

UCLA

UCLA Electronic Theses and Dissertations

Title

Improving Power and Uncertainty Quantification in Spatiotemporal Climate Applications

Permalink

<https://escholarship.org/uc/item/9xb8j0n8>

ISBN

9798247928560

Author

McEvoy, Kyle Roberts

Publication Date

2026-05-29

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Improving Power and Uncertainty Quantification in Spatiotemporal Climate Applications

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Statistics

by

Kyle Roberts McEvoy

ABSTRACT OF THE DISSERTATION

Improving Power and Uncertainty Quantification in Spatiotemporal Climate Applications

by

Kyle Roberts McEvoy

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2026

Karen A. McKinnon, Chair

Performing statistical analyses with climate data requires accounting for many types of variability. This variability is complex and spatiotemporal in nature, and the variability often integrates overlapping processes at multiple spatial and temporal scales. While this makes climate a challenging setting for applying statistical methods, we can use scientific expertise to properly account for and leverage these spatiotemporal structures in our analyses. This dissertation develops statistical methods for performing large scale multiple hypothesis testing and uncertainty quantification in the climate setting. Chapter 1 introduces the climate setting and some of the particular challenges that methods face in this setting. Chapter 2 focuses on multiple hypothesis testing in a common regression setting in climate, where a univariate response variable is regressed against a spatial grid of covariates. The false discovery rate (FDR) control framework provides a practical method for controlling for multiple hypothesis testing, but popular FDR controlling methods, such as Benjamini-Hochberg, exhibit weak power when applied to data that is highly spatially correlated, as is often the case with climate data. This limitation motivates a proposed spatial smoothing method using learned local covariance structures to smooth covariate data as pre-processing step before the analysis. We demonstrate that our method still controls the FDR rate empirically, while substantially increasing power in simulation studies. Then, we apply the method on regressions of central US temperatures against sea surface temperatures. In Chapter 3, we consider the problem

of quantifying uncertainty in estimates of precipitation extremes over the conterminous United States (CONUS). Due to the inherently rare nature of extreme events, extreme quantiles are difficult to estimate, and this process is made more challenging by the temporal structure of precipitation data, which is influenced by modes of variability. We expand upon observational large ensemble methods in the literature to generate synthetic spatiotemporal fields of precipitation over CONUS. These synthetic fields incorporate variability related to large scale modes of climate variability as well as unrelated short term variability to produce fields with realistic precipitation variability. This method is validated using data from climate model initial condition large ensembles, and we use these ensembles to estimate uncertainty in estimates of monthly extreme quantiles of the precipitation distribution across CONUS. In Chapter 4, we apply this observational large ensemble approach with regression models to better quantify variability in Colorado River streamflows at Lees Ferry. Streamflows of the Colorado River are heavily influenced by temperature and precipitation over the Upper Colorado River Basin. In order to understand how the variability in temperature and precipitation can influence the variability in streamflows, we apply the observational large ensemble to generate synthetic precipitation and temperature records for the Upper Colorado River Basin. And, using these records with the regression models, we produce a synthetic ensemble of streamflows. This synthetic ensemble is used to estimate annual variability for streamflows that properly integrate the variability of temperature and precipitation records. Then, we use estimates of future warming to explore how streamflows might change by 2050. Finally, Chapter 5 concludes the thesis with a discussion of these methods and future work.

The dissertation of Kyle Roberts McEvoy is approved.

Andrew N. Poppick

Frederic Paik Schoenberg

Mark S. Handcock

Karen A. McKinnon, Committee Chair

University of California, Los Angeles

2026

Contents

Abstract	ii
List of Figures	vii
Acknowledgments	ix
Vita	x
1 Introduction	1
2 Increasing Power When Controlling Multiple Hypothesis Testing with Climate Data via Covariate Smoothing	6
2.1 Introduction	7
2.2 Methods	10
2.3 Simulation Study	17
2.4 Real Covariate Data and Application	28
2.5 Summary, Discussion, and Conclusions	32
2.6 Data and Code Availability	35
3 Quantifying Uncertainty in Extreme Precipitation Quantiles over the Conterminous United States using an Observational Large Ensemble	36
3.1 Introduction	37
3.2 Methods	39
3.3 Validation	45
3.4 Results	50
3.5 Discussion	57
4 Quantifying Interannual Variability in Colorado River Flows at Lees Ferry	60
4.1 Introduction	61
4.2 Methods	64
4.3 Results	74
4.4 Discussion and Conclusion	84
5 Discussion	88
Appendices	92

A Supplemental Materials Chapter 2	93
B Supplemental Materials Chapter 3	99
C Supplemental Materials Chapter 4	106

