall records are not inconsistent with a white noise or an AR(1) process (i.e. 'short-term persistence') (Sun et al., 2018). In contrast, the proxy-informed ACF in Figure 6-5, and other proxy records (e.g. Markonis and Koutsoyiannis, 2016), clearly demonstrate centennial-scale variability or 'long-term persistence'. From a water management perspective, if short-term persistence is incorrectly assumed, drought risk will be underestimated, even over 50-year planning horizons (Figure 6-15). However, is such an assumption 'incorrect' if the best available source of information (i.e., instrumental records) does not contain contrary evidence? From a water management perspective, this highlights a key use for palaeoclimate data: it can reject 'short-term' persistence models due to better defined persistence.

Although proxy records can better define persistence and reduce parameter uncertainty, a key caveat is that proxy information is potentially biased. These biases can arise due to (a) non-climatic signals in the proxy record; (b) confounding climate signals, for example a temperature signal that increases proxy timeseries persistence (Franke et al., 2013b); and/or (c) statistical processing of 'raw' proxy measurements, which can modify the climate signal (Razavi and Vogel, 2018). This means that proxy records should be evaluated for biases before stochastic model calibration.

In terms of ice core records and persistence, snow accumulation records are unbiased and sea salt records have a positive bias (Chapter 2). However, these bias evaluations are limited by observation length. This is because 'long' rainfall records (e.g., 100-150 years) are still subject to considerable sampling uncertainty, meaning that small biases will not be detected.

For the proposed Bayesian framework, there is a clear trade-off between reducing parameter uncertainty (meaning persistence is better defined) and calibrating models with potentially biased proxy data. The relative importance of either is, in our opinion, a value judgement to be made by the modeller. For this case study, proxy persistence was well defined and clearly indicative of long-term persistence. In contrast, instrumental persistence was poorly defined

and not inconsistent with white noise. Because instrumental persistence was so poorly defined, we were willing to accept some proxy bias.

Aside from better defined persistence, an additional advantage to the proposed Bayesian approach is that no assumptions of stationary climate-proxy relationships are made. This assumption underpins most existing palaeoclimate reconstructions (and the kNN method). For remote proxies, a stationary climate-proxy relationship requires the atmospheric 'links' between sites to remain relatively unchanged in the pre-instrumental period. This is difficult to validate. Instead, the Bayesian approach makes more (somewhat) testable assumptions of (a) persistence parameters being stationarity (demonstrated in Results Chapter 4); and (b) common persistence signals across large regions (demonstrated by Fatichi, Ivanov and Caporali, 2012; Iliopoulou *et al.*, 2018; and Tyralis *et al.*, 2018). Insufficient evidence against stationary, and common, persistence over the mid-latitudes and Antarctica allows the incorporation of centennial-scale variability in stochastic modelling.

Various, different proxy records provide evidence for centennial-scale variability – so, what causes this variability? The most common explanation links rainfall variability with low-frequency changes in sea surface temperature (SST) anomalies (e.g., ENSO and the IPO). These changes in SST anomalies can then be linked with aerosol forcing from pulses of volcanic activity (Mann et al., 2021, 2020). Another explanation links climate variability with low-frequency changes in solar irradiance (Ait Brahim et al., 2018; Raspopov et al., 2008). Finally, various work shows that statistical descriptions of centennial-scale variability can be derived from the Principle of Maximum Entropy (Koutsoyiannis, 2011, 2005). Irrespective of what causes centennial-scale climate variability, it occurs. Therefore, it must be accounted for in water management – this study provides a framework for which this variability can be quantified, which can inform subsequent management philosophies and decisions.

Any question of how to manage water under centennial-scale climate variability must first consider the rate at which climate can change – will it be gradual, or sudden? More specifically, for a specific infrastructure planning horizon, at which time lags are historic observations useful for inferring future risk? The very nature of 'low-frequency' variability suggests that change, and therefore changing risk, occurs gradually. This was highlighted in Chapter 5, which demonstrated that, when inferring the mean and standard deviation of a timeseries under historic climate variability, (a) 100 years of observations can be used as a proxy for risk in the

next 100 years and (b) 1,000 years of observations *cannot* be used as a proxy for risk in the next 1,000 years. Points (a) and (b) suggest that climate risk is historically non-stationary, but not over typical infrastructure planning horizons.

Climate non-stationarity and centennial-scale variability highlight limitations with inferring climate risk using long-term 'baseline' simulations (e.g. Figure 6-14). Instead of designing a system that is robust under centennial-scale variability, it may be preferable to monitor climate, re-condition risk estimates based on recent observations, and, if necessary, adapt the system so it satisfices some management goals over a planning horizon. This monitor, re-condition, and adapt approach is aligned with the philosophies of info-gap decision theory and dynamic adaptive policy pathways (Ben-Haim, 2010; Haasnoot et al., 2013).

Irrespective of how climate risk is managed, this study suggests that over a planning horizon, the range of plausible future hydroclimatic trajectories is irreducibly 'wide'. This 'wide' uncertainty is driven by (a) the somewhat random nature of climate variability (i.e. aleatory uncertainty), (b) parameter uncertainty, and (c) centennial-scale variability. These factors mean that water supply systems and management plans must be robust under a future range of drought risk that is irreducibly 'wide', hence the term 'wide uncertainty'.

As well as 'wide' uncertainty, managing water under future climate risk must also contend with 'deep' uncertainty under anthropogenic climate change (Hallegatte et al., 2012; Kwakkel et al., 2016a). Deep uncertainty describes how we cannot reasonably assign probabilities to future climate risk factors, such as future socio-economic development and associated greenhouse gas emissions or land use changes (Lempert et al., 2006).

There are numerous risk management paradigms that examine system-specific risks and vulnerability, all with common themes of 'bottom-up' and 'scenario-neutral' risk assessments (Ben-Haim, 2006; Culley et al., 2016; Hall et al., 2012; Lempert et al., 2006). These paradigms were originally developed to manage water under 'deep' uncertainty – naturally, they can be applied to 'wide and deep' uncertainty.

A key theme linking various paradigms for managing water under 'wide and deep' uncertainty is the need to explore some future climate risk space - stochastic models are an ideal tool for doing so. Stochastic models can rapidly generate a wide range of climate information, which

is necessary to identify system vulnerabilities (Fowler et al., 2022). Stochastic model parameters can also be perturbed, which can produce hypothetical climate timeseries that incorporate climate change signals (Guo et al., 2018; McInerney et al., 2023). Furthermore, stochastic models can incorporate palaeoclimate variability or have non-stationary climate variables as a covariate (Kiem et al., 2021). In particular, the Bayesian framework presented in this study can be extended to forecast future climate and include additional climate information (e.g. climate model projections), meaning it can be used to characterise 'wide and deep' uncertainty.

Paleo-stochastic models are well suited for use in various existing risk assessment and management frameworks. However, the paleo-stochastic model used matters. This study highlights limitations with the kNN and CIMSS frameworks. Both produced similar required storages to the standard stochastic model (i.e., the ARMA(1,1) model calibrated using only instrumental data). When using these models in the study catchment, the inclusion of palaeoclimate information offered minimal benefits over simpler, rainfall-only methods.

The kNN model is constrained to resample instrumental data based on a single proxy sequence. However, due to the chaotic and random nature of climate variability, different sequences are possible. Furthermore, the instrumental-period of the proxy sequence may be relatively drier than the pre-instrumental period. This means that kNN method will repeatedly resample the same, wetter pre-instrumental period. In contrast, the proposed Bayesian model can generate novel sequences that are wetter and drier than the instrumental period, leading to a more robust characterisation of climate risk.

The CIMSS framework was limited by the weak influence of the IPO on catchment rainfall. In the Williams River, the IPO 'wet' phase had a slightly higher mean rainfall than the 'dry' phase. Despite the use of proxy data to simulate wet and dry 'run-lengths', these wet and dry runs still had similar mean states, resulting in similar required storages to the standard ARMA(1,1) model. Unlike CIMSS, the proposed Bayesian method does not depend on large-scale climate drivers when inferring wet and dry values. Because wet and dry values are not defined based on poorly separated instrumental sub-periods, the proposed Bayesian model can extrapolate to more severe wet and dry values. This also explains why the Bayesian model produced high statistical uncertainty than the CIMSS model (even though parameter uncertainty was smaller).

Due to the kNN and CIMSS results, we emphasise that future development of paleo-stochastic models should include a comparison with a 'standard' stochastic model (i.e., a stochastic model calibrated to instrumental measurements, not proxy data). Such a comparison is necessary to demonstrate the added benefits (or not) of using proxy data. Ideally, the comparison will focus on the perceived benefits of proxy data. Therefore, when comparing proxy-informed and standard stochastic models, the following guiding questions may be useful:

- Can the proxy-informed model preserve instrumental-period variance?
- Are low-frequency hydrological statistics more extreme or better defined in the proxy-informed model? This is contingent on the fidelity of the proxy low-frequency signal.
- Is parameter uncertainty reduced in the proxy-informed model? This is contingent on the stochastic modelling framework used.

When comparing proxy-informed and standard stochastic models, it is also crucial to compare synthetic timeseries of equal-length. In this study, both short-term and extended synthetic timeseries were examined. Comparing extended timeseries is an important and often overlooked part of model evaluation (and analysis of paleoclimate records in general). This accounts for the increased likelihood of longer timeseries having larger extremes, purely due to chance (e.g., the maximum of 1,000 draws from a standard normal distribution will be greater than the maximum of 100 draws). Perhaps instrumental and paleoclimate timeseries are statistically consistent, but the longer paleoclimate record coincidentally contains larger extremes?

Issues of model evaluation aside, the proposed modelling framework is flexible and can be extended. The key concept is using additional climate data to inform stochastic model calibration. This can be applied to any stochastic model, parameter, or calibration method. Furthermore, different proxy types could be used to calibrate different stochastic model parameters or used as stochastic model covariates. This would reproduce non-stationarities contained in the covariate proxy record. These potential applications highlight how the modelling framework can leverage the respective strengths of different proxies (i.e., unbiased persistence signal in ice cores, catchment-scale non-stationarities in tree-rings) without relying on predictive skill-based selection criteria. Such criteria can overlook lower skill records that still contain useful, and different, information about hydroclimatic variability (e.g., ice cores).

The general flexibility of the Bayesian framework means that additional case studies using different proxies and catchments can be conducted quickly. Will other proxy records and catchments produce similar results? Such replications are needed to better understand if and how centennial-scale climate variability can impact water security.

6.10 Conclusion

In this study, a Bayesian framework for calibrating a stochastic rainfall model using palaeoclimate proxy data was presented. The framework uses proxy data to define rainfall persistence priors. The use of proxy-informed priors serves two purposes: (1) it constrains persistence parameter uncertainty and (2) it incorporates proxy centennial-scale variability that is missing in the short instrumental rainfall record. This results in stochastic rainfall simulations that incorporate realistic centennial-scale climate variability. The framework also explicitly preserves rainfall coefficient of variation, which in turn results in rainfall outputs that can be used in operational water management. Finally, the framework simulates more severe droughts than other existing proxy-informed stochastic modelling methods.

Chapter 7. Final discussion

Before discussing the climate risk and water management implications of this thesis, it is important to discuss the assumptions made when using palaeoclimate proxy records to infer historic climate variability and climate risk.

Underpinning this thesis, and all palaeoclimatology, is an assumption that the proxy-climate relationship is stationary and can be inferred from the instrumental-period. This assumption is both necessary and potentially flawed.

When considering limitations with this assumption, note that numerous factors influence proxy formation and properties. The relative influence of these different factors may change between instrumental and pre-instrumental periods (e.g. forest stand dynamics might mean that some periods are more conducive to tree growth than others - Cook, 1985). This means that the statistical model calibrated from the instrumental period may not be wholly suitable for some pre-instrumental periods (D'Arrigo et al., 2008). In contrast to local proxies, remote proxy reconstructions assume that the atmospheric processes linking the proxy with the target climate variable remain relatively unchanged in the pre-instrumental period. This is hard to validate and may not be likely – remote proxies may be skilful predictors of climate for some periods and not others (Kiem et al., 2020).

For standard statistical reconstructions, assuming a stationary proxy-climate relationship can be particularly limiting. However, even reconstruction methods which use, in part, physically based climate models require proxy forward models (through a process of data assimilation) (Dee et al., 2016; Steiger et al., 2017, 2014). The parameters of these proxy forward models are informed by instrumental-period statistical relationships (Tolwinski-Ward et al., 2011). No matter the method, look close enough and you will see that, at some point, an assumption is made that pre-instrumental proxy-climate relationships are like those observed in the instrumental period.

In a sense, this assumption is 'the worm at the core' of paleoclimatology. It is easily hidden, hard to remove, and somewhat unsavoury.

Calling this assumption the 'worm at the core'[1] of paleoclimatology may be considered unfair. After all, won't any scientific field based on passive observation and measurement, as opposed to controlled experiments, also have a 'worm at the core'? Perhaps, but the 'worm at the core' of paleoclimatology is particularly pernicious when one considers the implications of proxy variability. More specifically, that pre-instrumental climate, and by extension future climate, might be substantially different to instrumental climate.

Various reconstruction studies have discussed pre-instrumental climate in terms of 'megadroughts' (Cook et al., 2022; Helama et al., 2009; Routson et al., 2011; Stevenson et al., 2022). Naturally, a megadrought invokes a concern for water security.

Megadroughts have quite severe social and economic impacts (Fernández et al., 2023; Muñoz et al., 2020), but adapting a water supply system to mitigate the risks posed by megadrought also has social and economic impacts (Gober et al., 2016). Therefore, any water security concerns about megadrought should be viewed with respect to the underlying assumption made when producing the reconstruction – the 'worm at the core'.

When examining pre-instrumental megadroughts, and paleoclimate reconstructions in general, note that a key implication of the 'worm at the core' is that the reconstruction skill remains constant across instrumental and pre-instrumental periods. Most reconstructions have an R^2 between 0.3 and 0.6, meaning they explain 30-60% of the variability in the target climate variable. What about the remaining 40-70%? Quantifying this unexplained variance is crucial for accurately inferring climate risk and managing water.

Linear reconstruction methods assume that this unexplained variance has a mean of zero and is independent and identically distributed. This means that, for an R^2 between 0.3 and 0.6, around half of the reconstruction is assumed white noise. Is this a reasonable assumption? What if, during the megadrought, this unexplained variance has a structure that counter acts the megadrought? This cannot be answered! So, considering this limitation, is it reasonable to

[1] The phrase 'worm at the core' was taken from a book describing the psychological sub-field called Terror Management Theory (Solomon et al., 2015)

adapt a water supply system based on the megadroughts implied by palaeoclimate reconstructions?

Despite the 'worm at the core', there are physical explanations, and corresponding evidence from different proxy types, indicating that climate varies at centennial and millennial timescales (i.e. megadroughts are possible). From a water security perspective, such variability is concerning because the statistical models used to infer climate risk (e.g. stochastic models), and the instrumental data these models were calibrated to, cannot account for centennial and millennial scale variability (Chapter 3).

Therefore, irrespective of proxy limitations, there was a clear need to use palaeoclimate information to re-evaluate and update the statistical models/assumptions used to infer climate risk in water management. This was the key motivation of this thesis.

When exploring how to use palaeoclimate data in climate risk assessment and water management, throughout this thesis we have been mindful of the 'worm at the core' of paleoclimatology. Instead of using a remote proxy to predict catchment rainfall, we assessed the fidelity of remote proxy persistence (Chapter 2), the stationarity of that persistence (Chapter 5), and then used proxy persistence to inform the calibration of a catchment-scale stochastic rainfall model capable of reproducing proxy variability (Chapter 3 and Chapter 6). In doing so, we sidestepped the limited skill of remote proxies and produced stochastic rainfall data with a low-frequency signal that, based on Chapter 2 and Chapter 4, we have insufficient evidence to say is unrealistic.

Aside from the modelling framework presented in Chapter 6, examining proxy records (as opposed to reconstructions) also lead to additional insights on the nature of climate risk and the limitations with using stochastic models to infer climate risk. For example, Chapter 5 examined the stationarity assumption underpinning 'traditional' water management. Marginal evidence against stationarity was found, which raises questions as to if and how historic non-stationarity should be considered in climate risk modelling.

From our perspective, the key issue facing climate risk modelling is not removing (or hiding) all stationarity assumptions. Rather, when assuming a particular parameter or relationship is stationary, how wrong are we prepared to be? This is a somewhat subjective judgement, which

depends on (a) the modelling task, (b) whether non-stationary models are demonstrably better than a stationary model, and (c) the consequences of wrongly assuming stationarity.

From a climate risk perspective, the consequences of wrongly assuming stationarity are, perhaps, most important. Chapter 5 provided evidence that, under historic climate variability, (a) stochastic model mean and standard deviation posteriors are similar at centennial timescales, but not multi-centennial and millennial timescales and (b) stochastic model persistence posteriors are similar across centennial, multi-centennial, and millennial timescales. This means that under historic climate variability and over a 100-year planning horizon, assuming stationarity can still produce reasonable climate risk estimates.

Evidence for stationary persistence, combined with the large uncertainty in persistence inferred from instrumental records, motivated the method presented in Chapter 6, which aimed to reduce the parameter uncertainty in stochastic model persistence. Although the method reduced parameter uncertainty, it also produced larger statistical uncertainty for various hydrological statistics when compared with a standard calibration method. The larger statistical uncertainty, also referred to as 'stochastic' or 'aleatory' uncertainty, arose due to the incorporation of proxy centennial-scale climate variability in stochastic model outputs. This uncertainty is unavoidable and must be managed. Centennial-scale variability, aleatory uncertainty, and parameter uncertainty means that water supply systems must be robust under a future range of drought risk that is irreducibly 'wide', hence the term 'wide uncertainty'.

'Wide' uncertainty stems from parameter uncertainty, aleatory uncertainty, and centennial-scale climate variability – however, managing water under future climate risk must also contend with 'deep' uncertainty under anthropogenic climate change (Hallegatte et al., 2012; Kwakkel et al., 2016a). Deep uncertainty describes how we cannot reasonably assign probabilities to future climate risk factors, such as future socio-economic development and associated greenhouse gas emissions or land use changes (Lempert et al., 2006).

'Wide and deep' uncertainty poses challenges to the traditional engineering approaches of inferring climate risk, then designing systems to mitigate said risk (i.e. 'predict then act' approaches). For example, a water supply system may have been designed to mitigate a 1-in-10,000-year drought. Typically, this drought was defined using a single stochastic model parameter set derived from an instrumental record (meaning parameter uncertainty and

centennial-scale variability was not considered). When updating system infrastructure or rules, a water manager may want to calculate a 1-in-10,000-year drought under 'wide and deep' uncertainty. However, what does a 1-in-10,000-year drought even look like under centennial-scale climate variability, non-stationarity, and deep uncertainty? This is hard to define using conventional statistical methods (Read and Vogel, 2015). Instead, over a particular planning horizon, it may be preferrable to look at system-specific risks of failure under a wide variety of scenarios (Borgomeo et al., 2018; Serinaldi, 2015).

There are numerous risk management paradigms that examine system-specific risks and vulnerability, all with common themes of 'bottom-up' and 'scenario-neutral' risk assessments (Ben-Haim, 2006; Culley et al., 2016; Hall et al., 2012; Lempert et al., 2006). These paradigms were originally developed to manage water under 'deep' uncertainty – naturally, they can be applied to 'wide and deep' uncertainty. The overarching philosophy of these paradigms is to 'stress-test' the system under a wide variety of plausible climate scenarios, identify system vulnerabilities under these scenarios, then develop operational rules and infrastructure whereby (a) system performance is insensitive to climate scenario (i.e. are 'robust') (Shortridge et al., 2017; Stanton and Roelich, 2021) or (b) the system is adapted pursuant to some decision threshold being crossed (Haasnoot et al., 2020, 2013).

When considering water management under 'wide and deep' uncertainty, results from Chapter 6 highlight potential difficulties with achieving system robustness under the long-term baseline risk posed by centennial-scale variability. For a high demand scenario, a 'robust' reservoir may require hundreds of years of storage! Such storage is clearly infeasible.

Instead of designing a system that is robust under centennial-scale variability, under 'wide' uncertainty it may be preferable to monitor climate, re-condition risk estimates based on recent observations, and, if necessary, adapt the system so it satisfices some management goals over a planning horizon. This monitor, re-condition, and adapt approach is aligned with the philosophies of info-gap decision theory and dynamic adaptive policy pathways (Ben-Haim, 2010; Haasnoot et al., 2013).

Although scenario-neutral, bottom-up, info-gap, and dynamic adaptation approaches are useful frameworks for understanding and managing risk under 'wide and deep' uncertainty, key questions related to these approaches remain. Four questions are listed below:

1. How should you select the scenarios when stress testing a system or identifying trigger points?

Despite being described as ‘scenario neutral’, scenario selection influences which system adaptations are considered robust (Quinn et al., 2020). For a truly ‘scenario neutral’ framework, system robustness should not be overly sensitive to scenario selection. However, is scenario neutrality even possible and, if so, how do you achieve it under ‘wide and deep’ uncertainty?

2. When monitoring decision thresholds using ‘signposts’, how do you know that a particular signpost has been crossed?

Under ‘wide’ uncertainty, this is a particularly difficult question to answer. Climate naturally fluctuates and trends over centennial-scales – a potential signpost should be relatively robust to these fluctuations (Haasnoot et al., 2018). Identifying signposts with a low signal-to-noise ratio is important but, under ‘wide’ uncertainty, can we expect such a signpost? Further research is needed that considers how making an adaptation decision based on a signpost will depend on (a) the relative consequences of a false alarm versus a 'miss' and (b) how palatable these relative consequences are (i.e. the ‘risk profile’ of the decision-maker). Instead of passively monitoring a signpost to inform decision-making, it may be preferable to couple the signpost with various decision-making risk profiles. These risk profiles are, in turn, informed by the socio-political context in which the adaptation decision is being made.

3. How should the water system boundaries be defined during modelling and risk assessment?

Westra and Zscheischler, 2023 refer to this issue as boundary critique (taken from systems engineering concepts). Boundary critique acknowledges that water supply systems do not exist or operate in isolation – they are embedded within broader social and economic systems (Falkenmark, 1977; Wang et al., 2023). Boundary critique is a process of identifying trade-offs between the need to consider (or not) these broader systems and the impossibility of modelling ‘everything’ (Westra and Zscheischler, 2023). In essence, what level of model complexity is necessary for given objectives of a risk assessment, and how can this be identified? Global sensitivity analysis seems an ideal tool for boundary critique.

4. How might adaptation decisions made today, and the risk assessment/modelling methods informing these decisions, impact future adaptation decisions, and the risk assessment/modelling methods informing future decisions?

There are various economic and socio-political factors that can dictate what modelling methods and adaptation options are permissible (Hurlimann and Dolnicar, 2010). These economic and socio-political factors are, in turn, influenced by stakeholder engagement during model development and adaptation scoping (Ross et al., 2014). This mutual influence will determine 'institutional capacity for adaptation', a concept often discussed in the development studies literature (Domorenok et al., 2021; Mortreux and Barnett, 2017; Smit and Wandel, 2006). A key, and often overlooked, outcome of this mutual influence is that a legacy of prior stakeholder engagements and modelling methods will influence policy and modelling updates (Lim et al., 2023). In short, institutional capacity for adaptation has a 'memory' that can either limit or enhance the risk assessment and modelling process (Barnett et al., 2015). However, institutional capacity for adaptation is, typically, only implicitly considered during the initial scoping of system risks and potential adaptations. Furthermore, sociological analyses of institutional capacity tend to overlook the overarching role of risk assessment and modelling in defining and building said capacity. Therefore, explicit consideration of institutional capacity before, during, and after the risk assessment and modelling process may be needed to ensure institutions can proactively manage and adapt to risks posed by 'wide and deep' uncertainty. This is a specific example of what Westra and Zscheischler (2023) call 'embedding second-order learning within risk assessments' – in essence, using the concept of institutional capacity to frame and guide second-order learning during the risk assessment and modelling process.

Questions 3 and 4 stem from the inherent complexity of coupled dynamic systems, highlighting that decision-making and climate risk assessment exist within and interact with socio-political systems. The nature of system complexity means that (a) future system trajectories are influenced by innocuous, seemingly random perturbations to the current state of the system; (b) there are various, potentially unknown feedbacks and interactions between coupled sub-systems (e.g. the coupling of socio-political decision making systems with physical environmental systems); and (c) risk can compound, cascade, and emerge due to the interactions between sub-systems (Lorenz, 1969; Lux, 1998; Simpson et al., 2021).

Uncertainties arising from system complexity are, in our opinion, best described as 'deep'. However, 'deep' uncertainty exists on a spectrum from 'clear' to 'murky' that describes the fidelity with which the deeply uncertain feature can be described. For example, future land use changes, although deeply uncertain, can be described. Moreover, future rainfall changes, although deeply uncertain, can also be described within some user-defined upper and lower bound (Brown et al., 2012; Mortazavi-Naeini et al., 2015; Turner et al., 2014). Therefore, both future rainfall and land use changes are deeply uncertain, however, they can be defined clearly. In contrast, the deep uncertainty associated with complex systems, particularly those that include socio-political sub-systems, is inherently 'murky'.

'Murky' uncertainty describes the deeply uncertain features and interactions that are either (a) poorly understood or (b) hard to objectively define. Poorly understood features include, but are not limited to, those arising from system complexity. Hard to objectively define features include, but are not limited to, social values and goals.

An example of a poorly understood system feature arising from system complexity is the so-called 'Irrigation Efficiency Paradox'. This paradox refers to empirical evidence that increasing irrigation efficiency at a farm-scale does not reduce water consumption at the basin-scale (Burt et al., 1997). Grafton *et al.*, 2018 proposed two causes of the paradox: (1) drip irrigation leading to a reduction in recoverable, reusable, return flows and (2) water savings being offset by the expansion of irrigated areas.

Now, consider a risk assessment exercise that, in part, wants to consider the impact of increasing irrigation efficiency on overall river system outcomes. This assessment will likely involve modelling of the river system. Empirical evidence indicates that, in a realistic model, an increase in farm-scale irrigation efficiency will not reduce basin-scale consumption (the Irrigation Efficiency Paradox). However, there has been limited research on modelling the Irrigation Efficiency Paradox. Some research proposes more detailed surface water-ground water modelling to represent recoverable and non-recoverable return flows (Xiong et al., 2021). Other research proposes incorporating the changing behaviour of irrigators within a coupled socio-environmental model (Ilyas et al., 2021). Crucially, further research is certain! So, what are we to do when there is reasonable evidence that a process influences system behaviour and risk, but there is limited understanding or agreement on how to model this process? The

subsequent modelling choice, and the sensitivity of the model results and inferred risk to this choice, gives rise to a very 'murky' uncertainty.

In contrast to poorly understood features, which may be better understood after further research, there are also deeply uncertain features that are inherently hard to define. For example, a suitable water system adaptation may depend on satisficing some social goals, such as economic 'fairness'. Social goals are hard to define, let alone model and optimise. Attempts to define, model, and optimise some heuristic social goal will require various trade-offs between multiple stakeholder values, data available to measure a potential goal heuristic, and the need for conventional system modelling and optimisation tools to represent goals clearly (Wu et al., 2023). These trade-offs will influence model results. However, the corresponding sensitivity and uncertainty is impossible to define – giving rise to 'murky' uncertainty.

For a climate risk assessment, modelling a system subject to 'murky' uncertainty requires two subjective choices. First, we must choose the system features and interactions to model – perhaps through a process of sensitivity analysis and boundary critique, or perhaps based on the expertise and capabilities of the modeller. Second, we must define hard to define features – perhaps through a process of stakeholder engagement, or perhaps based on data and modelling constraints. Regardless of process, the nature and subjectivity of both choices means that (a) no model set-up can be considered 'optimal' and (b) model goals are shaped by potentially conflicting stakeholder values. This can make model development highly contested. Crucially, the modelling choices made under 'murky' uncertainty will significantly influence model outputs and, by extension, outcomes.

The concept of 'murky' uncertainty invokes similar issues to that of a 'wicked' problem - but what is a 'wicked' problem? A 'wicked' problem is one involving multiple stakeholders and decision-makers with conflicting interests and values, whereby the problem definition is contested and ill-formed (Rittel and Webber, 1973). The inability to clearly and objectively define the problem, in turn, makes potential solutions to 'wicked' problems highly dependent on the problem heuristic (Kwakkel et al., 2016b). In a way, 'wickedness' and 'wide, deep and murky uncertainty' are two sides of the same coin, 'wickedness' being described through the lens of problem framing, 'wide, deep, and murky' uncertainty being described through the lens of systems modelling.

In climate risk assessment and water system modelling, considering ‘wickedness’ and ‘wide, deep, and murky’ uncertainty is not new. Nor is considering system complexity. Therefore, we find ‘wide, deep, and murky’ uncertainty useful for framing and communicating climate risk assessment and model development. Clearly framing these concepts to both technical and non-technical stakeholders is important, especially when considering the pivotal role of risk assessment and modelling in solving ‘wicked’ problems (Table 7-1 and Figure 7-1).

Table 7-1: Technical and 'layman' framing of 'wide, deep, and murky uncertainty'

	Technical explanation	Non-technical explanation
Wide uncertainty	Long-term climate variability, aleatory uncertainty, and parameter uncertainty means that the space of future climate risk is irreducibly ‘wide’.	A wide range of things can happen, purely due to the somewhat random nature of climate.
Deep uncertainty	We cannot reasonably assign probabilities to future climate risk factors, such as future socio-economic development and associated greenhouse gas emissions or land use changes	Many things could happen, and we don’t know which of these are more or less likely.
Murky uncertainty	a. There is reasonable evidence that certain processes influences system behaviour and risk, but there is limited understanding or agreement on how to model these processes. b. Social values and goals will shape how we assess and respond to risk, but these values and goals are hard to define.	There are things we don’t know and there are things we can’t define objectively.

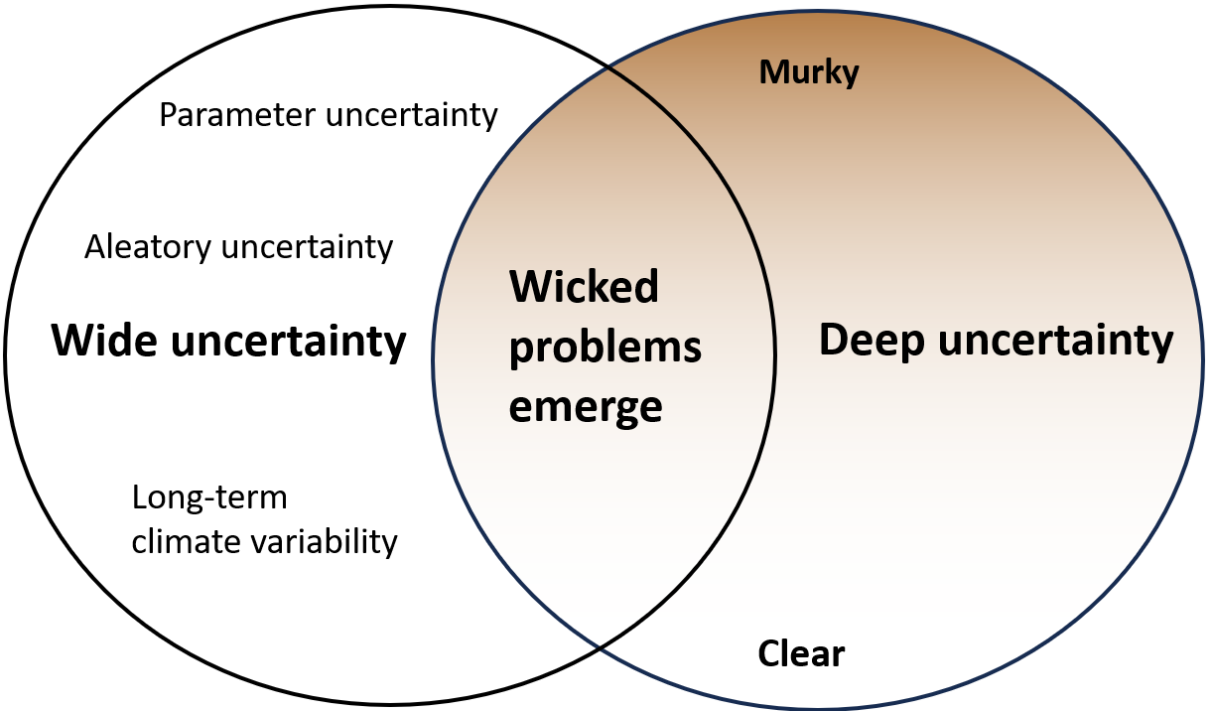

Figure 7-1: Schematic of 'wide, deep, and murky' uncertainty within the context of a 'wicked' problem. Additional descriptors of 'wide' uncertainty have been added to contextualise the key findings from this thesis.

This framing of 'wide, deep, and murky' uncertainty can also be used to contextualise the key findings of this thesis. Viewing these studies together leads to an improved understanding of what 'wide' uncertainty is and how it can be modelled in climate risk assessments. Within a 'wide' uncertainty context, the key findings from this thesis are:

1. Ice core records contain relatively unbiased signals of long-term climate variability. This means that ice cores, which are much longer than instrumental hydrological measurements, can be used to understand, constrain, and model uncertainty arising from long-term, centennial-scale climate variability.
2. A stochastic model calibrated to instrumental measurements cannot simulate long-term, centennial-scale climate variability. This means that traditional stochastic modelling approaches are unable to simulate risk arising from aleatory uncertainty and centennial-scale climate variability.
3. Stochastic model mean and standard deviation are likely (a) non-stationary at multi-centennial and millennial timescales and (b) stationary at centennial timescales.
4. Stochastic model persistence is likely stationary over centennial, multi-centennial, and millennial timescales.

Chapter 6 then ties these four key findings together and presents a stochastic modelling framework that incorporates the three key sources of 'wide' uncertainty in climate risk estimation.

These studies improve understanding of 'wide' uncertainty in climate risk. This is just one component of the various sources of uncertainty to account for when using risk assessment and modelling to solve 'wicked' problems. Further research is needed to better understand and account for these sources of uncertainty. Some broader, philosophical research questions related to these sources of uncertainty were discussed earlier. However, there are also more immediate, follow-up research tasks that can be conducted relatively quickly. These include:

- Stochastically modelling multiple Antarctic ice cores with a common persistence signal.

This would involve calibrating an ARMA(1,1) model to multiple, extended ice core records within a Hierarchical modelling framework. The persistence parameters would be drawn from an underlying hyper-distribution. The calibrated hyper-distribution will, hopefully, contain a robust signal of regional persistence. If successful, the calibrated hyper-distribution can be used in a similar manner to the proxy-prior presented in the Empirical Bayes framework. This means ice core information can be used across the entire mid-latitude Southern Hemisphere.

- Incorporating ice core persistence within scenario-neutral, bottom-up stochastic modelling approaches.

These approaches fall into two broad categories: (1) perturbing stochastic model parameters based on potential climate changes(Guo et al., 2018; McInerney et al., 2023) and (2) using climate data, such as temperature, as a covariate to simulate non-stationarity (Kiem et al., 2021). Both approaches need realistic representations of hydroclimate persistence, which can be derived from ice cores using the methods presented in this thesis. This is particularly important, given (a) the influence of persistence on water system performance (McMahon et al., 2007a; Vogel and Bolognese, 1995) and (b) that climate model simulations underestimate regional persistence (Henley et al., 2017; Rocheta et al., 2014).

- Evaluating if water system performance and potential adaptation options are sensitive to using either 'static baseline' climate inputs or conditional, potentially trending, climate inputs.

There is a key difference between the parameter perturbation approach and the non-stationary, covariate approach used for scenario-neutral, bottom-up stochastic modelling. The perturbation approach produces 'static' timeseries with a stationary long-term mean. The covariate approach can produce transient, trending timeseries (Kiem et al., 2021). Different adaptation approaches may favour static or trending timeseries (Haasnoot et al., 2015). However, it is possible to consider both. Such an important feature of a climate timeseries may influence the resultant climate risk and the robustness of different adaptations (e.g. likely time of failure over the planning horizon - Henley et al. (2013)). Identifying (or not) this sensitivity will guide how stochastic models are used to inform decision making under 'wide, deep, and murky' uncertainty.

These research ideas reflect immediate, practical follow-up work stemming from this thesis. There is a bigger question underpinning these technical proposals – how can we solve 'wicked' problems under 'wide, deep, and murky' uncertainty?

Solving 'wicked' problems under 'wide, deep, and murky' uncertainty is the defining challenge of future water management. Doing so will require, among other things, modelling tools to explore the space of future climate risk. This thesis has, hopefully, improved our understanding of 'wide' uncertainty and demonstrated how it can be modelled statistically.

Chapter 8. References

Ait Brahim, Y., Wassenburg, J.A., Cruz, F.W., Sifeddine, A., Scholz, D., Bouchaou, L., Dassié, E.P., Jochum, K.P., Edwards, R.L., Cheng, H., 2018. Multi-decadal to centennial hydro-climate variability and linkage to solar forcing in the Western Mediterranean during the last 1000 years. Sci Rep 8, 17446. https://doi.org/10.1038/s41598-018-35498-x

Armstrong, M.S., Kiem, A.S., Vance, T.R., 2020. Comparing instrumental, palaeoclimate, and projected rainfall data: Implications for water resources management and hydrological modelling. Journal of Hydrology: Regional Studies 31, 100728. https://doi.org/10.1016/j.ejrh.2020.100728

Ault, T.R., Cole, J.E., Overpeck, J.T., Pederson, G.T., Meko, D.M., 2014. Assessing the Risk of Persistent Drought Using Climate Model Simulations and Paleoclimate Data. Journal of Climate 27, 7529–7549.

Ault, T.R., George, S.S., Smerdon, J.E., Coats, S., Mankin, J.S., Carrillo, C.M., Cook, B.I., Stevenson, S., 2018. A Robust Null Hypothesis for the Potential Causes of Megadrought in Western North America. Journal of Climate 31, 3–24. https://doi.org/10.1175/JCLI-D-17-0154.1

Barnett, J., Evans, L.S., Gross, C., Kiem, A.S., Kingsford, R.T., Palutikof, J.P., Pickering, C.M., Smithers, S.G., 2015. From barriers to limits to climate change adaptation: path dependency and the speed of change. Ecology and Society 20.

Ben-Haim, Y., 2010. Info-Gap Economics. Palgrave Macmillan UK, London. https://doi.org/10.1057/9780230277328

Ben-Haim, Y., 2006. Info-gap decision theory: decisions under severe uncertainty. Elsevier.