List of Figures

2.1	Simulation Results	23
2.2	Sea Surface Temperature MLE parameters	29
2.3	Inference with Resampled Noise	30
2.4	SST Application Rejections	32
3.1	Precipitation Model R^2 Values	41
3.2	EOF Comparison	45
3.3	Precipitation Composites	47
3.4	Quantile Variability Validation	49
3.5	Confidence Intervals for 0.95 Quantiles	51
3.6	Confidence Intervals for 0.05 Quantiles	52
3.7	Precipitation Composites with Multiple Modes	52
3.8	A Case Study of Modes and Extremes	54
3.9	California Precipitation and Dry Stretches	56
4.1	Upper Colorado River Basin Elevation and Average Precipitation	67
4.2	Basin Temperature and Precipitation Correlations	69
4.3	Basin Monthly Temperature Trends	74
4.4	Flow Model Regression Diagnostics	76
4.5	Trends in Temperature Forced Response	78
4.6	Upper Colorado River Basin ENSO Composites	80
4.7	Upper Colorado River Basin PDO Composites	81
4.8	Upper Colorado River Basin PNA Composites	82
4.9	Upper Colorado River Basin NAO Composites	83
4.10	Modes and Synthetic Flows	84
4.11	Natural Flow Variability and Trends	85
4.12	Three-year Flow Distributions	86
A.1	Null t -Statistics	93
A.2	True Rejection Scatterplots	94
A.3	Double-Dipping Sensitivity	95
B.1	Univariate IAAFT Lagged Correlations	99
B.2	Multivariate IAAFT Lagged Correlations	100
B.3	Quantile Variability Validation for 0.05 Quantile	101
B.4	Quantile Variability Bias and λ	102

B.5	Values of λ in GPCC and CESM2-LE	102
B.6	California Quantile Validation	103
B.7	California Conditional Quantile Validation	104
B.8	Validation of Dry Stretches	105
C.1	Bootstrap Confidence Interval for Mean Synthetic Flow	106
C.2	Bootstrap Confidence Interval for 0.975 Quantile	107
C.3	Bootstrap Confidence Interval for 0.025 Quantile	108
C.4	Comparing Bootstrap Prediction Intervals with Interannual Variability . . .	109

Acknowledgments

First and foremost, I would like to acknowledge my incredible wife, Lauren Woodfill, without whom this dissertation would not have been possible. Thank you for your love and support on this journey! To my parents, thank you for encouraging my curiosity as a child.

I would also like to thank my advisor, Karen McKinnon, for all of her guidance and mentorship throughout the PhD, including the times when I was quite stressed out. I learned a lot from you that made this PhD possible. And, I would like to thank Andy Poppick for his statistical guidance and, especially, for his mentorship during winter and spring of 2023 that helped me continue to make progress toward this goal. Thank you to Mark Handcock and Rick Schoenberg for being on my thesis committee and for your feedback on my work.

Finally, thank you to Nathan Hwangbo, Mady Beckner, and Talia Anderson for hanging out in the office and talking through research problems. You made this process a lot more fun.

Chapter 2 is a version of Kyle R. McEvoy and Karen A. McKinnon. Increasing Power When Controlling Multiple Hypothesis Testing with Climate Data via Covariate Smoothing. *Journal of Agricultural, Biological and Environmental Statistics*, May 2026. doi:<https://doi.org/10.1007/s13253-026-00738-5>.

This work was supported by funding from the Packard Foundation.

Vita

Education

University of California, Los Angeles	September 2021 - June 2026
Ph.D. in Statistics (Expected)	
University of California, Berkeley	August 2008 - May 2012
B.S. in Molecular Environmental Biology	

Fellowships and Awards

LEAP Summer Momentum Fellowship	May - August 2024
Graduate Dean's Scholar Award	September 2021 - August 2023

Chapter 1

Introduction

Statistical analysis in climate science is complicated by short observational records, high-dimensionality, non-stationarity, and complex spatiotemporal correlation structures. These factors make application of standard statistical methods more difficult and require novel solutions that respect the unique structure of this data. This dissertation proposes methodological approaches for handling some of these complexities, allowing for more robust statistical inference in the climate domain.

Climate records are often short, lasting under 100 years, though some longer records begin in the 19th century. While some even older records based on ships logbooks do exist (García-Herrera et al., 2005), these records are often sparse through space and time, making them hard to analyze. Even over more recent time frames, such as the early 20th century, generating spatially and temporally complete records requires various statistical infilling procedures that result in their own estimation biases and uncertainties (e.g. Huang et al., 2017). These uncertainties increase as we head further back in time. Newer satellite-based datasets with increased spatial coverage only extend back about 50 years as satellites became more common. While these record lengths would be sufficient for many statistical problems, the structure of climate data poses unique challenges.

One challenge with climate data is that the number of locations considered in analyses

is often very large relative to the shortness of the records. For example, in regression or correlation analyses where the relationship between a univariate response variable and a covariate measured at many locations is examined there might be only 100 years of data with thousands of locations (for example Ting and Wang, 1997; McKinnon et al., 2016). This is a difficult high-dimensional inference problem with a relatively small sample. While some climate scientists focus on predicting the outcome variable with regularized regression methods (Sippel et al., 2024), the connections between specific locations, or spatial patterns of locations, and the response variable are often of scientific interest, requiring an alternative approach.

One strategy to resolving this challenge is to perform a dimension reduction method, such as principal components analysis, termed empirical orthogonal functions analysis in the climate literature, on the covariates before regression. This method can lead to scientifically interesting results, especially if the resulting spatial principal components can be linked to known modes of variability in climate science, such as the El Niño-Southern Oscillation. However, if individual locations are of interest, the problem is often resolved by performing isolated marginal analyses at each location and living with the high-dimensionality. Performing hypothesis tests on these many marginal locations naturally leads to issues with multiple hypothesis testing, which have not always been corrected for in the climate literature (Wilks, 2016). Even when corrections for multiple testing are used in this spatially correlated setting, the corrections can result in weak power. In Chapter 2 of this work, we propose a method for leveraging the spatial information in the covariate data to improve the power of this multiple testing approach with the potential to result in more discoveries while still empirically controlling for false discoveries. In this way we use the strong underlying structure of the climate data to our advantage in inference. This chapter has been published in the *Journal of Agricultural, Biological and Environmental Statistics* (McEvoy and McKinnon, 2026).

Another challenge in climate is that understanding the data requires a deep understanding of complex spatiotemporal variability. Climate scientists often decompose spatiotemporal

fields into two components: the forced response and the internal variability of the system. The forced response functions like an overall mean state and comes in response to a mixture of “natural” drivers, such as the amount of incoming solar radiation or volcanic eruptions, and anthropogenic drivers of the system, such as greenhouse gas emissions. The internal variability of the climate, fluctuations around this mean state, can result in dramatically different climate records. Even over longer scales, on the order of a century or more, this internal variability can hinder inference with climate records, increasing uncertainty in trends (McKinnon et al., 2017) or the estimation of the impacts of important climate modes of variability like the El Niño-Southern Oscillation (Deser et al., 2018).

Understanding this internal variability is further muddled by the complex interplay between short term atmospheric circulation events and longer term processes that relate to multi-year, decadal, or even multi-decadal shifts in sea surface temperatures that result in downstream circulation changes (Dong et al., 2018; Deser et al., 2004; Frankcombe et al., 2010). Since these climate modes of variability are understood to have important downstream impacts on climate variables of interest, such as temperature and precipitation, fully understanding the variability of the climate system requires accounting for the frequency, persistence, and impacts of the climate modes.

Climate models are one useful tool for analyzing the forced response and the internal variability of the system, but these models have their own limitations. Earth system models, such as the Community Earth System Model version 2 (CESM2) (Danabasoglu et al., 2020), solve physical equations that govern the behavior of the climate system in order to simulate climate outcomes. The forced drivers of the system, such as solar radiation, anthropogenic greenhouse gas emissions, and aerosols can be specified and the model can be run to sample from the model simulation of the internal variability. These models can be run for hundreds of years, but the non-stationarity of the climate with respect to the forced response and the significant contribution of internal variability poses challenges.

One way that climate scientists handle these problems is to use initial-condition large

ensembles to sample from the internal variability multiple times (Deser et al., 2020). In initial condition large ensembles, small atmospheric temperature perturbations are added to the initial conditions of a climate model before it is run, and, due to the highly chaotic nature of the climate system, these small perturbations quickly lead to decorrelation between the model runs. Despite the intense computational resources required, this process can be repeated many times leading to initial condition large ensembles, such as the CESM2 large ensemble (Rodgers et al., 2021). Using these large ensembles, climate scientists can use the ensemble mean to estimate the forced response and use the variations around that mean to estimate the internal variability of the system. While a powerful tool, climate models rely on coarse spatial scales and approximations that can cause biases in the simulated climate (Danabasoglu et al., 2020). This suggests a need for appropriate estimates of internal variability that are observationally based.

One approach in the literature to derive and sample from observationally-based estimates of the variability is the Observational Large Ensemble approach (McKinnon et al., 2017; McKinnon and Deser, 2018, 2021). This method involves estimating climate responses to the modes of variability, and then uses those estimates along with statistical resampling methods to generate synthetic spatiotemporal records of climate variables with realistic variability. In Chapter 3 of this work, we refine the method of McKinnon and Deser (2021), which applied the Observational Large Ensemble to precipitation, and use this approach to quantify uncertainty in extreme precipitation quantile estimates over 100 year records. By using this approach, we avoid the precipitation biases of climate models, while still developing a meaningful sample of the system’s internal variability. An important refinement over the method of McKinnon and Deser (2021), is that synthetic climate modes are generated in a multivariate fashion, preserving a more realistic cross-correlation structure across the modes. Since the modes drive important features of the climate response, this leads to more realistic synthetic records.

Finally, in Chapter 4 we explore how the Observational Large Ensemble method can be

used to generate synthetic records of Colorado River water. Streamflows of the Colorado River show strong interannual variability, but have also been shown to have a key dependence on temperature and precipitation over the Colorado River Basin (Vano et al., 2012; Udall and Overpeck, 2017; Lehner et al., 2019; Hoerling et al., 2024). So, in order to meaningfully account for the internal variability in these streamflows, it is important to generate realistic synthetic records of temperature and precipitation. We do so in a spatiotemporal fashion, allowing for analyses of the relationship between climate modes and the pattern of temperature and precipitation responses over the basin. The spatiotemporal records preserve important correlations between temperature and precipitation and allow for meaningful estimates of internal variability in the streamflows. Finally, we combine these estimates of internal variability with estimates of the forced response to explore how this important water source’s variability could change by 2050.