Beran, J., 2017. Statistics for long-memory processes. Routledge.

Berghout, B., Henley, B.J., Kuczera, G., 2017. Impact of hydroclimate parameter uncertainty on system yield. Australasian Journal of Water Resources 21, 53–62. https://doi.org/10.1080/13241583.2017.1404550

Betancourt, M., 2018. A Conceptual Introduction to Hamiltonian Monte Carlo. https://doi.org/10.48550/arXiv.1701.02434

Bezerra, B., Veiga, Á., Barroso, L.A., Pereira, M., 2017. Stochastic Long-Term Hydrothermal Scheduling With Parameter Uncertainty in Autoregressive Streamflow Models. IEEE Transactions on Power Systems 32, 999–1006. https://doi.org/10.1109/TPWRS.2016.2572722

Boes, D.C., Salas, J.D., 1978. Nonstationarity of the Mean and the Hurst Phenomenon. Water Resources Research 14, 135–143. https://doi.org/10.1029/WR014i001p00135

Borgomeo, E., Mortazavi-Naeini, M., Hall, J.W., Guillod, B.P., 2018. Risk, Robustness and Water Resources Planning Under Uncertainty. Earth's Future 6, 468–487. https://doi.org/10.1002/2017EF000730

Box, G.E.P., Cox, D.R., 1964. An Analysis of Transformations. Journal of the Royal Statistical Society. Series B (Methodological) 26, 211–252.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 1970. Time series analysis: forecasting and control. John Wiley & Sons.

Brooks, S., 1998. Markov chain Monte Carlo method and its application. Journal of the Royal Statistical Society: Series D (The Statistician) 47, 69–100. https://doi.org/10.1111/1467-9884.00117

Brooks, S.P., Gelman, A., 1998. General Methods for Monitoring Convergence of Iterative Simulations. Journal of Computational and Graphical Statistics 7, 434–455. https://doi.org/10.1080/10618600.1998.10474787

Brown, C., Ghile, Y., Laverty, M., Li, K., 2012. Decision scaling: Linking bottom-up vulnerability analysis with climate projections in the water sector. Water Resources Research 48. https://doi.org/10.1029/2011WR011212

Buckley, B.M., Wilson, R.J.S., Kelly, P.E., Larson, D.W., Cook, E.R., 2004. Inferred summer precipitation for southern Ontario back to AD 610, as reconstructed from ring widths of Thuja occidentalis. Canadian Journal of Forest Research 34, 2541–2553. https://doi.org/10.1139/x04-129

Budner, D., Cole-Dai, J., 2003. The number and magnitude of large explosive volcanic eruptions between 904 and 1865 A.D.: Quantitative evidence from a new South Pole ice core, in: Robock, A., Oppenheimer, C. (Eds.), Geophysical Monograph Series. American Geophysical Union, Washington, D. C., pp. 165–176. https://doi.org/10.1029/139GM10

Buizert, C., Adrian, B., Ahn, J., Albert, M., Alley, R.B., Baggenstos, D., Bauska, T.K., Bay, R.C., Bencivengo, B.B., Bentley, C.R., Brook, E.J., Chellman, N.J., Clow, G.D., Cole-Dai, J., Conway, H., Cravens, E., Cuffey, K.M., Dunbar, N.W., Edwards, J.S., Fegyveresi, J.M., Ferris, D.G., Fitzpatrick, J.J., Fudge, T.J., Gibson, C.J., Gkinis, V., Goetz, J.J., Gregory, S., Hargreaves, G.M., Iverson, N., Johnson, J.A., Jones, T.R., Kalk, M.L., Kippenhan, M.J., Koffman, B.G., Kreutz, K., Kuhl, T.W., Lebar, D.A., Lee, J.E., Marcott, S.A., Markle, B.R., Maselli, O.J., McConnell, J.R., McGwire, K.C., Mitchell, L.E., Mortensen, N.B., Neff, P.D., Nishiizumi, K., Nunn, R.M., Orsi, A.J., Pasteris, D.R., Pedro, J.B., Pettit, E.C., Buford Price, P., Priscu, J.C., Rhodes, R.H., Rosen, J.L., Schauer, A.J., Schoenemann, S.W., Sendelbach, P.J.,

Severinghaus, J.P., Shturmakov, A.J., Sigl, M., Slawny, K.R., Souney, J.M., Sowers, T.A., Spencer, M.K., Steig, E.J., Taylor, K.C., Twickler, M.S., Vaughn, B.H., Voigt, D.E., Waddington, E.D., Welten, K.C., Wendricks, A.W., White, J.W.C., Winstrup, M., Wong, G.J., Woodruff, T.E., WAIS Divide Project Members, 2015. Precise interpolar phasing of abrupt climate change during the last ice age. Nature 520, 661–665. https://doi.org/10.1038/nature14401

Büntgen, U., Allen, K., Anchukaitis, K.J., Arseneault, D., Boucher, É., Bräuning, A., Chatterjee, S., Cherubini, P., Churakova (Sidorova), O.V., Corona, C., Gennaretti, F., Grießinger, J., Guillet, S., Guiot, J., Gunnarson, B., Helama, S., Hochreuther, P., Hughes, M.K., Huybers, P., Kirdyanov, A.V., Krusic, P.J., Ludescher, J., Meier, W.J.-H., Myglan, V.S., Nicolussi, K., Oppenheimer, C., Reinig, F., Salzer, M.W., Seftigen, K., Stine, A.R., Stoffel, M., St. George, S., Tejedor, E., Trevino, A., Trouet, V., Wang, J., Wilson, R., Yang, B., Xu, G., Esper, J., 2021. The influence of decision-making in tree ring-based climate reconstructions. Nature Communications 12, 3411. https://doi.org/10.1038/s41467-021-23627-6

Burt, C.M., Clemmens, A.J., Strelkoff, T.S., Solomon, K.H., Bliesner, R.D., Hardy, L.A., Howell, T.A., Eisenhauer, D.E., 1997. Irrigation Performance Measures: Efficiency and Uniformity. Journal of Irrigation and Drainage Engineering 123, 423–442. https://doi.org/10.1061/(ASCE)0733-9437(1997)123:6(423)

Cahill, N., Croke, J., Campbell, M., Hughes, K., Vitkovsky, J., Kilgallen, J.E., Parnell, A., 2023. A Bayesian time series model for reconstructing hydroclimate from multiple proxies. Environmetrics 34, e2786. https://doi.org/10.1002/env.2786

Cannon, M.J., Percival, D.B., Caccia, D.C., Raymond, G.M., Bassingthwaighte, J.B., 1997. Evaluating scaled windowed variance methods for estimating the Hurst coefficient of time series. Physica A: Statistical Mechanics and its Applications 241, 606–626. https://doi.org/10.1016/S0378-4371(97)00252-5

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A., 2017. Stan : A Probabilistic Programming Language. Journal of Statistical Software 76. https://doi.org/10.18637/jss.v076.i01

Case, R.A., MacDonald, G.M., 2003. TREE RING RECONSTRUCTIONS OF STREAMFLOW FOR THREE CANADIAN PRAIRIE RIVERS1. JAWRA Journal of the American Water Resources Association 39, 703–716. https://doi.org/10.1111/j.1752-1688.2003.tb03686.x

Cook, B.I., Smerdon, J.E., Cook, E.R., Williams, A.P., Anchukaitis, K.J., Mankin, J.S., Allen, K., Andreu-Hayles, L., Ault, T.R., Belmecheri, S., Coats, S., Coulthard, B., Fosu, B., Grierson, P., Griffin, D., Herrera, D.A., Ionita, M., Lehner, F., Leland, C., Marvel, K., Morales, M.S., Mishra, V., Ngoma, J., Nguyen, H.T.T., O'Donnell, A., Palmer, J., Rao, M.P., Rodriguez-Caton, M., Seager, R., Stahle, D.W., Stevenson, S., Thapa, U.K., Varuolo-Clarke, A.M., Wise, E.K., 2022. Megadroughts in the Common Era and the Anthropocene. Nat Rev Earth Environ 3, 741–757. https://doi.org/10.1038/s43017-022-00329-1

Cook, E., 1985. A TIME SERIES ANALYSIS APPROACH TO TREE RING STANDARDIZATION (DENDROCHRONOLOGY, FORESTRY, DENDROCLIMATOLOGY, AUTOREGRESSIVE PROCESS) (Ph.D.). ProQuest Dissertations and Theses. The University of Arizona, Ann Arbor.

Cook, E.R., Briffa, K.R., Meko, D.M., Graybill, D.A., Funkhouser, G., 1995. The "segment length curse" in long tree-ring chronology development for palaeoclimatic studies. The Holocene 5, 229–237. https://doi.org/10.1177/095968369500500211

Cook, E.R., Seager, R., Kushnir, Y., Briffa, K.R., Büntgen, U., Frank, D., Krusic, P.J., Tegel, W., van der Schrier, G., Andreu-Hayles, L., Baillie, M., Baittinger, C., Bleicher, N., Bonde, N., Brown, D., Carrer, M., Cooper, R., Čufar, K., Dittmar, C., Esper, J., Griggs, C., Gunnarson, B., Günther, B., Gutierrez, E., Haneca, K., Helama, S., Herzig, F., Heussner, K.-U., Hofmann, J., Janda, P., Kontic, R., Köse, N., Kyncl, T., Levanič, T., Linderholm, H., Manning, S., Melvin, T.M., Miles, D., Neuwirth, B., Nicolussi, K., Nola, P., Panayotov, M., Popa, I., Rothe, A., Seftigen, K., Seim, A., Svarva, H., Svoboda, M., Thun, T., Timonen, M., Touchan, R., Trotsiuk, V., Trouet, V., Walder, F., Ważny, T., Wilson, R., Zang, C., 2015. Old World megadroughts and pluvials during the Common Era. Science Advances 1, e1500561. https://doi.org/10.1126/sciadv.1500561

Crockart, C.K., Vance, T.R., Fraser, A.D., Abram, N.J., Criscitiello, A.S., Curran, M.A.J., Favier, V., Gallant, A.J.E., Kittel, C., Kjær, H.A., Klekociuk, A.R., Jong, L.M., Moy, A.D., Plummer, C.T., Vallelonga, P.T., Wille, J., Zhang, L., 2021. El Niño--Southern Oscillation signal in a new East Antarctic ice core, Mount Brown South. Climate of the Past 17, 1795–1818. https://doi.org/10.5194/cp-17-1795-2021

Croke, J., Vítkovský, J., Hughes, K., Campbell, M., Amirnezhad-Mozhdehi, S., Parnell, A., Cahill, N., Dalla Pozza, R., 2021. A palaeoclimate proxy database for water security planning in Queensland Australia. Sci Data 8, 292. https://doi.org/10.1038/s41597-021-01074-8

Culley, S., Noble, S., Yates, A., Timbs, M., Westra, S., Maier, H.R., Giuliani, M., Castelletti, A., 2016. A bottom-up approach to identifying the maximum operational adaptive capacity of water resource systems to a changing climate. Water Resources Research 52, 6751–6768. https://doi.org/10.1002/2015WR018253

Curran, M.A.J., van Ommen, T.D., Morgan, V.I., Phillips, K.L., Palmer, A.S., 2003. Ice Core Evidence for Antarctic Sea Ice Decline Since the 1950s. Science 302, 1203–1206. https://doi.org/10.1126/science.1087888

Dansgaard, W., Johnsen, S.J., 1969. A Flow Model and a Time Scale for the Ice Core from Camp Century, Greenland. Journal of Glaciology 8, 215–223. https://doi.org/10.3189/S0022143000031208

D'Arrigo, R., Wilson, R., Liepert, B., Cherubini, P., 2008. On the 'Divergence Problem' in Northern Forests: A review of the tree-ring evidence and possible causes. Global and Planetary Change 60, 289–305. https://doi.org/10.1016/j.gloplacha.2007.03.004

Dätwyler, C., Grosjean, M., Steiger, N.J., Neukom, R., 2020. Teleconnections and relationship between the El Niño–Southern Oscillation (ENSO) and the Southern Annular Mode (SAM) in reconstructions and models over the past millennium. Climate of the Past 16, 743–756.

Davini, P., von Hardenberg, J., Corti, S., Christensen, H.M., Juricke, S., Subramanian, A., Watson, P.A.G., Weisheimer, A., Palmer, T.N., 2017. Climate SPHINX: evaluating the impact of resolution and stochastic physics parameterisations in the EC-Earth global climate model. Geoscientific Model Development 10, 1383–1402. https://doi.org/10.5194/gmd-10-1383-2017

Dee, S.G., Steiger, N.J., Emile-Geay, J., Hakim, G.J., 2016. On the utility of proxy system models for estimating climate states over the common era. Journal of Advances in Modeling Earth Systems 8, 1164–1179. https://doi.org/10.1002/2016MS000677

DeRose, R.J., Bekker, M.F., Wang, S.-Y., Buckley, B.M., Kjelgren, R.K., Bardsley, T., Rittenour, T.M., Allen, E.B., 2015. A millennium-length reconstruction of Bear River stream flow, Utah. Journal of Hydrology 529, 524–534. https://doi.org/10.1016/j.jhydrol.2015.01.014

Deser, C., Lehner, F., Rodgers, K.B., Ault, T., Delworth, T.L., DiNezio, P.N., Fiore, A., Frankignoul, C., Fyfe, J.C., Horton, D.E., Kay, J.E., Knutti, R., Lovenduski, N.S., Marotzke, J., McKinnon, K.A., Minobe, S., Randerson, J., Screen, J.A., Simpson, I.R., Ting, M., 2020. Insights from Earth system model initial-condition large ensembles and future prospects. Nature Climate Change 10, 277–286. https://doi.org/10.1038/s41558-020-0731-2

Dimitriadis, P., Koutsoyiannis, D., 2015. Climacogram versus autocovariance and power spectrum in stochastic modelling for Markovian and Hurst–Kolmogorov processes. Stoch Environ Res Risk Assess 29, 1649–1669. https://doi.org/10.1007/s00477-015-1023-7

Dixon, B.C., Tyler, J.J., Lorrey, A.M., Goodwin, I.D., Gergis, J., Drysdale, R.N., 2017. Low-resolution Australasian palaeoclimate records of the last 2000 years. Climate Of The Past 13, 1403–1433. https://doi.org/10.5194/cp-13-1403-2017

Domorenok, E., Graziano, P., Polverari, L., 2021. Introduction: policy integration and institutional capacity: theoretical, conceptual and empirical challenges. Policy and Society 40, 1–18. https://doi.org/10.1080/14494035.2021.1902058

Durre, I., Menne, M.J., Gleason, B.E., Houston, T.G., Vose, R.S., 2010. Comprehensive Automated Quality Assurance of Daily Surface Observations. Journal of Applied Meteorology and Climatology 49, 1615–1633. https://doi.org/10.1175/2010JAMC2375.1

Emile-Geay, J., McKay, N.P., Kaufman, D.S., von Gunten, L., Wang, J., Anchukaitis, K.J., Abram, N.J., Addison, J.A., Curran, M.A.J., Evans, M.N., Henley, B.J., Hao, Z., Martrat, B., McGregor, H.V., Neukom, R., Pederson, G.T., Stenni, B., Thirumalai, K., Werner, J.P., Xu, C., Divine, D.V., Dixon, B.C., Gergis, J., Mundo, I.A., Nakatsuka, T., Phipps, S.J., Routson, C.C., Steig, E.J., Tierney, J.E., Tyler, J.J., Allen, K.J., Bertler, N.A.N., Björklund, J., Chase, B.M., Chen, M.-T., Cook, E., de Jong, R., DeLong, K.L., Dixon, D.A., Ekaykin, A.A., Ersek, V., Filipsson, H.L., Francus, P., Freund, M.B., Frezzotti, M., Gaire, N.P., Gajewski, K., Ge, Q., Goosse, H., Gornostaeva, A., Grosjean, M., Horiuchi, K., Hormes, A., Husum, K., Isaksson, E., Kandasamy, S., Kawamura, K., Kilbourne, K.H., Koç, N., Leduc, G., Linderholm, H.W., Lorrey, A.M., Mikhalenko, V., Mortyn, P.G., Motoyama, H., Moy, A.D., Mulvaney, R., Munz, P.M., Nash, D.J., Oerter, H., Opel, T., Orsi, A.J., Ovchinnikov, D.V., Porter, T.J., Roop, H.A., Saenger, C., Sano, M., Sauchyn, D., Saunders, K.M., Seidenkrantz, M.-S., Severi, M., Shao, X., Sicre, M.-A., Sigl, M., Sinclair, K., St. George, S., St. Jacques, J.-M., Thamban, M., Kuwar Thapa, U., Thomas, E.R., Turney, C., Uemura, R., Viau, A.E., Vladimirova, D.O., Wahl, E.R., White, J.W.C., Yu, Z., Zinke, J., PAGES2k Consortium, 2017. A global multiproxy database for temperature reconstructions of the Common Era. Scientific Data 4, 170088. https://doi.org/10.1038/sdata.2017.88

Engle, R.F., 1982. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. Econometrica 50, 987–1007. https://doi.org/10.2307/1912773

Engle, R.F., Bollerslev, T., 1986. Modelling the persistence of conditional variances. null 5, 1–50. https://doi.org/10.1080/07474938608800095

Erkyihun, S.T., Rajagopalan, B., Zagona, E., Lall, U., Nowak, K., 2016. Wavelet-based time series bootstrap model for multidecadal streamflow simulation using climate indicators. Water Resources Research 52, 4061–4077. https://doi.org/10.1002/2016WR018696