Throughout this work, solutions to various problems in statistical inference and uncertainty quantification in climate applications are explored. First, the spatial structure of climate data is leveraged as an asset, allowing for the borrowing of information across space to increase power in multiple hypothesis testing problems. Next, resampling techniques are explored that enable meaningful estimates of the internal variability of spatiotemporal climate fields based on observational data. These techniques are used to explore precipitation extremes over the conterminous United States and the internal variability of Colorado River streamflows with significant implications for water resources planning. By leveraging novel statistical approaches, we can improve the quality of climate analyses.

Chapter 2

Increasing Power When Controlling Multiple Hypothesis Testing with Climate Data via Covariate Smoothing

Whenever many hypothesis tests are performed simultaneously, the risks of falsely rejecting some null hypotheses must be accounted for in statistical analyses. Methods controlling the possibility of any false rejections become very conservative as the number of tests grows, which motivated the False Discovery Rate (FDR) control framework that limits the average proportion of rejections that are false. However, FDR control methods, such as Benjamini-Hochberg, often have undesirably low power when working in spatially correlated settings. Here we focus on regression problems in large climate datasets involving a univariate response variable and spatially gridded covariate data. Given that nearby locations are more likely to share a signal, we propose a method in which the covariates are spatially smoothed with locally fitted covariance functions before applying the hypothesis testing procedure. Simulation results show that by using Benjamini-Hochberg with this smoothed data, power

The contents of this chapter are published in the *Journal of Agricultural, Biological and Environmental Statistics* (McEvoy and McKinnon, 2026).

is increased while still maintaining empirical FDR control. We then apply the technique to real sea surface temperature covariates with both simulated and real responses. The real data example, linking sea surface temperatures to July temperatures in the central US, demonstrates that the smoothing procedure allows for signals to be identified when none pass traditional Benjamini-Hochberg control. The proposed procedure provides climate scientists with a simple yet powerful tool for extracting more information from their data, while correcting for multiple hypothesis testing.

2.1 Introduction

Multiple hypothesis testing is a fundamental problem when working with large spatiotemporal datasets, which is typical in the climate sciences. A motivating and common type of study involves investigating a relationship between a univariate response variable and a spatial grid of covariates that is quantified through correlation, regression, or composite maps (averages conditional on certain values of the response variable). The resulting maps are then assigned an estimate of significance at each gridpoint. Because the number of predictors often far exceed the length of the time series, the analysis is typically done marginally. This approach has been used to identify, for example, sea surface temperature predictors of summertime precipitation (Ting and Wang, 1997), heatwaves (McKinnon et al., 2016), and the state of the stratospheric polar vortex (Kretschmer et al., 2017). Since the number of hypotheses being tested in this way can number in the thousands, failing to control for multiple hypotheses will result in many spurious rejections with coherent spatial structure that can falsely suggest interpretability. Researchers who are not familiar with correcting for multiple hypothesis testing can take these rejections at face value, drawing incorrect conclusions, and the volume of spurious rejections can drown out interesting results. Correcting for multiple hypothesis testing can resolve many of these problems, but can be underpowered in cases when the underlying covariate field is spatially correlated (Wilks, 2016), as discussed below.

Early work within climate science proposed to address this issue via quantifying “field significance”, wherein the full spatial field is identified to contain a large enough number of significant locations that they are unlikely to have arisen by chance (Livezey and Chen, 1983). However, this method has low power and would typically preclude finding localized regions of strong signals (Wilks, 2016). An alternative is to control for the False Discovery Rate (FDR), defined as the fraction of “discoveries”, or rejections of the null, that are false (i.e. the null hypothesis is true). In cases where the p -values generated by the hypothesis testing procedure are independent from each other, the FDR can be controlled by a simple procedure based on ordering the p -values of the tests (Benjamini and Hochberg, 1995); this is often termed the “Benjamini-Hochberg method”. Benjamini and Yekutieli (2001) then showed that the Benjamini–Hochberg procedure controls the false discovery rate for test statistics satisfying the “positive regression dependence on a subset” (PRDS) condition. However, this condition is often violated in settings with correlated test statistics, particularly for two-sided tests. Finally, they introduced a modified procedure that guarantees FDR control under arbitrary dependence, but it requires substantially more stringent p -value thresholds, leading to reduced statistical power in practice.

Although there are not theoretical guarantees of FDR control using the standard Benjamini-Hochberg method when the tests are correlated, simulation studies generally show that FDR control generally holds empirically (Farcomeni, 2006; Wilks, 2016), at least for differences-in-mean tests. That said, the control was generally conservative, and the extent to which it reduced power depended on the correlation structure in the data.

Prior attempts to increase power using FDR in the correlated setting have had mixed success. Motivated by correlated climate data, Yekutieli and Benjamini (1999) proposed a resampling-based approach to improve power. While this approach led to small gains in power, it also required more computation (Ventura et al., 2004, and Table A.1), so it would be unlikely to be used as a straightforward replacement for the standard procedure. Ventura et al. (2004) then sought to improve the power further by increasing the target FDR control level

by a scaling factor related to the estimated proportion of true null hypotheses. Their method successfully increases the power, but in their simulations they found that when over 30% of the alternative hypotheses were true and the test statistics were highly correlated there were some failures in FDR control. It has also been suggested to increase power in the correlated setting by clustering the covariates, thereby sharing information across space (Benjamini and Heller, 2007); however, when the data are spatially-smooth, as is often the case with geophysical variables, the clustering will inherently induce artificial boundaries. Furthermore, the clustering process could lead to undesirable effects at the boundaries between regions, where a location may or may not have been rejected depending on which cluster it had been assigned.

One key challenge in the climate regression setting is that the sample size, n , is almost always much smaller than the number of covariates, p . For example, a typical gridded dataset could have on the order of tens of thousands of points, while a typical climate dataset is on the order of 50-100 years long. This precludes the use of regression methods that try to consider the joint relationship between the response and the full set of covariates, such as the knockoffs method proposed in Barber and Candès (2015).

A number of other methods have been proposed to try to leverage the spatial structure of the testing problem to improve power. Some of these methods operate under the signal-plus-noise model, where spatially correlated noise is added to a mean field (Sun et al., 2015; Zhang et al., 2011). But applying this model directly to the regression coefficients, as in Sun et al. (2015), requires defining a reasonable specification of the noise model for fields of regression coefficients, which may not be straightforward in applications by climate scientists. Even if we do not fully trust that the regression statistics arise from the signal-plus-noise model, methods designed under this model that operate directly on the p -values, such as the FDR_L method of Zhang et al. (2011), could still be applied. However, in applying the FDR_L method to regression problems in simulation studies we found that the method made many true rejections but lost FDR control (Table A.2).

Other commonly used spatial approaches for increasing power assume the two groups model (e.g. Tansey et al., 2018), where test statistics are assumed to arise from a mixture model of a null density and a signal density. In the climate setting, where we might expect the distribution of test statistics to vary smoothly in space, interpolating between two distinct densities may be unable to represent the full range of test statistic distributions at different locations. Furthermore, when Tansey et al. (2018) applied their method to a spatially varying signal in simulation studies, they found small failures of FDR control.

In order to improve power in the spatially correlated setting, here we propose a localized spatial smoothing pre-processing step for the covariates. The smoothed covariate data can then be used in standard regression approaches combined with Benjamini-Hochberg for empirical FDR control. This method has few assumptions, is not computationally demanding, and can be easily incorporated into existing hypothesis testing procedures. These benefits make the procedure a useful tool for researchers to extract more information from datasets, while empirically controlling the FDR.

The remainder of the paper proceeds as follows. Section 2.2 defines the setting and introduces the smoothing method. Section 2.3 evaluates the method using simulation studies. Section 2.4 investigates the method with synthetic responses and real sea surface temperature (SST) covariates and then applies the method to a regression of Central United States temperatures on SST covariates. We conclude in Section 2.5 with a discussion of our findings.

2.2 Methods

2.2.1 Model and Assumptions

We focus on the linear relationships between a single univariate response variable, $Y_i \in \mathbb{R}$, and a spatial grid of covariates, $\{X_{ij}\}$, $X_{ij} \in \mathbb{R}$, where $i \in \{1, 2, \dots, n\}$ indexes the sample through time and $j \in \{1, 2, \dots, p\}$ indexes the location, for some $n, p \in \mathbb{N}$. Under our model, Y_i and each of the X_{ij} are random variables. Throughout, it is assumed that the covariate vectors,

$X_{i\cdot} = (X_{i1}, X_{i2}, \dots, X_{ip}) \in \mathbb{R}^p$, follow a multivariate normal distribution, $X_{i\cdot} \stackrel{iid}{\sim} N_p(0, \Sigma)$, with mean vector 0 and a covariance matrix, $\Sigma \in \mathbb{R}^{p \times p}$, that can be locally approximated by a spatial covariance function. In practice, we work with covariates that have been standardized by removing the mean and dividing by the standard deviation. The equations presented here assume zero-centered covariates but could be modified to include intercept terms if desired. We consider a model for the response variable of the form

$$Y_i = \beta^T X_{i\cdot} + \varepsilon_i, \quad (2.1)$$

where $\beta \in \mathbb{R}^p$ is a coefficient vector and $\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma_\varepsilon^2)$ is independent of the set of covariate vectors. These assumptions jointly imply that the n random vectors, $(X_{i1}, X_{i2}, \dots, X_{ip}, Y_i)$, are i.i.d. samples. The inferential goal is to select a subset of the locations where there are likely to be significant linear relationships under the FDR control framework.

In the climate regression setting of our work, the X_{ij} are often standardized anomalies of the same variable, for example sea surface temperature (SST), observed across a latitude-longitude grid. These anomalies typically have strong spatial correlations between nearby locations, and in many climate applications the number of grid locations, p , may number in the tens of thousands. We expect β to have many zero entries because the response variable is likely to be sensitive to only some of the covariates. For example, for SST anomalies, only specific regions of the ocean influence the downstream climate variable of interest.

The assumption that the samples are i.i.d. implies that there should not be seasonality, autocorrelation, or a trend in the samples. This set-up is reasonable for detrended climate variables with limited inter-annual memory when focusing on measurements made during the same month each year, but this is a strong assumption in general. In cases where there is autocorrelation in the error terms the analysis should be adjusted.

A common method in climate science is to perform p simple ordinary least squares linear regressions in parallel, regressing the response variable Y_i on each covariate X_{ij} in isolation.