Esper, J., Frank, D., Büntgen, U., Verstege, A., Luterbacher, J., Xoplaki, E., 2007. Long-term drought severity variations in Morocco. Geophysical Research Letters 34. https://doi.org/10.1029/2007GL030844
Evans, J.P., Ekström, M., Ji, F., 2012. Evaluating the performance of a WRF physics ensemble over South-East Australia. Clim Dyn 39, 1241–1258. https://doi.org/10.1007/s00382-011-1244-5
Falkenmark, M., 1977. Water and Mankind: A Complex System of Mutual Interaction. Ambio 6, 3–9.
Fatichi, S., Ivanov, V.Y., Caporali, E., 2012. Investigating Interannual Variability of Precipitation at the Global Scale: Is There a Connection with Seasonality? Journal of Climate 25, 5512–5523. https://doi.org/10.1175/JCLI-D-11-00356.1
Fernández, A., Muñoz, A., González-Reyes, Á., Aguilera-Betti, I., Toledo, I., Puchi, P., Sauchyn, D., Crespo, S., Frene, C., Mundo, I., González, M., Vignola, R., 2018. Dendrohydrology and water resources management in south-central Chile: lessons from the Río Imperial streamflow reconstruction. Hydrology and Earth System Sciences 22, 2921–2935. https://doi.org/10.5194/hess-22-2921-2018
Fernández, F.J., Vásquez-Lavín, F., Ponce, R.D., Garreaud, R., Hernández, F., Link, O., Zambrano, F., Hanemann, M., 2023. The economics impacts of long-run droughts: Challenges, gaps, and way forward. Journal of Environmental Management 344, 118726. https://doi.org/10.1016/j.jenvman.2023.118726
Ferris, D.G., Cole-Dai, J., Reyes, A.R., Budner, D.M., 2011. South Pole ice core record of explosive volcanic eruptions in the first and second millennia A.D. and evidence of a large eruption in the tropics around 535 A.D. Journal of Geophysical Research: Atmospheres 116. https://doi.org/10.1029/2011JD015916
Fiering, M.B., 2013. Streamflow Synthesis, in: Streamflow Synthesis. Harvard University Press. https://doi.org/10.4159/harvard.9780674189287
Flack, A.L., Kiem, A.S., Vance, T.R., Tozer, C.R., Roberts, J.L., 2020. Comparison of published palaeoclimate records suitable for reconstructing annual to sub-decadal hydroclimatic variability in eastern Australia: implications for water resource management and planning. Hydrology and Earth System Sciences Discussions 2020, 1–25. https://doi.org/10.5194/hess-2020-314
Fowler, K., Ballis, N., Horne, A., John, A., Nathan, R., Peel, M., 2022. Integrated framework for rapid climate stress testing on a monthly timestep. Environmental Modelling & Software 150, 105339. https://doi.org/10.1016/j.envsoft.2022.105339
Fraley, C., Leisch, F., Maechler, M., Reisen, V., Lemonte, A., 2012. fracdiff: Fractionally differenced ARIMA aka ARFIMA (p, d, q) models. R package version 1.
Franke, J., Frank, D., Raible, C.C., Esper, J., Brönnimann, S., 2013a. Spectral biases in tree-ring climate proxies. Nature Climate Change 3, 360–364. https://doi.org/10.1038/nclimate1816
Franke, J., Frank, D., Raible, C.C., Esper, J., Brönnimann, S., 2013b. Spectral biases in tree-ring climate proxies. Nature Climate Change 3, 360–364. https://doi.org/10.1038/nclimate1816
Frost, A.J., Thyer, M.A., Srikanthan, R., Kuczera, G., 2007. A general Bayesian framework for calibrating and evaluating stochastic models of annual multi-site hydrological data. Journal of Hydrology 340, 129–148. https://doi.org/10.1016/j.jhydrol.2007.03.023
Galelli, S., Nguyen, H.T.T., Turner, S.W.D., Buckley, B.M., 2021. Time to Use Dendrohydrological Data in Water Resources Management? Journal of Water Resources Planning and Management 147, 01821001. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001422
Gangopadhyay, S., Harding, B.L., Rajagopalan, B., Lukas, J.J., Fulp, T.J., 2009. A nonparametric approach for paleohydrologic reconstruction of annual streamflow ensembles. Water Resources Research 45. https://doi.org/10.1029/2008WR007201
Geay, J.E., Tingley, M., 2016. Inferring climate variability from nonlinear proxies:application to palaeo-ENSO studies. Climate of the Past 12, 31–50.
Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis, Third Edition. CRC Press.
Gelman, A., Rubin, D.B., 1992. Inference from Iterative Simulation Using Multiple Sequences. Statistical Science 7, 457–472. https://doi.org/10.1214/ss/1177011136
Gershman, S.J., Blei, D.M., 2012. A tutorial on Bayesian nonparametric models. Journal of Mathematical Psychology 56, 1–12. https://doi.org/10.1016/j.jmp.2011.08.004
Geweke, J., Porter-Hudak, S., 1983. The estimation and application of long memory time series models. Journal of Time Series Analysis 4, 221–238. https://doi.org/10.1111/j.1467-9892.1983.tb00371.x
Gober, P., Sampson, D.A., Quay, R., White, D.D., Chow, W.T.L., 2016. Urban adaptation to mega-drought: Anticipatory water modeling, policy, and planning for the urban Southwest. Sustainable Cities and Society 27, 497–504. https://doi.org/10.1016/j.scs.2016.05.001
Goodwin, M.J., Verdon-Kidd, D.C., Hua, Q., English, N.B., Haines, H.A., Allen, K.J., 2022. Hydroclimate proxies for eastern Australia using stable isotopes in grey mangroves (Avicennia marina). Global and Planetary Change 208, 103691. https://doi.org/10.1016/j.gloplacha.2021.103691

Grafton, R.Q., Williams, J., Perry, C.J., Molle, F., Ringler, C., Steduto, P., Udall, B., Wheeler, S.A., Wang, Y., Garrick, D., Allen, R.G., 2018. The paradox of irrigation efficiency. Science 361, 748–750. https://doi.org/10.1126/science.aat9314

Granger, C.W.J., Joyeux, R., 1980. AN INTRODUCTION TO LONG-MEMORY TIME SERIES MODELS AND FRACTIONAL DIFFERENCING. Journal of Time Series Analysis 1, 15–29. https://doi.org/10.1111/j.1467-9892.1980.tb00297.x

Graves, T., Gramacy, R., Watkins, N., Franzke, C., 2017. A Brief History of Long Memory: Hurst, Mandelbrot and the Road to ARFIMA, 1951-1980. Entropy 19, 437.

Griffin, D., Anchukaitis, K.J., 2014. How unusual is the 2012–2014 California drought? Geophysical Research Letters 41, 9017–9023. https://doi.org/10.1002/2014GL062433

Grimm, A.M., Barros, V.R., Doyle, M.E., 2000. Climate Variability in Southern South America Associated with El Niño and La Niña Events. Journal of Climate 13, 35–58. https://doi.org/10.1175/1520-0442(2000)013<0035:CVISSA>2.0.CO;2

Grose, M., Bhend, J., Argueso, D., Ekstrom, M., Dowdy, A., Hoffman, P., Evans, J., Timbal, B., 2015. Comparison of various climate change projections of eastern Australian rainfall. Australian Meteorological and Oceanographic Journal 65, 90–106.

Grose, M.R., Narsey, S., Delage, F.P., Dowdy, A.J., Bador, M., Boschat, G., Chung, C., Kajtar, J.B., Rauniyar, S., Freund, M.B., Lyu, K., Rashid, H., Zhang, X., Wales, S., Trenham, C., Holbrook, N.J., Cowan, T., Alexander, L., Arblaster, J.M., Power, S., 2020. Insights From CMIP6 for Australia's Future Climate. Earth's Future 8, e2019EF001469. https://doi.org/10.1029/2019EF001469

Guo, D., Westra, S., Maier, H.R., 2018. An inverse approach to perturb historical rainfall data for scenario-neutral climate impact studies. Journal of Hydrology 556, 877–890. https://doi.org/10.1016/j.jhydrol.2016.03.025

Haasnoot, M., Kwakkel, J.H., Walker, W.E., ter Maat, J., 2013. Dynamic adaptive policy pathways: A method for crafting robust decisions for a deeply uncertain world. Global Environmental Change 23, 485–498. https://doi.org/10.1016/j.gloenvcha.2012.12.006

Haasnoot, M., Schellekens, J., Beersma, J.J., Middelkoop, H., Kwadijk, J.C.J., 2015. Transient scenarios for robust climate change adaptation illustrated for water management in The Netherlands. Environ. Res. Lett. 10, 105008. https://doi.org/10.1088/1748-9326/10/10/105008

Haasnoot, M., van Aalst, M., Rozenberg, J., Dominique, K., Matthews, J., Bouwer, L.M., Kind, J., Poff, N.L., 2020. Investments under non-stationarity: economic evaluation of adaptation pathways. Climatic Change 161, 451–463. https://doi.org/10.1007/s10584-019-02409-6

Haasnoot, M., van 't Klooster, S., van Alphen, J., 2018. Designing a monitoring system to detect signals to adapt to uncertain climate change. Global Environmental Change 52, 273–285. https://doi.org/10.1016/j.gloenvcha.2018.08.003

Hall, J.W., Lempert, R.J., Keller, K., Hackbarth, A., Mijere, C., McInerney, D.J., 2012. Robust Climate Policies Under Uncertainty: A Comparison of Robust Decision Making and Info-Gap Methods. Risk Analysis 32, 1657–1672. https://doi.org/10.1111/j.1539-6924.2012.01802.x

Hallegatte, S., Shah, A., Lempert, R., Brown, C., Gill, S., 2012. Investment Decision Making under Deep Uncertainty - Application to Climate Change, Policy Research Working Papers. The World Bank. https://doi.org/10.1596/1813-9450-6193

Hamed, K.H., 2009. Enhancing the effectiveness of prewhitening in trend analysis of hydrologic data. Journal of Hydrology 368, 143–155. https://doi.org/10.1016/j.jhydrol.2009.01.040

Hamed, K.H., 2007. Improved finite-sample Hurst exponent estimates using rescaled range analysis. Water Resources Research 43. https://doi.org/10.1029/2006WR005111

Haslett, J., Raftery, A.E., 1989. Space-Time Modelling with Long-Memory Dependence: Assessing Ireland's Wind Power Resource. Journal of the Royal Statistical Society: Series C (Applied Statistics) 38, 1–21. https://doi.org/10.2307/2347679

Helama, S., Meirläinen, J., Tuomenvirta, H., 2009. Multicentennial megadrought in northern Europe coincided with a global El Nino-Southern Oscillation drought pattern during the Medieval Climate Anomaly. Geology 37, 175–178. https://doi.org/10.1130/G25329A.1

Henley, B.J., Meehl, G., Power, S.B., Folland, C.K., King, A.D., Brown, J.N., Karoly, D.J., Delage, F., Gallant, A.J.E., Freund, M., Neukom, R., 2017. Spatial and temporal agreement in climate model simulations of the Interdecadal Pacific Oscillation. Environmental Research Letters 12, 44011. https://doi.org/10.1088/1748-9326/aa5cc8

Henley, B.J., Thyer, M.A., Kuczera, G., 2013. Climate driver informed short-term drought risk evaluation. Water Resources Research 49, 2317–2326. https://doi.org/10.1002/wrcr.20222

Henley, B.J., Thyer, M.A., Kuczera, G., Franks, S.W., 2011. Climate-informed stochastic hydrological modeling: Incorporating decadal-scale variability using paleo data. Water Resources Research 47. https://doi.org/10.1029/2010WR010034

Hessl, A.E., Anchukaitis, K.J., Jelsema, C., Cook, B., Byambasuren, O., Leland, C., Nachin, B., Pederson, N., Tian, H., Hayles, L.A., 2018. Past and future drought in Mongolia. Science Advances 4. https://doi.org/10.1126/sciadv.1701832

Higgins, P.A., Palmer, J.G., Andersen, M.S., Turney, C.S.M., Johnson, F., 2023. Extreme events in the multi-proxy South Pacific drought atlas. Climatic Change 176, 105. https://doi.org/10.1007/s10584-023-03585-2

Ho, M., Kiem, A.S., Verdon-Kidd, D.C., 2015a. A paleoclimate rainfall reconstruction in the Murray-Darling Basin (MDB), Australia: 2. Assessing hydroclimatic risk using paleoclimate records of wet and dry epochs. Water Resources Research 51, 8380–8396. https://doi.org/10.1002/2015WR017059

Ho, M., Kiem, A.S., Verdon-Kidd, D.C., 2015b. A paleoclimate rainfall reconstruction in the Murray-Darling Basin (MDB), Australia: 1. Evaluation of different paleoclimate archives, rainfall networks, and reconstruction techniques. Water Resources Research 51, 8362–8379. https://doi.org/10.1002/2015WR017058

Homan, M.D., Gelman, A., 2014. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. J. Mach. Learn. Res. 15, 1593–1623.

Hong, S.-Y., Noh, Y., Dudhia, J., 2006. A New Vertical Diffusion Package with an Explicit Treatment of Entrainment Processes. Monthly Weather Review 134, 2318–2341. https://doi.org/10.1175/MWR3199.1

Hosking, J.R.M., 1984. Modeling persistence in hydrological time series using fractional differencing. Water Resources Research 20, 1898–1908. https://doi.org/10.1029/WR020i012p01898

Hu, J., Emile-Geay, J., Partin, J., 2017. Correlation-based interpretations of paleoclimate data – where statistics meet past climates. Earth and Planetary Science Letters 459, 362–371. https://doi.org/10.1016/j.epsl.2016.11.048

Huang, Y., Wang, Y., Xue, L., Wei, X., Zhang, L., Li, H., 2020. Comparison of three microphysics parameterization schemes in the WRF model for an extreme rainfall event in the coastal metropolitan City of Guangzhou, China. Atmospheric Research 240, 104939. https://doi.org/10.1016/j.atmosres.2020.104939

Hughes, J.P., Guttorp, P., Charles, S.P., 1999. A non-homogeneous hidden Markov model for precipitation occurrence. Journal of the Royal Statistical Society: Series C (Applied Statistics) 48, 15–30. https://doi.org/10.1111/1467-9876.00136

Hurlimann, A., Dolnicar, S., 2010. When public opposition defeats alternative water projects – The case of Toowoomba Australia. Water Research 44, 287–297. https://doi.org/10.1016/j.watres.2009.09.020

Hurst, H., 1951. Long-Term Storage Capacity of Reservoirs. Transactions of the American Society of Civil Engineers 116, 770–799. https://doi.org/10.1061/TACEAT.0006518

Huybrechts, P., Rybak, O., Pattyn, F., Ruth, U., Steinhage, D., 2007. Ice thinning, upstream advection, and non-climatic biases for the upper 89% of the EDML ice core from a nested model of the Antarctic ice sheet. Climate of the Past 3, 577–589. https://doi.org/10.5194/cp-3-577-2007

Hyndman, R.J., Khandakar, Y., 2007. Automatic time series forecasting: the forecast package for R. Journal of Statistical Software. https://doi.org/DOI:

Iliopoulou, T., Papalexiou, S.M., Markonis, Y., Koutsoyiannis, D., 2018. Revisiting long-range dependence in annual precipitation. Journal of Hydrology 556, 891–900. https://doi.org/10.1016/j.jhydrol.2016.04.015

Ilyas, A., Manzoor, T., Muhammad, A., 2021. A Dynamic Socio-Hydrological Model of the Irrigation Efficiency Paradox. Water Resources Research 57, e2021WR029783. https://doi.org/10.1029/2021WR029783

Jackson, S.L., Vance, T.R., Crockart, C., Moy, A., Plummer, C., Abram, N.J., 2023. Climatology of the Mount Brown South ice core site in East Antarctica: implications for the interpretation of a water isotope record. Climate of the Past 19, 1653–1675. https://doi.org/10.5194/cp-19-1653-2023

Jeffrey, S.J., Carter, J.O., Moodie, K.B., Beswick, A.R., 2001. Using spatial interpolation to construct a comprehensive archive of Australian climate data. Environmental Modelling & Software 16, 309–330. https://doi.org/10.1016/S1364-8152(01)00008-1

Jones, D., Wang, W., Fawcett, R., 2009. High-quality spatial climate data-sets for Australia. Australian Meteorological and Oceanographic Journal 58, 233.

Jones, G.L., Qin, Q., 2022. Markov Chain Monte Carlo in Practice. Annual Review of Statistics and Its Application 9, 557–578. https://doi.org/10.1146/annurev-statistics-040220-090158

Jones, P., Briffa, K., Osborn, T., Lough, J., van Ommen, T., Vinther, B., Luterbacher, J., Wahl, E., Zwiers, F., Mann, M., Schmidt, G., Ammann, C., Buckley, B., Cobb, K., Esper, J., Goosse, H., Graham, N., Jansen, E., Kiefer, T., Kull, C., Küttel, M., Mosley-Thompson, E., Overpeck, J., Riedwyl, N., Schulz, M., Tudhope, A., Villalba, R., Wanner, H., Wolff, E., Xoplaki, E., 2009. High-resolution palaeoclimatology of the last millennium: a review of current status and future prospects. The Holocene 19, 3–49. https://doi.org/10.1177/0959683608098952

Jong, L.M., Plummer, C.T., Roberts, J.L., Moy, A.D., Curran, M.A.J., Vance, T.R., Pedro, J., Long, C., Nation, M., Mayewski, P.A., van Ommen, T.D., 2022. 2000 years of annual ice core data from Law Dome, East Antarctica. Earth System Science Data Discussions 2022, 1–26. https://doi.org/10.5194/essd-2021-408

Kavetski, D., Kuczera, G., Franks, S.W., 2006. Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. Water Resources Research 42. https://doi.org/10.1029/2005WR004368

Kendziorski, C.M., Bassingthwaighte, J.B., Tonellato, P.J., 1999. Evaluating maximum likelihood estimation methods to determine the Hurst coefficient. Physica A: Statistical Mechanics and its Applications 273, 439–451. https://doi.org/10.1016/S0378-4371(99)00268-X

Kiem, A.S., Franks, S.W., 2004. Multi-decadal variability of drought risk, eastern Australia. Hydrological Processes 18, 2039–2050. https://doi.org/10.1002/hyp.1460

Kiem, A.S., Franks, S.W., Kuczera, G., 2003. Multi-decadal variability of flood risk. Geophysical Research Letters 30. https://doi.org/10.1029/2002GL015992

Kiem, A.S., Kuczera, G., Kozarovski, P., Zhang, L., Willgoose, G., 2021. Stochastic Generation of Future Hydroclimate Using Temperature as a Climate Change Covariate. Water Resources Research 57, 2020WR027331. https://doi.org/10.1029/2020WR027331

Kiem, A.S., Vance, T.R., Tozer, C.R., Roberts, J.L., Pozza, R., Vitkovsky, J., Smolders, K., Curran, M.A.J., 2020. Learning from the past – Using palaeoclimate data to better understand and manage drought in South East Queensland (SEQ), Australia. Journal of Hydrology: Regional Studies 29, 100686. https://doi.org/10.1016/j.ejrh.2020.100686

Kim, H., Park, J., Yoo, J., Kim, T.-W., 2015. Assessment of drought hazard, vulnerability, and risk: A case study for administrative districts in South Korea. Journal of Hydro-environment Research 9, 28–35. https://doi.org/10.1016/j.jher.2013.07.003

Klemeš, V., 1989. The improbable probabilities of extreme floods and droughts, in: Hydrology of Disasters: Proceedings of the World Meteorological Organization Technical Conference Held in Geneva, November 1988. Routledge, pp. 43–51.

Klippel, L., Krusic, P.J., Brandes, R., Hartl, C., Belmecheri, S., Dienst, M., Esper, J., 2018. A 1286-year hydro-climate reconstruction for the Balkan Peninsula. Boreas 47, 1218–1229. https://doi.org/10.1111/bor.12320

Knief, U., Forstmeier, W., 2021. Violating the normality assumption may be the lesser of two evils. Behav Res 53, 2576–2590. https://doi.org/10.3758/s13428-021-01587-5

Knight, T.A., Meko, D.M., Baisan, C.H., 2010. A Bimillennial-Length Tree-Ring Reconstruction of Precipitation for the Tavaputs Plateau, Northeastern Utah. Quaternary Research 73, 107–117. https://doi.org/DOI: 10.1016/j.yqres.2009.08.002

Koutsoyiannis, D., 2011. Hurst–Kolmogorov dynamics as a result of extremal entropy production. Physica A: Statistical Mechanics and its Applications 390, 1424–1432. https://doi.org/10.1016/j.physa.2010.12.035

Koutsoyiannis, D., 2010. HESS Opinions "A random walk on water." Hydrology and Earth System Sciences 14, 585–601. https://doi.org/10.5194/hess-14-585-2010

Koutsoyiannis, D., 2006. Nonstationarity versus scaling in hydrology. Journal of Hydrology 324, 239–254. https://doi.org/10.1016/j.jhydrol.2005.09.022

Koutsoyiannis, D., 2005. Uncertainty, entropy, scaling and hydrological stochastics. 2. Time dependence of hydrological processes and time scaling / Incertitude, entropie, effet d'échelle et propriétés stochastiques hydrologiques. 2. Dépendance temporelle des processus hydrologiques et échelle temporelle. Hydrological Sciences Journal 50, null-426. https://doi.org/10.1623/hysj.50.3.405.65028

Koutsoyiannis, D., 2003. Climate change, the Hurst phenomenon, and hydrological statistics. Hydrological Sciences Journal 48, 3–24. https://doi.org/10.1623/hysj.48.1.3.43481

Koutsoyiannis, D., 2002. The Hurst phenomenon and fractional Gaussian noise made easy. Hydrological Sciences Journal 47, 573–595. https://doi.org/10.1080/02626660209492961

Koutsoyiannis, D., 2000. A generalized mathematical framework for stochastic simulation and forecast of hydrologic time series. Water Resources Research 36, 1519–1533. https://doi.org/10.1029/2000WR900044

Koutsoyiannis, D., Montanari, A., 2015. Negligent killing of scientific concepts: the stationarity case. Hydrological Sciences Journal 60, 1174–1183. https://doi.org/10.1080/02626667.2014.959959

Koutsoyiannis, D., Montanari, A., 2007. Statistical analysis of hydroclimatic time series: Uncertainty and insights. Water Resources Research 43. https://doi.org/10.1029/2006WR005592

Kuczera, G., 1992. Water supply headworks simulation using network linear programming. Advances in Engineering Software 14, 55–60. https://doi.org/10.1016/0965-9978(92)90084-S

Kwakkel, J.H., Haasnoot, M., Walker, W.E., 2016a. Comparing Robust Decision-Making and Dynamic Adaptive Policy Pathways for model-based decision support under deep uncertainty. Environmental Modelling & Software 86, 168–183. https://doi.org/10.1016/j.envsoft.2016.09.017

Kwakkel, J.H., Walker, W.E., Haasnoot, M., 2016b. Coping with the Wickedness of Public Policy Problems: Approaches for Decision Making under Deep Uncertainty. J. Water Resour. Plann. Manage. 142, 01816001. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000626

Kwon, H.-H., Lall, U., Khalil, A.F., 2007. Stochastic simulation model for nonstationary time series using an autoregressive wavelet decomposition: Applications to rainfall and temperature. Water Resources Research 43. https://doi.org/10.1029/2006WR005258

Lall, U., Sharma, A., 1996. A Nearest Neighbor Bootstrap For Resampling Hydrologic Time Series. Water Resources Research 32, 679–693. https://doi.org/10.1029/95WR02966

Lambert, F., Delmonte, B., Petit, J.R., Bigler, M., Kaufmann, P.R., Hutterli, M.A., Stocker, T.F., Ruth, U., Steffensen, J.P., Maggi, V., 2008. Dust-climate couplings over the past 800,000 years from the EPICA Dome C ice core. Nature 452, 616–619. https://doi.org/10.1038/nature06763

Lambert, M.F., Whiting, J.P., Metcalfe, A.V., 2003. A non-parametric hidden Markov model for climate state identification. Hydrology and Earth System Sciences 7, 652–667.

Lee, T., Ouarda, T.B.M.J., 2012. Stochastic simulation of nonstationary oscillation hydroclimatic processes using empirical mode decomposition. Water Resources Research 48. https://doi.org/10.1029/2011WR010660

Lempert, R.J., Groves, D.G., Popper, S.W., Bankes, S.C., 2006. A General, Analytic Method for Generating Robust Strategies and Narrative Scenarios. Management Science 52, 514–528. https://doi.org/10.1287/mnsc.1050.0472

Lim, T.C., Glynn, P.D., Shenk, G.W., Bitterman, P., Guillaume, J.H.A., Little, J.C., Webster, D.G., 2023. Recognizing political influences in participatory social-ecological systems modeling. Socio-Environmental Systems Modelling 5, 18509–18509. https://doi.org/10.18174/sesmo.18509

Ljungqvist, F.C., Krusic, P.J., Sundqvist, H.S., Zorita, E., Brattström, G., Frank, D., 2016. Northern Hemisphere hydroclimate variability over the past twelve centuries. Nature 532, 94–98. https://doi.org/10.1038/nature17418

Lorenz, E.N., 1969. Atmospheric Predictability as Revealed by Naturally Occurring Analogues. Journal of the Atmospheric Sciences 26, 636–646. https://doi.org/10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2

Loucks, D.P., Van Beek, E., 2017. Water resource systems planning and management: An introduction to methods, models, and applications. Springer.

Ludescher, J., Bunde, A., Büntgen, U., Schellnhuber, H.J., 2020. Setting the tree-ring record straight. Climate Dynamics 55, 3017–3024. https://doi.org/10.1007/s00382-020-05433-w

Lux, T., 1998. The socio-economic dynamics of speculative markets: interacting agents, chaos, and the fat tails of return distributions. Journal of Economic Behavior & Organization 33, 143–165. https://doi.org/10.1016/S0167-2681(97)00088-7

MacDonald, G.M., 2007. Severe and sustained drought in southern California and the West: Present conditions and insights from the past on causes and impacts. Quaternary International 173–174, 87–100. https://doi.org/10.1016/j.quaint.2007.03.012

MacDonald, G.M., Kremenetski, K.V., Hidalgo, H.G., 2008. Southern California and the perfect drought: Simultaneous prolonged drought in southern California and the Sacramento and Colorado River systems. Quaternary International, The 22nd Pacific Climate Workshop 188, 11–23. https://doi.org/10.1016/j.quaint.2007.06.027

Macias-Fauria, M., Grinsted, A., Helama, S., Holopainen, J., 2012. Persistence matters: Estimation of the statistical significance of paleoclimatic reconstruction statistics from autocorrelated time series. Dendrochronologia 30, 179–187. https://doi.org/10.1016/j.dendro.2011.08.003

Maier, H.R., Guillaume, J.H.A., van Delden, H., Riddell, G.A., Haasnoot, M., Kwakkel, J.H., 2016. An uncertain future, deep uncertainty, scenarios, robustness and adaptation: How do they fit together? Environmental Modelling and Software Software 81, 154–164. https://doi.org/10.1016/j.envsoft.2016.03.014

Malevich, S.B., Woodhouse, C.A., Meko, D.M., 2013. Tree-ring reconstructed hydroclimate of the Upper Klamath basin. Journal of Hydrology 495, 13–22. https://doi.org/10.1016/j.jhydrol.2013.04.048

Mandelbrot, B.B., 1971. A Fast Fractional Gaussian Noise Generator. Water Resources Research 7, 543–553. https://doi.org/10.1029/WR007i003p00543

Mandelbrot, B.B., Wallis, J.R., 1969. Robustness of the rescaled range R/S in the measurement of noncyclic long run statistical dependence. Water Resources Research 5, 967–988. https://doi.org/10.1029/WR005i005p00967

Mann, M.E., Steinman, B.A., Brouillette, D.J., Miller, S.K., 2021. Multidecadal climate oscillations during the past millennium driven by volcanic forcing. Science 371, 1014–1019. https://doi.org/10.1126/science.abc5810

Mann, M.E., Steinman, B.A., Miller, S.K., 2020. Absence of internal multidecadal and interdecadal oscillations in climate model simulations. Nature Communications 11, 49. https://doi.org/10.1038/s41467-019-13823-w

Markle, B.R., Steig, E.J., Roe, G.H., Winckler, G., McConnell, J.R., 2018. Concomitant variability in high-latitude aerosols, water isotopes and the hydrologic cycle. Nature Geosci 11, 853–859. https://doi.org/10.1038/s41561-018-0210-9

Markonis, Y., Koutsoyiannis, D., 2016. Scale-dependence of persistence in precipitation records. Nature Climate Change 6, 399–401. https://doi.org/10.1038/nclimate2894

Markonis, Y., Moustakis, Y., Nasika, C., Sychova, P., Dimitriadis, P., Hanel, M., Máca, P., Papalexiou, S.M., 2018. Global estimation of long-term persistence in annual river runoff. Advances in Water Resources 113, 1–12. https://doi.org/10.1016/j.advwatres.2018.01.003

Mason, S.J., Jury, M.R., 1997. Climatic variability and change over southern Africa: a reflection on underlying processes. Progress in Physical Geography: Earth and Environment 21, 23–50. https://doi.org/10.1177/030913339702100103

Matalas, N.C., 1997. Stochastic Hydrology in the Context of Climate Change. Climatic Change 37, 89–101. https://doi.org/10.1023/A:1005374000318

Matalas, N.C., 1967. Mathematical assessment of synthetic hydrology. Water Resources Research 3, 937–945.

Maxwell, R.S., Hessl, A.E., Cook, E.R., Pederson, N., 2011. A multispecies tree ring reconstruction of Potomac River streamflow (950–2001). Water Resources Research 47. https://doi.org/10.1029/2010WR010019

McInerney, D., Kavetski, D., Thyer, M., Lerat, J., Kuczera, G., 2019. Benefits of Explicit Treatment of Zero Flows in Probabilistic Hydrological Modeling of Ephemeral Catchments. Water Resources Research 55, 11035–11060. https://doi.org/10.1029/2018WR024148

McInerney, D., Westra, S., Leonard, M., Bennett, B., Thyer, M., Maier, H.R., 2023. A climate stress testing method for changes in spatially variable rainfall. Journal of Hydrology 129876. https://doi.org/10.1016/j.jhydrol.2023.129876

McKinnon, K.A., Deser, C., 2021. The Inherent Uncertainty of Precipitation Variability, Trends, and Extremes due to Internal Variability, with Implications for Western U.S. Water Resources. Journal of Climate 34, 9605–9622. https://doi.org/10.1175/JCLI-D-21-0251.1

McLeod, A.I., Hipel, K.W., 1978. Preservation of the rescaled adjusted range: 1. A reassessment of the Hurst Phenomenon. Water Resources Research 14, 491–508. https://doi.org/10.1029/WR014i003p00491

McMahon, T.A., Kiem, A.S., Peel, M.C., Jordan, P.W., Pegram, G.G.S., 2008. A New Approach to Stochastically Generating Six-Monthly Rainfall Sequences Based on Empirical Mode Decomposition. Journal of Hydrometeorology 9, 1377–1389. https://doi.org/10.1175/2008JHM991.1

McMahon, T.A., Pegram, G.G.S., Vogel, R.M., Peel, M.C., 2007a. Revisiting reservoir storage–yield relationships using a global streamflow database. Advances in Water Resources 30, 1858–1872. https://doi.org/10.1016/j.advwatres.2007.02.003

McMahon, T.A., Vogel, R.M., Pegram, G.G.S., Peel, M.C., Etkin, D., 2007b. Global streamflows – Part 2: Reservoir storage–yield performance. Journal of Hydrology 347, 260–271. https://doi.org/10.1016/j.jhydrol.2007.09.021

Mejia, J.M., Rodriguez-Iturbe, I., Dawdy, D.R., 1972. Streamflow simulation: 2. The broken line process as a potential model for hydrologic simulation. Water Resources Research 8, 931–941. https://doi.org/10.1029/WR008i004p00931

Meko, D., 1997. Dendroclimatic Reconstruction with Time Varying Predictor Subsets of Tree Indices. Journal of Climate 10, 687–696. https://doi.org/10.1175/1520-0442(1997)010<0687:DRWTVP>2.0.CO;2

Meko, D.M., Woodhouse, C.A., Baisan, C.A., Knight, T., Lukas, J.J., Hughes, M.K., Salzer, M.W., 2007. Medieval drought in the upper Colorado River Basin. Geophysical Research Letters 34. https://doi.org/10.1029/2007GL029988

Meko, D.M., Woodhouse, C.A., Winitsky, A.G., 2022. Tree-Ring Perspectives on the Colorado River: Looking Back and Moving Forward. JAWRA Journal of the American Water Resources Association n/a. https://doi.org/10.1111/1752-1688.12989

Meneghini, B., Simmonds, I., Smith, I.N., 2007. Association between Australian rainfall and the Southern Annular Mode. International Journal of Climatology 27, 109–121. https://doi.org/10.1002/joc.1370

Menne, M.J., Durre, I., Vose, R.S., Gleason, B.E., Houston, T.G., 2012. An Overview of the Global Historical Climatology Network-Daily Database. Journal of Atmospheric and Oceanic Technology 29, 897–910. https://doi.org/10.1175/JTECH-D-11-00103.1

Milly, P.C.D., Betancourt, J., Falkenmark, M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaier, D.P., Stouffer, R.J., 2008. Stationarity Is Dead: Whither Water Management? Science 319, 573–574. https://doi.org/10.1126/science.1151915

Milly, P.C.D., Betancourt, J., Falkenmark, M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaier, D.P., Stouffer, R.J., Dettinger, M.D., Krysanova, V., 2015. On Critiques of "Stationarity is Dead: Whither Water Management?" Water Resources Research 51, 7785–7789. https://doi.org/10.1002/2015WR017408

Montanari, A., Koutsoyiannis, D., 2014. Modeling and mitigating natural hazards: Stationarity is immortal! Water Resources Research 50, 9748–9756. https://doi.org/10.1002/2014WR016092

Montanari, A., Rosso, R., Taqqu, M.S., 1997. Fractionally differenced ARIMA models applied to hydrologic time series: Identification, estimation, and simulation. Water Resources Research 33, 1035–1044. https://doi.org/10.1029/97WR00043

Mortazavi-Naeini, M., Kuczera, G., Kiem, A.S., Cui, L., Henley, B., Berghout, B., Turner, E., 2015. Robust optimization to secure urban bulk water supply against extreme drought and uncertain climate change. Environmental Modelling and Software 69, 437–451. https://doi.org/10.1016/j.envsoft.2015.02.021

Mortreux, C., Barnett, J., 2017. Adaptive capacity: exploring the research frontier. WIREs Climate Change 8, e467. https://doi.org/10.1002/wcc.467

Motizuki, Y., Motoyama, H., Nakai, Y., Suzuki, K., Iizuka, Y., Takahashi, K., 2017. Overview of the chemical composition and characteristics of Na^+ and Cl^- distributions in shallow samples from Antarctic ice core DF01 (Dome Fuji) drilled in 2001. Geochemical Journal 51, 293–298. https://doi.org/10.2343/geochemj.2.0458

Mundo, I.A., Masiokas, M.H., Villalba, R., Morales, M.S., Neukom, R., Le Quesne, C., Urrutia, R.B., Lara, A., 2012. Multi-century tree-ring based reconstruction of the Neuquén River streamflow, northern Patagonia, Argentina. Climate of the Past 8, 815–829. https://doi.org/10.5194/cp-8-815-2012

Muñoz, A.A., Klock-Barría, K., Alvarez-Garreton, C., Aguilera-Betti, I., González-Reyes, Á., Lastra, J.A., Chávez, R.O., Barría, P., Christie, D., Rojas-Badilla, M., LeQuesne, C., 2020. Water Crisis in Petorca Basin, Chile: The Combined Effects of a Mega-Drought and Water Management. Water 12, 648. https://doi.org/10.3390/w12030648

Nelder, J.A., Mead, R., 1965. A Simplex Method for Function Minimization. The Computer Journal 7, 308–313. https://doi.org/10.1093/comjnl/7.4.308

Nowak, K.C., Rajagopalan, B., Zagona, E., 2011. Wavelet Auto-Regressive Method (WARM) for multi-site streamflow simulation of data with non-stationary spectra. Journal of Hydrology 410, 1–12. https://doi.org/10.1016/j.jhydrol.2011.08.051

Nye, J.F., 1963. Correction Factor for Accumulation Measured by the Thickness of the Annual Layers in an Ice Sheet. Journal of Glaciology 4, 785–788. https://doi.org/10.3189/S0022143000028367

O'Connell, P.E., Koutsoyiannis, D., Lins, H.F., Markonis, Y., Montanari, A., Cohn, T., 2016. The scientific legacy of Harold Edwin Hurst (1880–1978). null 61, 1571–1590. https://doi.org/10.1080/02626667.2015.1125998

O'Connor, J.A., Henley, B.J., Brookhouse, M.T., Allen, K.J., 2022. Ring-width and blue-light chronologies of Podocarpus lawrencei from southeastern mainland Australia reveal a regional climate signal. Climate of the Past 18, 2567–2581. https://doi.org/10.5194/cp-18-2567-2022

O'Donnell, A.J., Renton, M., Allen, K.J., Grierson, P.F., 2021. The role of extreme rain events in driving tree growth across a continental-scale climatic range in Australia. Ecography 44, 1086–1097. https://doi.org/10.1111/ecog.05671

Palmer, J.G., Cook, E.R., Turney, C.S. ~M., Allen, K., Fenwick, P., Cook, B.I., O'Donnell, A., Lough, J., Grierson, P., Baker, P., 2015. Drought variability in the eastern Australia and New Zealand summer drought atlas (ANZDA, CE 1500-2012) modulated by the Interdecadal Pacific Oscillation. Environmental Research Letters 10, 124002. https://doi.org/10.1088/1748-9326/10/12/124002

Papalexiou, S.M., Koutsoyiannis, D., 2012. Entropy based derivation of probability distributions: A case study to daily rainfall. Advances in Water Resources, Space-Time Precipitation from Urban Scale to Global Change 45, 51–57. https://doi.org/10.1016/j.advwatres.2011.11.007

Parrenin, F., Rémy, F., Ritz, C., Siegert, M.J., Jouzel, J., 2004. New modeling of the Vostok ice flow line and implication for the glaciological chronology of the Vostok ice core. Journal of Geophysical Research: Atmospheres 109. https://doi.org/10.1029/2004JD004561

Pasteris, D., McConnell, J.R., Edwards, R., Isaksson, E., Albert, M.R., 2014. Acidity decline in Antarctic ice cores during the Little Ice Age linked to changes in atmospheric nitrate and sea salt concentrations. Journal of Geophysical Research: Atmospheres 119, 5640–5652. https://doi.org/10.1002/2013JD020377

Patskoski, J., Sankarasubramanian, A., 2015. Improved reservoir sizing utilizing observed and reconstructed streamflows within a Bayesian combination framework. Water Resources Research 51, 5677–5697. https://doi.org/10.1002/2014WR016189

Patskoski, J., Sankarasubramanian, A., Wang, H., 2015. Reconstructed streamflow using SST and tree-ring chronologies over the southeastern United States. Journal of Hydrology 527, 761–775. https://doi.org/10.1016/j.jhydrol.2015.05.041

Pelletier, J.D., Turcotte, D.L., 1997. Long-range persistence in climatological and hydrological time series: analysis, modeling and application to drought hazard assessment. Journal of Hydrology 203, 198–208. https://doi.org/10.1016/S0022-1694(97)00102-9

Peng, C. -K., Havlin, S., Stanley, H.E., Goldberger, A.L., 1995. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos 5, 82–87. https://doi.org/10.1063/1.166141

Plummer, C.T., Curran, M. a. J., van Ommen, T.D., Rasmussen, S.O., Moy, A.D., Vance, T.R., Clausen, H.B., Vinther, B.M., Mayewski, P.A., 2012. An independently dated 2000-yr volcanic record from Law Dome, East Antarctica, including a new perspective on the dating of the 1450s CE eruption of Kuwae, Vanuatu. Climate of the Past 8, 1929–1940. https://doi.org/10.5194/cp-8-1929-2012

Pohl, B., Favier, V., Wille, J., Udy, D.G., Vance, T.R., Pergaud, J., Dutrievoz, N., Blanchet, J., Kittel, C., Amory, C., Krinner, G., Codron, F., 2021. Relationship Between Weather Regimes and Atmospheric Rivers in East Antarctica. Journal of Geophysical Research: Atmospheres 126, e2021JD035294. https://doi.org/10.1029/2021JD035294

Power, S., Casey, T., Folland, C., Colman, A., Mehta, V., 1999. Inter-decadal modulation of the impact of ENSO on Australia. Climate Dynamics 15, 319–324. https://doi.org/10.1007/s003820050284

Prairie, J., Nowak, K., Rajagopalan, B., Lall, U., Fulp, T., 2008. A stochastic nonparametric approach for streamflow generation combining observational and paleoreconstructed data. Water Resources Research 44. https://doi.org/10.1029/2007WR006684

Quinn, J.D., Hadjimichael, A., Reed, P.M., Steinschneider, S., 2020. Can Exploratory Modeling of Water Scarcity Vulnerabilities and Robustness Be Scenario Neutral? Earth's Future 8, e2020EF001650. https://doi.org/10.1029/2020EF001650

Raspopov, O.M., Dergachev, V.A., Esper, J., Kozyreva, O.V., Frank, D., Ogurtsov, M., Kolström, T., Shao, X., 2008. The influence of the de Vries (~200-year) solar cycle on climate variations: Results from the Central Asian Mountains and their global link. Palaeogeography, Palaeoclimatology, Palaeoecology, The Paleoecological Record from Mountain Regions 259, 6–16. https://doi.org/10.1016/j.palaeo.2006.12.017

Raupach, M.R., Briggs, P.R., Haverd, V., King, E.A., Paget, M., Trudinger, C.M., 2009. Australian Water Availability Projcect (AWAP). CSIRO.