That is, fitting the regression model

$$Y_i = \kappa_j + \gamma_j X_{ij} + \nu_{ij}, \quad \nu_{ij} \stackrel{iid}{\sim} N(0, \sigma_j^2), \quad (2.2)$$

for $\kappa_j, \gamma_j \in \mathbb{R}$, $i = 1, \dots, n$ and for each $j = 1, \dots, p$, with ν_{ij} independent of X_{ij} . Note that with the multivariate normality of the covariates and the independence and normality of the errors, model (2.2) is compatible with model (2.1) for all j . For each j , let $\Sigma_{\cdot j} \in \mathbb{R}^p$ be the column of Σ containing the covariances between X_{ij} and other covariates. Then, for all $j \in \{1, 2, \dots, p\}$, $\text{Cov}(X_{ij}, Y_i) = \text{Cov}(X_{ij}, \beta^T X_i + \varepsilon_i) = \beta^T \Sigma_{\cdot j}$, using the independence of the errors from the covariates. The regression coefficient, γ_j , is given by

$$\gamma_j = \frac{\text{Cov}(Y_i, X_{ij})}{\Sigma_{jj}} = \frac{\beta^T \Sigma_{\cdot j}}{\Sigma_{jj}}. \quad (2.3)$$

The estimated coefficients, $\hat{\kappa}_j$ and $\hat{\gamma}_j$, are fit using ordinary least squares regression, and the standard t -statistic for testing the null hypotheses $H_0^j : \gamma_j = 0$ is calculated as

$$T_j = \frac{\hat{\gamma}_j}{\widehat{\text{se}}(\hat{\gamma}_j)}. \quad (2.4)$$

In this work, we do not treat the covariates X_{ij} as fixed; however, with a linear model and homoscedastic errors that are uncorrelated with X_{ij} , T_j will asymptotically converge to a normal distribution under the null hypothesis (a standard result, see e.g. Wooldridge (2010)). Therefore, the standard two-sided t -test with $n - 2$ degrees of freedom will be asymptotically valid test for testing the null hypotheses $H_0^j : \gamma_j = 0$ against the two-sided alternatives, $H_1^j : \gamma_j \neq 0$. As alternatives to the asymptotic t -test, either a permutation testing approach or a resampling-based method, such as the one in Yekutieli and Benjamini (1999), could be applied with our method. In finite samples these methods could potentially have better control of the type-I error rate than the asymptotic t -test. However, these methods are also more computationally demanding than the t -test, especially with thousands of locations.

If the assumption of homoscedasticity of errors is unreasonable due to a shift in the error variance or other heterogeneities in time, then the i.i.d. assumption could be relaxed by designing an appropriate permutation test, assuming sufficient knowledge about the structure of the heteroscedasticity.

2.2.2 Smoothing Method

In order to improve the power of the standard Benjamini-Hochberg procedure in this setting, we propose a localized smoothing procedure as a pre-processing step before hypothesis testing using a windowed, weighted-average of the covariate field. Under standard assumptions for spatial data, covariances between locations decline monotonically as the distance between two locations increases. Therefore, under Equation (2.3), the contribution of each nonzero element of β to γ_j will decline toward zero as distance between the nonzero locations and location j increases. This matches the expectation for geophysical data, where we expect that the signal (significant γ_j values) will exhibit spatially coherent clusters related to the covariance structure of the covariate field. Smoothed covariates that put more weight in the average on nearby signal locations will tend to exhibit stronger covariance with the predictand than they would before smoothing. This will increase the number of “discoveries” for locations that are false nulls but have a weak signal in isolation. Conversely, we do not expect a large increase in the number of false discoveries because locations that exhibit small p-values by chance will then be averaged with many neighboring locations that are unrelated to the predictor.

To describe the method, let x_{ij} , $1 \leq i \leq n$, $1 \leq j \leq p$, be the observed covariate data which has been standardized by subtracting the mean and dividing by the standard deviation. At location $k \in \{1, 2, \dots, p\}$, a local window of the data, centered at k , is selected. Now, let $W_k \subset \{1, 2, \dots, p\}$ be the subset of indices for locations that fall within this local window. Then, for any set of non-negative weights of $\{w_m^k\}_{m \in W_k}$, with one weight for each location in

W_k , we can define the weighted average at location k , x_i^k , as

$$x_i^k = \frac{\sum_{m \in W_k} w_m^k x_{im}}{\sum_{m \in W_k} w_m^k}. \quad (2.5)$$

This weighted average can be used in place of the original data in our regressions.

The weights are determined as a function of the local covariance structure in the covariates. This structure can vary substantially across space, so we fit covariance models in moving windows, within which we can reasonably assume that the covariance structures are stationary (Kuusela and Stein, 2018; Higdon, 1998; Haas, 1990). Once fit, the spatial covariance function gives a measure of how much similarity we expect between the gridboxes in the window. The choice of window size depends on the application of interest. Ideally, the window size needs to be large enough that the spatial correlation structures can be captured, while small enough that the local stationarity assumption is reasonable.