Razavi, S., Elshorbagy, A., Wheater, H., Sauchyn, D., 2015. Toward understanding nonstationarity in climate and hydrology through tree ring proxy records. Water Resources Research 51, 1813–1830. https://doi.org/10.1002/2014WR015696

Razavi, S., Vogel, R., 2018. Prewhitening of hydroclimatic time series? Implications for inferred change and variability across time scales. Journal of Hydrology 557, 109–115. https://doi.org/10.1016/j.jhydrol.2017.11.053

Read, L.K., Vogel, R.M., 2015. Reliability, return periods, and risk under nonstationarity. Water Resources Research 51, 6381–6398. https://doi.org/10.1002/2015WR017089

Reason, C.J.C., Jagadheesha, D., 2005. Relationships between South Atlantic SST Variability and Atmospheric Circulation over the South African Region during Austral Winter. Journal of Climate 18, 3339–3355. https://doi.org/10.1175/JCLI3474.1

Ren, P., Stewardson, M., Peel, M., Fowler, K., 2023. A modified Gould-Dincer method to assess yield of carry-over reservoirs with environmental water requirements. Journal of Hydrology 617, 129065. https://doi.org/10.1016/j.jhydrol.2023.129065

Risbey, J.S., Pook, M.J., McIntosh, P.C., Wheeler, M.C., Hendon, H.H., 2009. On the Remote Drivers of Rainfall Variability in Australia. Monthly Weather Review 137, 3233–3253. https://doi.org/10.1175/2009MWR2861.1

Rittel, H.W.J., Webber, M.M., 1973. Dilemmas in a general theory of planning. Policy Sci 4, 155–169. https://doi.org/10.1007/BF01405730

Roberts, J., Plummer, C., Vance, T., van Ommen, T., Moy, A., Poynter, S., Treverrow, A., Curran, M., George, S., 2015. A 2000-year annual record of snow accumulation rates for Law Dome, East Antarctica. Climate of the Past 11, 697–707.

Rocheta, E., Evans, J.P., Sharma, A., 2017. Can Bias Correction of Regional Climate Model Lateral Boundary Conditions Improve Low-Frequency Rainfall Variability? Journal of Climate 30, 9785–9806. https://doi.org/10.1175/JCLI-D-16-0654.1

Rocheta, E., Sugiyanto, M., Johnson, F., Evans, J., Sharma, A., 2014. How well do general circulation models represent low-frequency rainfall variability? Water Resources Research 50, 2108–2123. https://doi.org/10.1002/2012WR013085

Ross, V.L., Fielding, K.S., Louis, W.R., 2014. Social trust, risk perceptions and public acceptance of recycled water: Testing a social-psychological model. Journal of Environmental Management 137, 61–68. https://doi.org/10.1016/j.jenvman.2014.01.039

Routson, C.C., Woodhouse, C.A., Overpeck, J.T., 2011. Second century megadrought in the Rio Grande headwaters, Colorado: How unusual was medieval drought? Geophysical Research Letters 38. https://doi.org/10.1029/2011GL050015

Salzer, M.W., Kipfmuller, K.F., 2005. Reconstructed Temperature And Precipitation On A Millennial Timescale From Tree-Rings In The Southern Colorado Plateau, U.S.A. Climatic Change 70, 465–487. https://doi.org/10.1007/s10584-005-5922-3

Sauchyn, D., Vanstone, J., St. Jacques, J.-M., Sauchyn, R., 2015. Dendrohydrology in Canada's western interior and applications to water resource management. Journal of Hydrology 529, 548–558. https://doi.org/10.1016/j.jhydrol.2014.11.049

Serinaldi, F., 2015. Dismissing return periods! Stoch Environ Res Risk Assess 29, 1179–1189. https://doi.org/10.1007/s00477-014-0916-1

Serinaldi, F., Kilsby, C.G., 2015. Stationarity is undead: Uncertainty dominates the distribution of extremes. Advances in Water Resources 77, 17–36. https://doi.org/10.1016/j.advwatres.2014.12.013

Severi, M., Becagli, S., Caiazzo, L., Ciardini, V., Colizza, E., Giardi, F., Mezgec, K., Scarchilli, C., Stenni, B., Thomas, E.R., Traversi, R., Udisti, R., 2017. Sea salt sodium record from Talos Dome (East Antarctica) as a potential proxy of the Antarctic past sea ice extent. Chemosphere 177, 266–274. https://doi.org/10.1016/j.chemosphere.2017.03.025

Shao, X., Huang, L., Liu, H., Liang, E., Fang, X., Wang, L., 2005. Reconstruction of precipitation variation from tree rings in recent 1000 years in Delingha, Qinghai. Science in China Series D-Earth Sciences 48, 939–949. https://doi.org/10.1360/03yd0146

Sharma, A., Tarboton, D.G., Lall, U., 1997. Streamflow simulation: A nonparametric approach. Water Resources Research 33, 291–308. https://doi.org/10.1029/96WR02839

Sheppard, P.R., Tarasov, P.E., Graumlich, L.J., Heussner, K.-U., Wagner, M., Österle, H., Thompson, L.G., 2004. Annual precipitation since 515 BC reconstructed from living and fossil juniper growth of northeastern Qinghai Province, China. Climate Dynamics 23, 869–881. https://doi.org/10.1007/s00382-004-0473-2

Shortridge, J., Aven, T., Guikema, S., 2017. Risk assessment under deep uncertainty: A methodological comparison. Reliability Engineering & System Safety 159, 12–23. https://doi.org/10.1016/j.ress.2016.10.017

Sigl, M., Fudge, T.J., Winstrup, M., Cole-Dai, J., Ferris, D., McConnell, J.R., Taylor, K.C., Welten, K.C., Woodruff, T.E., Adolphi, F., Bisiaux, M., Brook, E.J., Buizert, C., Caffee, M.W., Dunbar, N.W., Edwards, R., Geng, L., Iverson, N., Koffman, B., Layman, L., 2016. The WAIS Divide deep ice core WD2014 chronology - Part 2: Annual-layer counting (0-31 kaBP). Climate of the Past 12, 769–786.

Sigl, M., McConnell, J.R., Toohey, M., Curran, M., Das, S.B., Edwards, R., Isaksson, E., Kawamura, K., Kipfstuhl, S., Krüger, K., Layman, L., Maselli, O.J., Motizuki, Y., Motoyama, H., Pasteris, D.R., Severi, M., 2014. Insights from Antarctica on volcanic forcing during the Common Era. Nature Clim Change 4, 693–697. https://doi.org/10.1038/nclimate2293

Sigl, M., Winstrup, M., McConnell, J.R., Welten, K.C., Plunkett, G., Ludlow, F., Büntgen, U., Caffee, M., Chellman, N., Dahl-Jensen, D., Fischer, H., Kipfstuhl, S., Kostick, C., Maselli, O.J., Mekhaldi, F., Mulvaney, R., Muscheler, R., Pasteris, D.R., Pilcher, J.R., Salzer, M., Schüpbach, S., Steffensen, J.P., Vinther, B.M., Woodruff, T.E., 2015. Timing and climate forcing of volcanic eruptions for the past 2,500 years. Nature 523, 543–549. https://doi.org/10.1038/nature14565

Simpson, N.P., Mach, K.J., Constable, A., Hess, J., Hogarth, R., Howden, M., Lawrence, J., Lempert, R.J., Muccione, V., Mackey, B., New, M.G., O'Neill, B., Otto, F., Pörtner, H.-O., Reisinger, A., Roberts, D., Schmidt, D.N., Seneviratne, S., Strongin, S., Van Aalst, M., Totin, E., Trisos, C.H., 2021. A framework for complex climate change risk assessment. One Earth 4, 489–501. https://doi.org/10.1016/j.oneear.2021.03.005

Sivakumar, B., 2000. Chaos theory in hydrology: important issues and interpretations. Journal of Hydrology 227, 1–20. https://doi.org/10.1016/S0022-1694(99)00186-9

Smit, B., Wandel, J., 2006. Adaptation, adaptive capacity and vulnerability. Global Environmental Change, Resilience, Vulnerability, and Adaptation: A Cross-Cutting Theme of the International Human Dimensions Programme on Global Environmental Change 16, 282–292. https://doi.org/10.1016/j.gloenvcha.2006.03.008

Solomon, S., Greenberg, J., Pyszczynski, T., 2015. The worm at the core: On the role of death in life. Random House.

Srikanthan, R., McMahon, T. ~A., 2001. Stochastic generation of annual, monthly and daily climate data: A review. Hydrology and Earth System Sciences 5, 653–670.

Stager, J.C., Mayewski, P.A., White, J., Chase, B.M., Neumann, F.H., Meadows, M.E., King, C.D., Dixon, D.A., 2012. Precipitation variability in the winter rainfall zone of South Africa during the last 1400 yr linked to the austral westerlies. Climate of the Past 8, 877–887. https://doi.org/10.5194/cp-8-877-2012

Stahle, D.K., Burnette, D.J., Stahle, D.W., 2013. A Moisture Balance Reconstruction for the Drainage Basin of Albemarle Sound, North Carolina. Estuaries and Coasts 36, 1340–1353. https://doi.org/10.1007/s12237-013-9643-y

Stahle, D.W., Burnette, D.J., Villanueva, J., Cerano, J., Fye, F.K., Griffin, R.D., Cleaveland, M.K., Stahle, D.K., Edmondson, J.R., Wolff, K.P., 2012. Tree-ring analysis of ancient baldcypress trees and subfossil wood. Quaternary Science Reviews 34, 1–15. https://doi.org/10.1016/j.quascirev.2011.11.005

Stahle, D.W., Cleaveland, M.K., Grissino-Mayer, H.D., Griffin, R.D., Fye, F.K., Therrell, M.D., Burnette, D.J., Meko, D.M., Diaz, J.V., 2009. Cool- and Warm-Season Precipitation Reconstructions over Western New Mexico. Journal of Climate 22, 3729–3750. https://doi.org/10.1175/2008JCLI2752.1

Stahle, D.W., Diaz, J.V., Burnette, D.J., Paredes, J.C., Heim Jr., R.R., Fye, F.K., Acuna Soto, R., Therrell, M.D., Cleaveland, M.K., Stahle, D.K., 2011. Major Mesoamerican droughts of the past millennium. Geophysical Research Letters 38. https://doi.org/10.1029/2010GL046472

Stanton, M.C.B., Roelich, K., 2021. Decision making under deep uncertainties: A review of the applicability of methods in practice. Technological Forecasting and Social Change 171, 120939. https://doi.org/10.1016/j.techfore.2021.120939

Stedinger, J.R., 1980. Fitting log normal distributions to hydrologic data. Water Resources Research 16, 481–490. https://doi.org/10.1029/WR016i003p00481

Stedinger, J.R., Taylor, M.R., 1982a. Synthetic streamflow generation: 2. Effect of parameter uncertainty. Water Resources Research 18, 919–924. https://doi.org/10.1029/WR018i004p00919

Stedinger, J.R., Taylor, M.R., 1982b. Synthetic streamflow generation: 1. Model verification and validation. Water Resources Research 18, 909–918. https://doi.org/10.1029/WR018i004p00909

Steiger, N.J., Hakim, G.J., Steig, E.J., Battisti, D.S., Roe, G.H., 2014. Assimilation of Time-Averaged Pseudoproxies for Climate Reconstruction. Journal of Climate 27, 426–441.

Steiger, N.J., Steig, E.J., Dee, S.G., Roe, G.H., Hakim, G.J., 2017. Climate reconstruction using data assimilation of water isotope ratios from ice cores. Journal of Geophysical Research: Atmospheres 122, 1545–1568. https://doi.org/10.1002/2016JD026011

Stephens, C.M., Marshall, L.A., Johnson, F.M., Lin, L., Band, L.E., Ajami, H., 2020. Is Past Variability a Suitable Proxy for Future Change? A Virtual Catchment Experiment. Water Resources Research e2019WR026275. https://doi.org/10.1029/2019WR026275

Stevenson, S., Coats, S., Touma, D., Cole, J., Lehner, F., Fasullo, J., Otto-Bliesner, B., 2022. Twenty-first century hydroclimate: A continually changing baseline, with more frequent extremes. Proceedings of the National Academy of Sciences 119, e2108124119. https://doi.org/10.1073/pnas.2108124119

Sun, F., Roderick, M.L., Farquhar, G.D., 2018. Rainfall statistics, stationarity, and climate change. Proceedings of the National Academy of Sciences 115, 2305–2310. https://doi.org/10.1073/pnas.1705349115

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., Gelman, A., 2020. Validating Bayesian Inference Algorithms with Simulation-Based Calibration. https://doi.org/10.48550/arXiv.1804.06788

Taqqu, M.S., Teverosky, V., Willinger, W., 1995. Estimators for long-range dependence: An empirical study. Fractals 03, 785–798. https://doi.org/10.1142/S0218348X95000692

Thomas, E.R., van Wessem, J.M., Roberts, J., Isaksson, E., Schlosser, E., Fudge, T.J., Vallelonga, P., Medley, B., Lenaerts, J., Bertler, N., van den Broeke, M.R., Dixon, D.A., Frezzotti, M., Stenni, B., Curran, M., Ekaykin, A.A., 2017. Regional Antarctic snow accumulation over the past 1000 years. Climate of the Past 13, 1491–1513.

Thomas, E.R., Vladimirova, D.O., Tetzner, D.R., Emanuelsson, B.D., Chellman, N., Dixon, D.A., Goosse, H., Grieman, M.M., King, A.C.F., Sigl, M., Udy, D.G., Vance, T.R., Winski, D.A., Winton, V.H.L., Bertler, N.A.N., Hori, A., Laluraj, C.M., McConnell, J.R., Motizuki, Y., Takahashi, K., Motoyama, H., Nakai, Y., Schwanck, F., Simões, J.C., Lindau, F.G.L., Severi, M., Traversi, R., Wauthy, S., Xiao, C., Yang, J., Mosely-Thompson, E., Khodzher, T.V., Golobokova, L.P., Ekaykin, A.A., 2023. Ice core chemistry database: an Antarctic compilation of sodium and sulfate records spanning the past 2000 years. Earth System Science Data 15, 2517–2532. https://doi.org/10.5194/essd-15-2517-2023

Thompson, L.G., Mosley-Thompson, E., Davis, M.E., Zagorodnov, V.S., Howat, I.M., Mikhalenko, V.N., Lin, P.-N., 2013. Annually Resolved Ice Core Records of Tropical Climate Variability over the Past ~1800 Years. Science 340, 945–950. https://doi.org/10.1126/science.1234210

Thyer, M., Frost, A.J., Kuczera, G., 2006. Parameter estimation and model identification for stochastic models of annual hydrological data: Is the observed record long enough? Journal of Hydrology 330, 313–328. https://doi.org/10.1016/j.jhydrol.2006.03.029

Thyer, M., Kuczera, G., 2000. Modeling long-term persistence in hydroclimatic time series using a hidden state Markov Model. Water Resources Research 36, 3301–3310. https://doi.org/10.1029/2000WR900157

Thyer, M., Kuczera, G., Wang, Q.J., 2002. Quantifying parameter uncertainty in stochastic models using the Box–Cox transformation. Journal of Hydrology 265, 246–257. https://doi.org/10.1016/S0022-1694(02)00113-0

Tingstad, A.H., Groves, D.G., Lempert, R.J., 2014. Paleoclimate scenarios to inform decision making in water resource management: example from Southern California's inland empire. Journal of Water Resources Planning and Management 140, 4014025.

Tolwinski-Ward, S., Evans, M., Hughes, M., Anchukaitis, K., 2011. An efficient forward model of the climate controls on interannual variation in tree-ring width. Climate Dynamics 36, 2419–2439.

Torrence, C., Compo, G.P., 1998. A Practical Guide to Wavelet Analysis. Bulletin of the American Meteorological Society 79, 61–78. https://doi.org/10.1175/1520-0477(1998)079<0061:APGTWA>2.0.CO;2

Touchan, R., Woodhouse, C.A., Meko, D.M., Allen, C., 2011. Millennial precipitation reconstruction for the Jemez Mountains, New Mexico, reveals changingb drought signal. International Journal of Climatology 31, 896–906. https://doi.org/10.1002/joc.2117

Tozer, C.R., Kiem, A.S., Vance, T.R., Roberts, J.L., Curran, M.A.J., Moy, A.D., 2018. Reconstructing pre-instrumental streamflow in Eastern Australia using a water balance approach. Journal of Hydrology 558, 632–646. https://doi.org/10.1016/j.jhydrol.2018.01.064

Tozer, C.R., Vance, T.R., Roberts, J., Kiem, A.S., Curran, M.A.J., Moy, A.D., 2016. An ice core derived 1013-year catchment scale annual rainfall reconstruction in subtropical eastern Australia. Hydrology and Earth System Sciences 20, 12483–12514.

Traufetter, F., Oerter, H., Fischer, H., Weller, R., Miller, H., 2004. Spatio-temporal variability in volcanic sulphate deposition over the past 2 kyr in snow pits and firn cores from Amundsenisen, Antarctica. Journal of Glaciology 50, 137–146. https://doi.org/10.3189/172756504781830222

Tsoukalas, I., Kossieris, P., Makropoulos, C., 2020. Simulation of Non-Gaussian Correlated Random Variables, Stochastic Processes and Random Fields: Introducing the anySim R-Package for Environmental Applications and Beyond. Water 12. https://doi.org/10.3390/w12061645

Tsoukalas, I., Makropoulos, C., Koutsoyiannis, D., 2018. Simulation of Stochastic Processes Exhibiting Any-Range Dependence and Arbitrary Marginal Distributions. Water Resources Research 54, 9484–9513. https://doi.org/10.1029/2017WR022462

Turner, S.W.D., Marlow, D., Ekström, M., Rhodes, B.G., Kularathna, U., Jeffrey, P.J., 2014. Linking climate projections to performance: A yield-based decision scaling assessment of a large urban water resources system. Water Resources Research 50, 3553–3567. https://doi.org/10.1002/2013WR015156

Tyralis, H., Dimitriadis, P., Koutsoyiannis, D., O'Connell, P.E., Tzouka, K., Iliopoulou, T., 2018. On the long-range dependence properties of annual precipitation using a global network of instrumental measurements. Advances in Water Resources 111, 301–318. https://doi.org/10.1016/j.advwatres.2017.11.010

Tyralis, H., Koutsoyiannis, D., 2011. Simultaneous estimation of the parameters of the Hurst–Kolmogorov stochastic process. Stochastic Environmental Research and Risk Assessment 25, 21–33. https://doi.org/10.1007/s00477-010-0408-x

Udy, D.G., Vance, T.R., Kiem, A.S., Holbrook, N.J., 2022. A synoptic bridge linking sea salt aerosol concentrations in East Antarctic snowfall to Australian rainfall. Commun Earth Environ 3, 1–11. https://doi.org/10.1038/s43247-022-00502-w

Udy, D.G., Vance, T.R., Kiem, A.S., Holbrook, N.J., Curran, M.A.J., 2021. Links between Large-Scale Modes of Climate Variability and Synoptic Weather Patterns in the Southern Indian Ocean. Journal of Climate 34, 883–899. https://doi.org/10.1175/JCLI-D-20-0297.1

Upadhyay, S.K., Singh, U., Dey, D.K., Loganathan, A., 2015. Current Trends in Bayesian Methodology with Applications. CRC Press.

Van Gael, J., Ghahramani, Z., 2011. Nonparametric hidden Markov models, in: Cemgil, A.T., Barber, D., Chiappa, S. (Eds.), Bayesian Time Series Models. Cambridge University Press, Cambridge, pp. 317–340. https://doi.org/10.1017/CBO9780511984679.016

van Ommen, T.D., Morgan, V., 2010. Snowfall increase in coastal East Antarctica linked with southwest Western Australian drought. Nature Geoscience 3, 267–272. https://doi.org/10.1038/ngeo761

Vance, T.R., Kiem, A.S., Jong, L.M., Roberts, J.L., Plummer, C.T., Moy, A.D., Curran, M.A.J., van Ommen, T.D., 2022. Pacific decadal variability over the last 2000 years and implications for climatic risk. Communications Earth & Environment 3, 33. https://doi.org/10.1038/s43247-022-00359-z

Vance, T.R., Roberts, J.L., Plummer, C.T., Kiem, A.S., van Ommen, T.D., 2015. Interdecadal Pacific variability and eastern Australian megadroughts over the last millennium. Geophysical Research Letters 42, 129–137. https://doi.org/10.1002/2014GL062447

Vance, T.R., van Ommen, T.D., Curran, M.A.J., Plummer, C.T., Moy, A.D., 2013. A Millennial Proxy Record of ENSO and Eastern Australian Rainfall from the Law Dome Ice Core, East Antarctica. Journal of Climate 26, 710–725.

Verdon, D.C., Wyatt, A.M., Kiem, A.S., Franks, S.W., 2004. Multidecadal variability of rainfall and streamflow: Eastern Australia. Water Resources Research 40. https://doi.org/10.1029/2004WR003234

Verdon-Kidd, D.C., Hancock, G.R., Lowry, J.B., 2017. A 507-year rainfall and runoff reconstruction for the Monsoonal North West, Australia derived from remote paleoclimate archives. Global and Planetary Change 158, 21–35. https://doi.org/10.1016/j.gloplacha.2017.09.003

Visser, I., 2011. Seven things to remember about hidden Markov models: A tutorial on Markovian models for time series. Journal of Mathematical Psychology 55, 403–415. https://doi.org/10.1016/j.jmp.2011.08.002

Visser, I., Speekenbrink, M., 2010. depmixS4: An R Package for Hidden Markov Models. Journal of Statistical Software 36, 1–21. https://doi.org/10.18637/jss.v036.i07
Visser, J.B., Kim, S., Wasko, C., Nathan, R., Sharma, A., 2022. The Impact of Climate Change on Operational Probable Maximum Precipitation Estimates. Water Resources Research 58, e2022WR032247. https://doi.org/10.1029/2022WR032247
Vogel, R.M., 2017. Stochastic watershed models for hydrologic risk management. Water Security 1, 28–35. https://doi.org/10.1016/j.wasec.2017.06.001
Vogel, R.M., Bolognese, R.A., 1995. Storage-Reliability-Resilience-Yield Relations for Over-Year Water Supply Systems. Water Resources Research 31, 645–654. https://doi.org/10.1029/94WR02972
Vogel, R.M., Fennessey, N.M., 1993. L moment diagrams should replace product moment diagrams. Water Resources Research 29, 1745–1752. https://doi.org/10.1029/93WR00341
Vogel, R.M., Lane, M., Ranjith, R., Kirshen, P., 1999. Storage Reservoir Behavior in the United States. Journal of Water Resources Planning and Management 125, 245–254. https://doi.org/10.1061/(ASCE)0733-9496(1999)125:5(245)
Vrugt, J.A., ter Braak, C.J.F., Diks, Robinson, B.A., Hyman, J.M., Higdon, D., 2009. Accelerating Markov Chain Monte Carlo Simulation by Differential Evolution with Self-Adaptive Randomized Subspace Sampling. International Journal of Nonlinear Sciences and Numerical Simulation 10, 273–290. https://doi.org/10.1515/IJNSNS.2009.10.3.273
Wallis, J.R., Matalas, N.C., 1970. Small Sample Properties of H and K—Estimators of the Hurst Coefficient h. Water Resources Research 6, 1583–1594. https://doi.org/10.1029/WR006i006p01583
Wang, H., Gao, X., Qian, L., Yu, S., 2012. Uncertainty analysis of hydrological processes based on ARMA-GARCH model. Science China Technological Sciences 55, 2321–2331. https://doi.org/10.1007/s11431-012-4909-3
Wang, H.-H., Van Voorn, G., Grant, W.E., Zare, F., Giupponi, C., Steinmann, P., Müller, B., Elsawah, S., van Delden, H., Athanasiadis, I.N., Sun, Z., Jager, W., Little, J.C., Jakeman, A.J., 2023. Scale decisions and good practices in socio-environmental systems modelling: Guidance and documentation during problem scoping and model formulation. Socio-Environmental Systems Modelling 5. https://doi.org/10.18174/sesmo.18563
Wang, Y., Thomas, E.R., Hou, S., Huai, B., Wu, S., Sun, W., Qi, S., Ding, M., Zhang, Y., 2017. Snow Accumulation Variability Over the West Antarctic Ice Sheet Since 1900: A Comparison of Ice Core Records With ERA-20C Reanalysis. Geophysical Research Letters 44, 11,411-482,490. https://doi.org/10.1002/2017GL075135
Wasko, C., Sharma, A., 2015. Steeper temporal distribution of rain intensity at higher temperatures within Australian storms. Nature Geosci 8, 527–529. https://doi.org/10.1038/ngeo2456
Weedon, G., 2003. Time Series Analysis and Cyclostratigraphy: Examining Stratigraphic Records of Environmental Cycles, Cambridge University Press: Cambridge. https://doi.org/10.1017/CBO9780511535482
Weron, R., 2002a. Estimating long-range dependence: finite sample properties and confidence intervals. Physica A: Statistical Mechanics and its Applications 312, 285–299. https://doi.org/10.1016/S0378-4371(02)00961-5
Weron, R., 2002b. Estimating long-range dependence: finite sample properties and confidence intervals. Physica A: Statistical Mechanics and its Applications 312, 285–299. https://doi.org/10.1016/S0378-4371(02)00961-5
Westra, S., Alexander, L.V., Zwiers, F.W., 2013. Global Increasing Trends in Annual Maximum Daily Precipitation. Journal of Climate 26, 3904–3918. https://doi.org/10.1175/JCLI-D-12-00502.1
Westra, S., Renard, B., Thyer, M., 2015. The ENSO–Precipitation Teleconnection and Its Modulation by the Interdecadal Pacific Oscillation. Journal of Climate 28, 4753–4773. https://doi.org/10.1175/JCLI-D-14-00722.1
Westra, S., Zscheischler, J., 2023. Accounting for systemic complexity in the assessment of climate risk. One Earth 6, 645–655. https://doi.org/10.1016/j.oneear.2023.05.005
Wigley, T.M.L., Briffa, K.R., Jones, P.D., 1984. On the Average Value of Correlated Time Series, with Applications in Dendroclimatology and Hydrometeorology. Journal of Applied Meteorology and Climatology 23, 201–213. https://doi.org/10.1175/1520-0450(1984)023<0201:OTAVOC>2.0.CO;2
Wille, J.D., Favier, V., Gorodetskaya, I.V., Agosta, C., Kittel, C., Beeman, J.C., Jourdain, N.C., Lenaerts, J.T.M., Codron, F., 2021. Antarctic Atmospheric River Climatology and Precipitation Impacts. Journal of Geophysical Research: Atmospheres 126, e2020JD033788. https://doi.org/10.1029/2020JD033788
Winski, D.A., Fudge, T.J., Ferris, D.G., Osterberg, E.C., Fegyveresi, J.M., Cole-Dai, J., Thundercloud, Z., Cox, T.S., Kreutz, K.J., Ortman, N., Buizert, C., Epifanio, J., Brook, E.J., Beaudette, R., Severinghaus, J., Sowers, T., Steig, E.J., Kahle, E.C., Jones, T.R., Morris, V., Aydin, M., Nicewonger, M.R., Casey, K.A., Alley, R.B., Waddington, E.D., Iverson, N.A., Dunbar, N.W., Bay, R.C., Souney, J.M., Sigl, M.,

McConnell, J.R., 2019. The SP19 chronology for the South Pole Ice Core -- Part 1: volcanic matching and annual layer counting. Climate of the Past 15, 1793–1808. https://doi.org/10.5194/cp-15-1793-2019

Winstrup, M., Vallelonga, P., Kjær, H.A., Fudge, T.J., Lee, J.E., Riis, M.H., Edwards, R., Bertler, N.A.N., Blunier, T., Brook, E.J., Buizert, C., Ciobanu, G., Conway, H., Dahl-Jensen, D., Ellis, A., Emanuelsson, B.D., Hindmarsh, R.C.A., Keller, E.D., Kurbatov, A.V., Mayewski, P.A., 2019. A 2700-year annual timescale and accumulation history for an ice core from Roosevelt Island, West Antarctica. Climate of the Past 15, 751–779.

Wolff, E.W., Fischer, H., Fundel, F., Ruth, U., Twarloh, B., Littot, G.C., Mulvaney, R., Röthlisberger, R., de Angelis, M., Boutron, C.F., Hansson, M., Jonsell, U., Hutterli, M.A., Lambert, F., Kaufmann, P., Stauffer, B., Stocker, T.F., Steffensen, J.P., Bigler, M., Siggaard-Andersen, M.L., Udisti, R., Becagli, S., Castellano, E., Severi, M., Wagenbach, D., Barbante, C., Gabrielli, P., Gaspari, V., 2006. Southern Ocean sea-ice extent, productivity and iron flux over the past eight glacial cycles. Nature 440, 491–496. https://doi.org/10.1038/nature04614

Woodhouse, C.A., Pederson, G.T., Gray, S.T., 2011. An 1800-yr record of decadal-scale hydroclimatic variability in the upper Arkansas River basin from bristlecone pine. Quaternary Research 75, 483–490. https://doi.org/10.1016/j.yqres.2010.12.007

Wu, W., Eamen, L., Dandy, G., Razavi, S., Kuczera, G., Maier, H.R., 2023. Beyond engineering: A review of reservoir management through the lens of wickedness, competing objectives and uncertainty. Environmental Modelling & Software 167, 105777. https://doi.org/10.1016/j.envsoft.2023.105777

Xiong, R., Zheng, Y., Han, F., Tian, Y., 2021. Improving the Scientific Understanding of the Paradox of Irrigation Efficiency: An Integrated Modeling Approach to Assessing Basin-Scale Irrigation Efficiency. Water Resources Research 57, e2020WR029397. https://doi.org/10.1029/2020WR029397

Yates, D., Gangopadhyay, S., Rajagopalan, B., Strzepek, K., 2003. A technique for generating regional climate scenarios using a nearest-neighbor algorithm. Water Resources Research 39. https://doi.org/10.1029/2002WR001769

Yuan, N., Xiong, F., Xoplaki, E., He, W., Luterbacher, J., 2021. A new approach to correct the overestimated persistence in tree-ring width based precipitation reconstructions. Climate Dynamics. https://doi.org/10.1007/s00382-021-06024-z

Yue, S., Pilon, P., Phinney, B., Cavadias, G., 2002. The influence of autocorrelation on the ability to detect trend in hydrological series. Hydrological Processes 16, 1807–1829. https://doi.org/10.1002/hyp.1095

Zhang, H., Yuan, N., Esper, J., Werner, J.P., Xoplaki, E., Büntgen, U., Treydte, K., Luterbacher, J., 2015. Modified climate with long term memory in tree ring proxies. Environ. Res. Lett. 10, 084020. https://doi.org/10.1088/1748-9326/10/8/084020

Zheng, Y., Jong, L.M., Phipps, S.J., Roberts, J.L., Moy, A.D., Curran, M.A.J., van Ommen, T.D., 2021. Extending and understanding the South West Western Australian rainfall record using a snowfall reconstruction from Law Dome, East Antarctica. Climate of the Past 17, 1973–1987. https://doi.org/10.5194/cp-17-1973-2021

Chapter 9. Appendices

9.1 Chapter 3 Appendix

9.1.1 Residual diagnostics

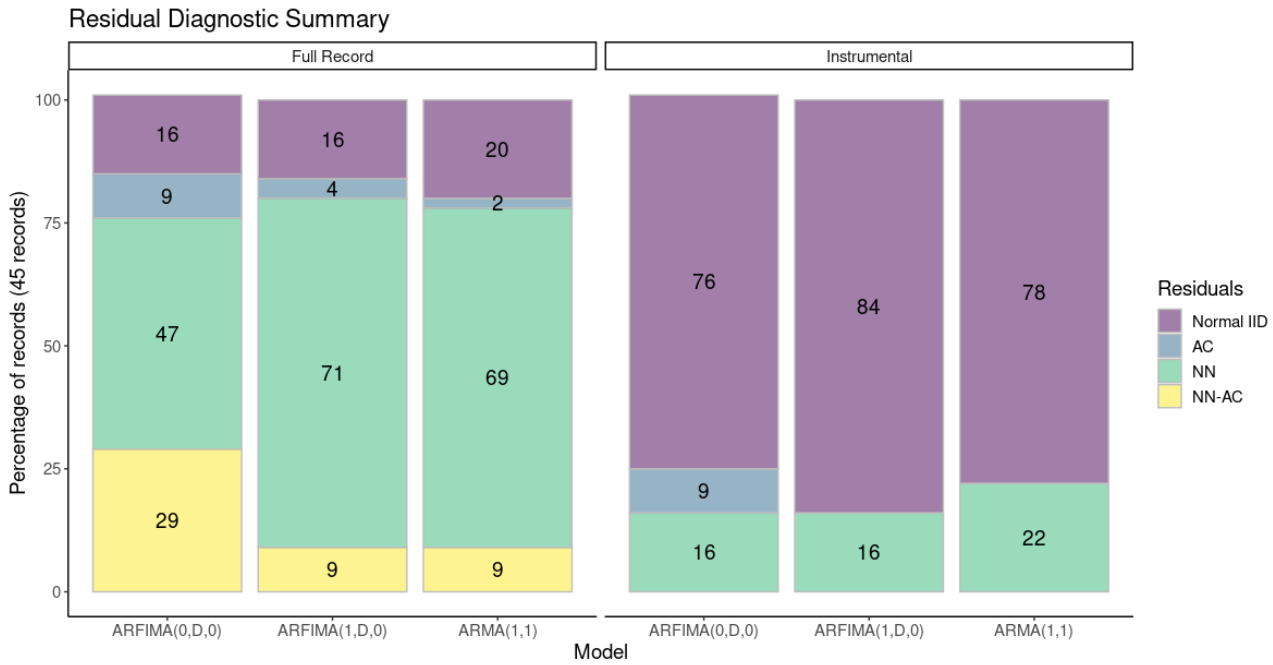

Figure 9-1: Summary of residual diagnostics for the ARFIMA(0,D,0), ARFIMA(0,D,0), and ARMA(1,1) models. The proportion of models with either normal and independent and identically distributed residuals (Normal IID); normal and autocorrelated residuals (AC); non-normal and independent residuals (NN); and non-normal and autocorrelated residuals (NN-AC) are shown. "Full record" models were used for Experiment 3, "Instrumental" models were used for Experiments 1 and 2. Normality was evaluated using a Shapiro-Wilks test and autocorrelation was evaluated using a Ljung-Box test.