Since we are focused on an application to SST anomalies, we choose a local $20^\circ \times 20^\circ$ longitude-latitude window around each location. For points that are closer than 10° latitude to the northern and southern borders of our spatial region ($60^\circ N$ and $60^\circ S$, respectively), the window is truncated at the boundary. In the east and west directions, the window wraps around the earth's surface. In our SST application, there are no measurements for gridboxes on land, but we do not truncate the window at these land boundaries. So, for example, the window can contain seas separated by land. This modeling decision could easily be modified by masking ocean gridboxes across a land mass.

While this window size may appear large, windows of this size have previously been used to fit local covariance functions for ocean heat content in Kuusela and Stein (2018). Furthermore, as shown in the fitted spatial models in Figure 2.2, SSTs have long correlation length scales with fitted e -folding distances for many locations over 2,000 km (around 18° at the equator), suggesting that this window size is necessary to accurately capture the covariance structure. Finally, as distance from the target location increases, the covariance in

the model will decrease, naturally decreasing the influence of distant locations in the weighted average.

For the window, W_k , around each gridbox, we model the data as being sampled from a multivariate normal distribution with mean 0, and a covariance matrix Σ_k , arising from a stationary spatial covariance function, f . That is, let X'_i be random vectors of length m_k , consisting of the variables for each of the locations in W_k . Then the model we are fitting is

$$X'_1, X'_2, \dots, X'_n \stackrel{iid}{\sim} N_{m_k}(0, \Sigma_k). \quad (2.6)$$

In order to define f , we first need to describe the distances between locations. Since climate data are often anisotropic, the distance was split into zonal (east-west) and meridional (north-south) coordinates, u_{jk} and v_{jk} respectively, with distances measured using great-circle distance. For u_{jk} the midpoint of the two gridbox latitudes is used when defining the arc on the sphere. Let $q_{jk} = (u_{jk}, v_{jk})^T$ be the coordinate lag vector, and let the anisotropy matrix A be defined as

$$A = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \psi^{-1} \end{bmatrix} \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}, \quad (2.7)$$

where $\theta \in [0, \pi)$, is the rotation angle, $\psi \in [1, \infty)$, is the anisotropy ratio. Then, the anisotropic distance can be calculated as the square root of the quadratic form of q_{jk} with A :

$$d_{jk}(\theta, \psi) = \sqrt{q_{jk}^T A q_{jk}} \quad (2.8)$$

The rotation angle, θ , determines the rotated axis along which the distances are maximized, and $\theta + \pi/2$ determines the direction along which distance values are shrunk the most, with the opposite effects on the resulting covariances in these directions.

When fitting f over the window W_k , all data that fall outside of the window are ignored. Additionally, the specific choice of the family of models, defined by f , will impact the resulting

weights. Therefore, before attempting to employ this method, researchers should carefully choose their spatial covariance function and the size of the window of data needed to fit it, and evaluate some of the local fits of the functions, before proceeding in their analysis. The impact of misspecifying f on the method is explored in the results. The commonly used Matérn family of functions provides a useful starting point for these evaluations, and, in this paper, the Matérn-3/2 covariance function was used for local fitting with the parameterization

$$f(d_{jk}(\theta, \psi), \phi) = \left(1 + \frac{\sqrt{3} d_{jk}(\theta, \psi)}{\phi}\right) \exp\left(-\frac{\sqrt{3} d_{jk}(\theta, \psi)}{\phi}\right), \quad (2.9)$$

where $\phi > 0$ is a correlation length scale parameter. As ϕ increases and the distance, $d_{jk}(\theta, \psi)$, between two locations stays fixed, the variables at those two locations become more correlated. Note that, since we have standardized the observed data, the marginal variance of this function has been fixed to equal 1. A nugget effect could be added to this model, if desired. However, in our applications using the Extended Reconstructed Sea Surface Temperature (ERSST) data of Huang et al. (2017) we found that including a nugget was unnecessary, since the averaging process used to produce the gridded dataset made the small-scale errors in the data negligible.

The local model over W_k of Equation 2.9 has three unknown parameters: a correlation length scale parameter, ϕ_k ; an anisotropy angle, θ_k ; and an anisotropy ratio, ψ_k . These parameters of f are fit by maximum likelihood, using the windowed sample of data, giving estimates $(\hat{\phi}_k, \hat{\theta}_k, \hat{\psi}_k)$ for each gridbox. See Gelfand et al. (2010) and Diggle (2007) for more details on maximum likelihood estimation in the spatial setting. Then, using these maximum likelihood estimates, we calculate the weights as

$$w_j^k = f(d_{jk}(\hat{\theta}_k, \hat{\psi}_k), \hat{\phi}_k),$$

and use these weights in Equation (2.5) to calculate the weighted averages, x_i^k .

To complete the hypothesis testing procedure, the observed response, y_i , is regressed

against x_i^k for each k , as in Model (2.2), the usual regression t -statistics are calculated, as in Equation (2.4), and Benjamini-Hochberg is applied to the p -values of the standard two-sided test to find rejections.