9.1.2 Experiment 3 with Na+ records removed

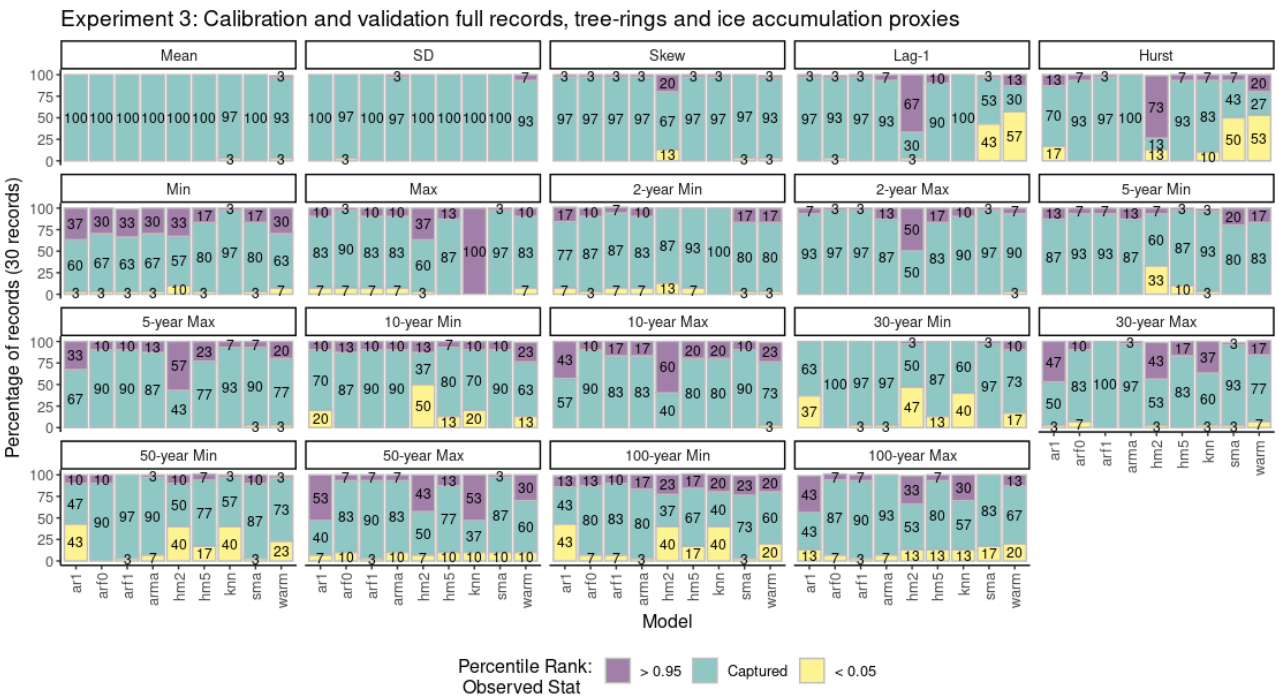

Figure 9-2: Same as *Figure 3-6*, but only ice core accumulation and tree-ring results are presented

9.2 Chapter 4 Appendix

9.2.1 Approximate likelihood function for skewed data

Bayesian inference requires estimation of parameter likelihoods; for both the ARFIMA(0,D,0) and ARMA(1,1) models, the likelihood is calculated assuming the data is normally distributed. However, hydroclimatic data (and the proxy records examined in this thesis) typically have skewed marginal distributions and a finite lower bound of zero (i.e., are non-normal). In order to remove this skew prior to model calibration, a Box-Cox transformation was used (Box and Cox, 1964). Although this transformation can distort a timeseries' autocovariance function (Montanari et al., 1997), it has been shown to (a) produce residuals that are normally distributed (if a model is fitted in transformed space); and (b) reproduce the skew of the timeseries marginal distribution (after back-transformation) (Srikanthan and McMahon, 2001). Therefore, this transformation is often used in operational water management. For a timeseries y, the Box-Cox transformation that produces the transformed timeseries z using the parameter **λ** in the following equation:

$$if\ \lambda \neq 0: z_t = \frac{y^{\lambda}{}_t - 1}{\lambda}$$
$$if\ \lambda = 0: z_t = log(y_t)$$

Equation 9-1

Although it is useful for removing skew, the Box-Cox transformation complicates inferring the stochastic model posteriors via MCMC methods. This is because the transformation introduces a strong dependence between potential $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}/\boldsymbol{\sigma}$ parameters, which prevents proper exploration of the posterior space (Thyer et al., 2002). To reduce this dependence when calculating the likelihood of a proposed parameter set, first order approximations of $\boldsymbol{\mu}/\boldsymbol{\sigma}$ in the Box-Cox transformed space can be used. Thyer et al. (2002) derived these approximations, with the transformed mean $\boldsymbol{\mu}_z$ being expressed in terms of the sample mean $\boldsymbol{\mu}_y$ and $\boldsymbol{\lambda}$:

$$if\ \lambda \neq 0: \mu_z = \frac{\mu^{\lambda}{}_y - 1}{\lambda}$$
$$if\ \lambda = 0: \mu_z = log(\mu_y) \qquad \text{Equation 9-2}$$

The first order approximation of the transformed variance $\boldsymbol{\sigma}^2{}_z$ was derived in terms of the sample mean $\boldsymbol{\mu}_y$; sample variance $\mathbf{s}^2{}_y$; and $\boldsymbol{\lambda}$, giving:

$$\sigma_z{}^2 = m_y{}^{2(\lambda-1)} s_y{}^2 \qquad \text{Equation 9-3}$$

Calculation of the likelihood function requires $\boldsymbol{\sigma}^2{}_z$ to be expressed in terms of the residual variance $\boldsymbol{\sigma}^2{}_\varepsilon$. For the ARMA(1,1) model, this is:

$$\sigma_z{}^2 = \frac{(1 + 2\phi\theta + \theta^2)\sigma_\epsilon{}^2}{1 - \phi^2} \qquad \text{Equation 9-4}$$

Inserting Equation 9-3 into Equation 9-4 then rearranging, we obtain a first order approximation of the residual variance $\boldsymbol{\sigma}^2{}_\varepsilon$ in terms of the sample mean $\boldsymbol{\mu}_y$; sample variance $\mathbf{s}^2{}_y$; the $\boldsymbol{\lambda}$ parameter; and the ARMA(1,1) $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ parameters:

$$\sigma_\epsilon{}^2 = \frac{\mu_y{}^{2(\lambda-1)} s_y{}^2 (1 - \phi^2)}{(1 + 2\phi\theta + \theta^2)} \qquad \text{Equation 9-5}$$

For the ARFIMA(0,D,0) model, $\boldsymbol{\sigma}^2{}_z$ can be expressed in terms of residual variance $\boldsymbol{\sigma}^2{}_\varepsilon$ and the D parameter as:

$$\sigma_z{}^2 = \frac{\sigma_\epsilon{}^2\,\Gamma(1-2D)}{[\Gamma(1-D)]^2} \qquad \text{Equation 9-6}$$

And, as with the ARMA(1,1) model, inserting Equation 9-5 into Equation 9-6 and rearranging, we obtain:

$$\sigma_\epsilon{}^2 = \frac{\mu_y{}^{2(\lambda-1)}\,s_y{}^2\,[\Gamma(1-D)]^2}{\Gamma(1-2D)} \qquad \text{Equation 9-7}$$

When calculating the likelihood of a proposed parameter set, an additional constraint introduced by the Box-Cox transformation is that it can only be applied to positive data. To ensure a successful back-transformation to the original timeseries scale, this means that the transformed data have upper/lower bounds that are a function of $\boldsymbol{\lambda}$ (also derived by Thyer et al. (2000)) and follow a Truncated Normal distribution such that:

$$z_t \mid z_{t-1}, \epsilon_{t-1} \sim \mathrm{TN}(\bar{z}_t, \sigma_\epsilon{}^2, \mathrm{lb}, \mathrm{ub}) \text{ for the ARMA(1,1) model and}$$
$$z_t \mid \epsilon_{t-1,\ldots,1} \sim \mathrm{TN}(\bar{z}_t, \sigma_\epsilon{}^2, \mathrm{lb}, \mathrm{ub}) \text{ for the ARFIMA(0,D,0) model} \qquad \text{Equation 9-8}$$

Where $\check{z}_t$ is the conditional mean of the transformed timeseries and lb and ub are respective lower and upper bounds such that:

$$if\ \lambda > 0: lb = \frac{-1}{\lambda}\ ;\ ub = \infty$$
$$if\ \lambda < 0: lb = -\infty\ ;\ ub = \frac{-1}{\lambda} \qquad \text{Equation 9-9}$$

These upper and lower bounds are derived to satisfy the constraint $z_t\boldsymbol{\lambda} + 1 > 0$

Considering these approximations/constraints, the likelihood of a proposed parameter set θ_p is:

$$P(\theta_p \mid y) \propto \prod_{t=2}^{n} y_t{}^{\lambda-1}\,\mathrm{TN}(\bar{z}_t, \sigma_\epsilon{}^2, \mathrm{lb}, \mathrm{ub}) \qquad \text{Equation 9-10}$$

Note that the $y_t^{\lambda-1}$ term leading the Truncated Normal density is a Jacobian adjustment that accounts for any distortions in the posterior density introduced by the non-linear Box-Cox transformation.

9.2.2 Selection of prior distributions

For both ARMA and ARFIMA models, non-informative priors were used (Frost et al., 2007).

$$\lambda \sim Uniform(-2, 2)$$
$$\mu_y \sim Normal(\overline{\mu}_y, \overline{\sigma}_y) \qquad \text{Equation 9-11}$$
$$\sigma_y \sim Inverse\ Gamma(1, \overline{\sigma}_y)$$

Where $\overline{\mu}_y$ and $\overline{\sigma}_y$ are the sample mean and standard deviation respectively. For the persistence parameters, both models assumed uniform persistence priors. For the ARMA(1,1) model, these were:

$$\phi \sim Uniform(-1, 1)$$
$$\theta \sim Uniform(-1, 1) \qquad \text{Equation 9-12}$$

For the ARFIMA(0,D,0) model, these were:

$$D \sim Uniform(-0.5, 0.5) \qquad \text{Equation 9-13}$$

9.3 Chapter 5 Appendix

9.3.1 CIMSS model likelihood

Respective IPO 'wet' and 'dry' phases (i.e., IPO negative and IPO neutral-positive) were assumed to follow a gamma distribution.

$$IPO\ wet \sim Gamma(\alpha_w, \beta_w)$$
$$IPO\ dry \sim Gamma(\alpha_d, \beta_d) \qquad \text{Equation 9-14}$$

An AR(1) model was then calibrated to respective wet/dry phases in the instrumental rainfall record. These phases have, separate mean, standard deviation, and Box-Cox **λ** parameters identified for each phase, but share the same persistence parameter This meant that, for IPO state *j* at time *t*, the transformed annual rainfall timeseries *z* at time *t* can be expressed as:

$$z_{jt} = \mu_j + \phi * (z_{j(t-1)} - \mu_{j(t-1)}) + \epsilon_{jt} \qquad \text{Equation 9-15}$$

$$\epsilon_{jt} \sim N(0, \sigma_j) \qquad \text{Equation 9-16}$$

Where

$$if\ \lambda_j \neq 0: z_{jt} = \frac{{y_t}^{\lambda j} - 1}{\lambda_j}$$
$$if\ \lambda_j = 0: z_{jt} = log(y_t) \qquad \text{Equation 9-17}$$

As with the proposed model, the CIMSS framework allows Bayesian calibration and inference of posteriors. The likelihood of the CIMSS framework used in this study can be expressed as:

$$P(\theta_{run}, \theta_{rain} | Y_{ipo}, Y_{rain}) = P(\theta_{run} | Y_{ipo})\, P(\theta_{rain} |\, Y_{rain}) \qquad \text{Equation 9-18}$$

$$P(\theta_{run} | Y_{ipo})\, P(\theta_{rain} |\, Y_{rain}) \qquad \text{Equation 9-19}$$
$$\propto P(Y_{ipo} |\, \alpha_w, \alpha_d, \beta_w, \beta_d)\, P(Y_{rain} | \mu_w, \mu_d, \sigma_w, \sigma_d . \lambda_w, \lambda_d, \phi)$$

Where *w* and *d* subscripts refer to parameters identified for wet/dry IPO phases respectively. Prior distributions for each parameter were selected following Frost et al. (2007) and posteriors inferred using the NUTS algorithm.

9.3.2 CIMSS model validation

Figure 9-3 demonstrates that the gamma distribution is an appropriate choice for respective IPO negative and IPO neutral-positive phases, Figure 9-4 demonstrates that CIMSS model residuals were independent and normally distributed, and Figure 9-5 demonstrates that the CIMSS model was able to reproduce key Williams River hydrological statistics.

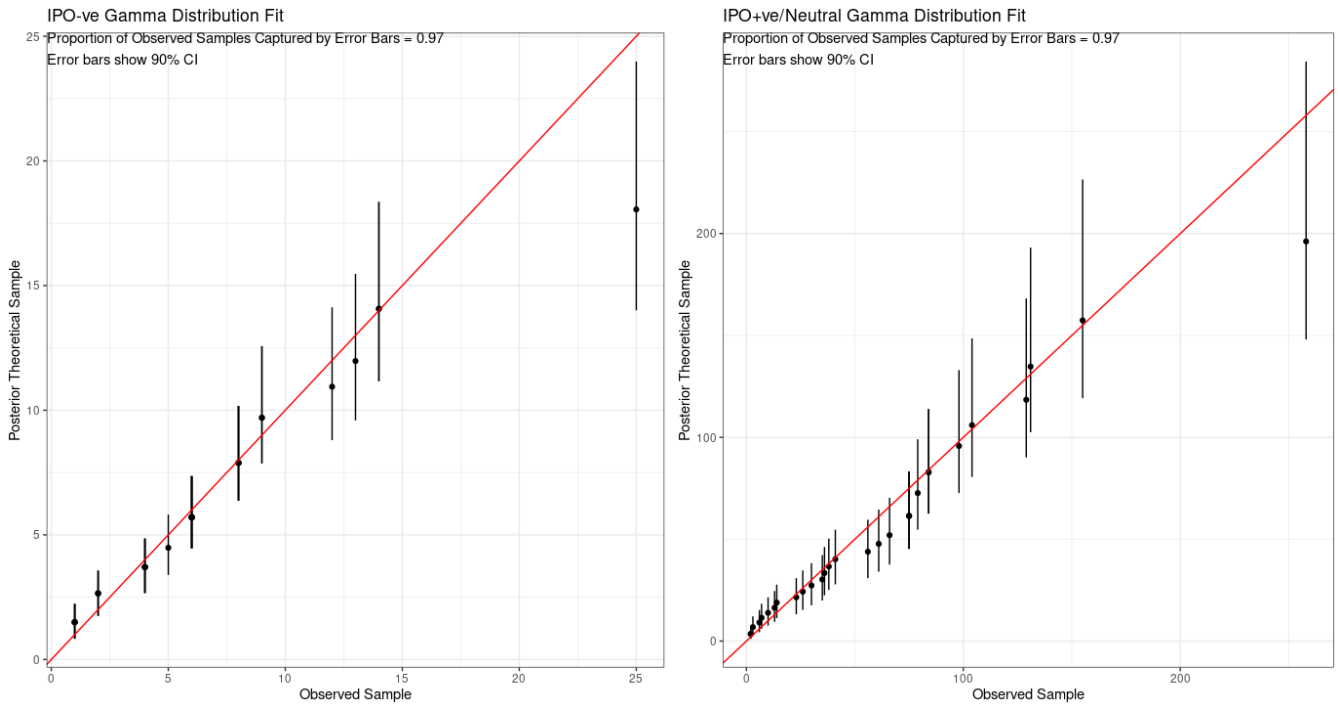

Figure 9-3: Posterior of Gamma distribution fitted to IPO run-lengths.

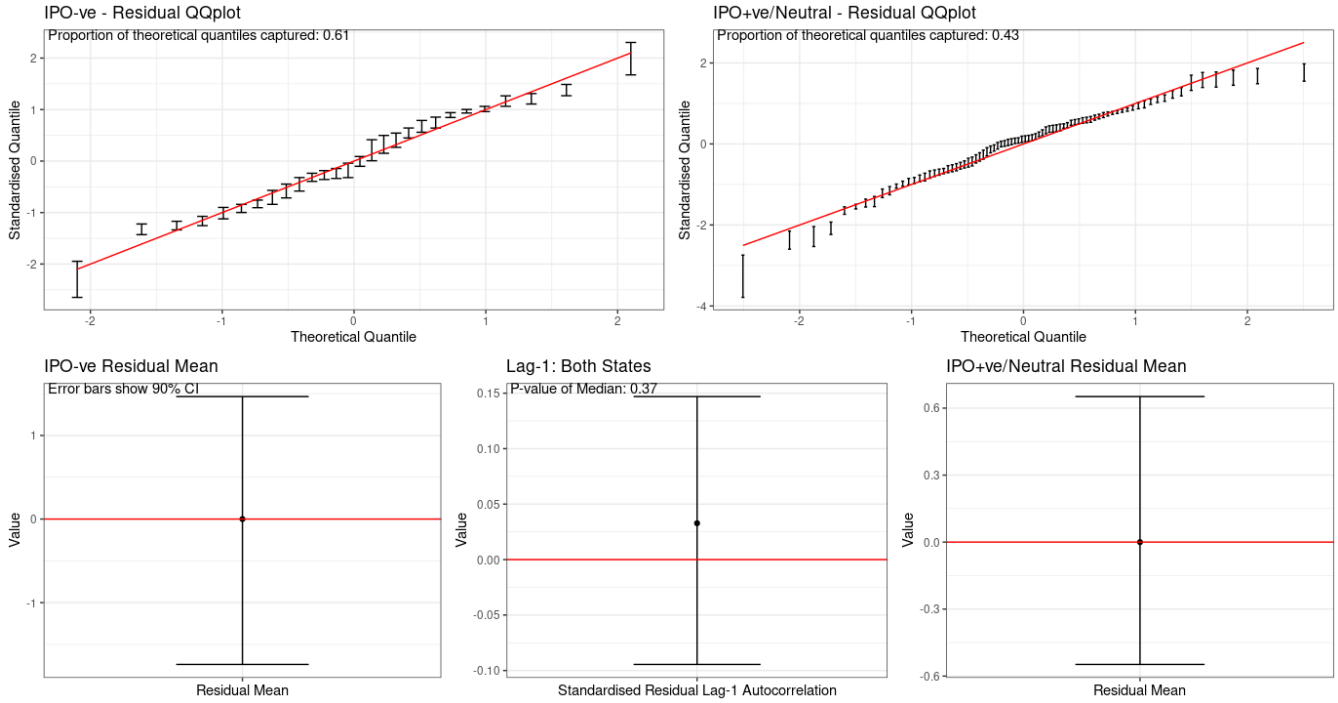

Figure 9-4: CIMSS residual diagnostics for different IPO phases.

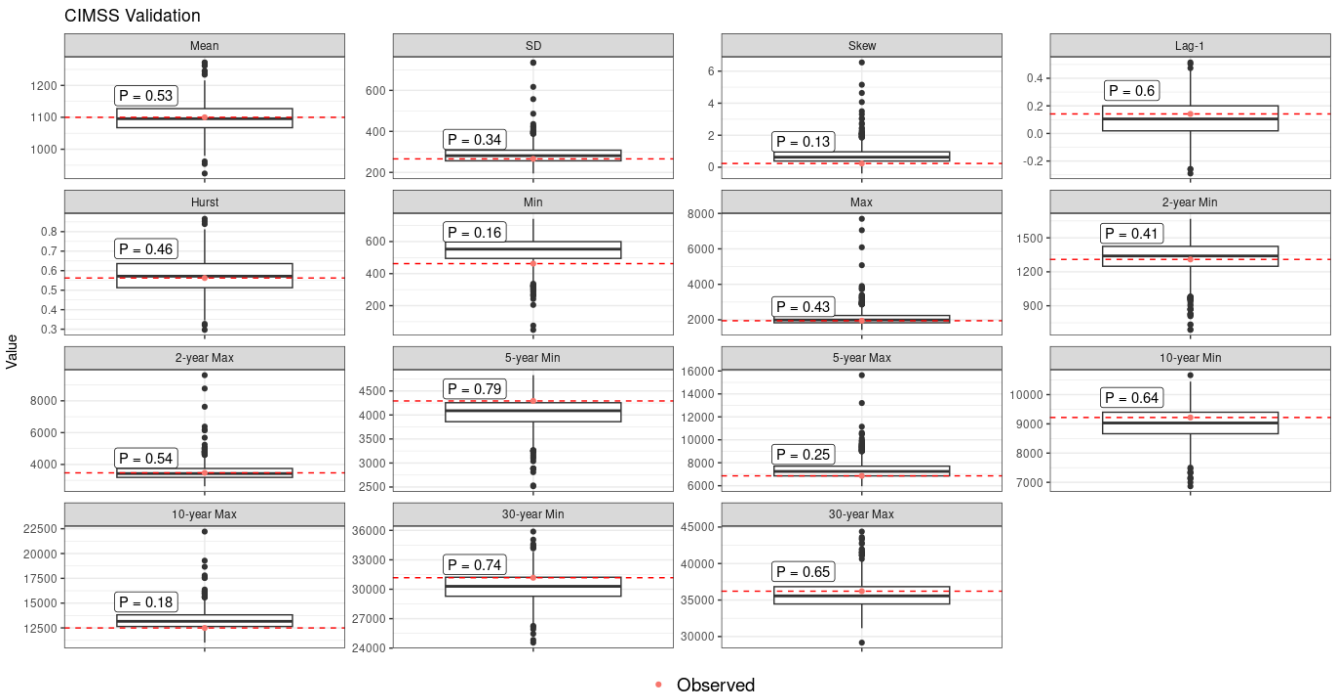

Figure 9-5: CIMSS statistics validated against Williams River rainfall.