2.3 Simulation Study

Our simulations follow the assumptions outlined in Section 2.2.1. For simplicity, we randomly select a single location, l^* , and we set $\beta_j = 0$ for $j \neq l^*$. Thus, Y_i is sampled from the equation $Y_i = \beta_{l^*} X_{il^*} + \varepsilon_i$. While we use a single point source signal to generate the data, the high correlation in the covariates makes Y_i highly correlated with the locations near l^* . For these nearby locations j , the regression coefficients γ_j will be close to the value of β_{l^*} . So, the effective signal will be spread over a regional cluster of data. For the spatial structure of the simulations, a $2^\circ \times 2^\circ$ latitude-longitude grid was created with latitudes ranging from -60 to 60 and longitudes spanning the sphere. This matches the resolution of our SST data in our application. This spatial grid contains a total of $p = 10,980$ locations at which to sample the covariates.

Equation (2.3) shows that the value of the regression coefficient at any location j , with $j \neq l^*$, will be nonzero if there is a nonzero correlation between the covariates at locations j and l^* . Thus, in order for the hypothesis testing problem to be meaningful we need to choose a covariance function for $\{X_{ij}\}$ that can generate exact zero values. One way to ensure this is to use the covariance tapering method of Furrer et al. (2006) to modify the standard Matérn covariance. For our simulations we choose the Matérn-3/2 covariance function, parameterized as in Equation (2.9) as our primary covariance function and taper this covariance function with Wendland-1 polynomials, as suggested in Furrer et al. (2006),

$$g(d_{jk}(\theta, \psi), \delta) = \left(\max \left\{ 0, \left(1 - \frac{1}{\delta} d_{jk}(\theta, \psi) \right) \right\} \right)^4 \left(1 + \frac{4}{\delta} d_{jk}(\theta, \psi) \right), \quad (2.10)$$

where $\delta > 0$ is the tapering distance. For distances greater than δ , g will equal 0. This gives

a tapered covariance function h :

$$h(d_{jk}(\theta, \psi), \phi, \delta) = f(d_{jk}(\theta, \psi), \phi) \cdot g(d_{jk}(\theta, \psi), \delta). \quad (2.11)$$

h retains the smoothness of f in the sense that both functions are twice differentiable, but h will decrease to 0 more quickly than f with the same value of ϕ .

The n vectors of simulated covariate data were then sampled i.i.d. from a multivariate normal distribution with mean 0 and covariance matrix Σ arising from h . As a baseline scenario, we sampled $n = 50$ samples from the distribution, corresponding approximately to the number of years of data available in the satellite era for many climate datasets. We use the default values $\beta_{l^*} = 1$ and $\sigma_\varepsilon^2 = 1$. This level of noise balances the difficulty of the problem with the relatively small sample size, though for some climate variables larger values may be more realistic.

2.3.1 Simulation Details and Evaluation Metrics

We now turn to evaluating the proposed method. The sample size, n ; noise s.d., σ_ε ; and the parameters of the spatial covariance function were varied across different experiments. The parameter choices can be seen in Table 2.1. The correlation length scale, ϕ , and tapering parameter, δ , were selected to roughly match the covariance functions to east-west and north-south variograms of the SST data at $\pm(20, 30, 40, 50)$ degrees latitude. The windows, W_k , are set to be 20° latitude by 20° longitude, centered at location k and inclusive of the edges. As mentioned previously, this window size matches the windows used in Kuusela and Stein (2018) for ocean temperature data and allows for reasonable fits of the covariance function to SST data, as seen in Figure 2.2. This window size is also used for the real SST covariates. In each experiment, 10 total spatial fields, $\mathbf{X}$, were generated, and for each of these spatial fields, 1,000 different response vectors, $\mathbf{y}$ were generated, giving a total of 10,000 simulations.

In the results we will refer to our method as the smoothing method and the original linear regressions with Benjamini-Hochberg FDR control as the base method. The same response vector $\mathbf{y}$, generated from the original covariates, was regressed against both the smoothed covariate data and the original data from which it was generated (after standardization), and the two-sided t -tests were rejected using the Benjamini-Hochberg method with targeted FDR control level $\alpha = 0.1$.

For each experiment, we evaluated the two methods using four criteria. The first was an estimate of the FDR using the mean FDP across the simulations. The second evaluation criterion was the empirical estimate of the false discovery exceedance (FDX). The FDX was defined in Perone Pacifico et al. (2004) as $\text{FDX}(\alpha) = \mathbb{P}(\text{FDP} > \alpha)$ and measures how frequently the false discoveries would exceed the desired control level in individual realizations of a process. As argued in Sun et al. (2015), the FDX is especially important in the spatial setting where dependence between tests can make the FDP more variable. The third metric was the mean difference in the number of true rejections between the smoothing method and the base method across the simulations. The fourth criterion was the mean sensitivity. The sensitivity is defined as the number of true rejections divided by the total number of false null hypotheses. Sensitivity gives a measure of how powerful the method is in rejecting false nulls. Ideally, a method would have high sensitivity, while maintaining empirical FDR control and keeping FDX low. The results for the experiments are collected in Tables 2.1 and 2.2.

2.3.2 Simulation Results

As seen in Table 2.1, across all simulation experiments, both methods conservatively controlled the average FDP, suggesting that empirical FDR control is maintained under both methods. The $\text{FDX}(0.1)$ values are also similar but in most simulation studies tended to be marginally lower for the smoothing method. Across all simulations, the empirical FDR control is conservative for the target of 0.1, suggesting there may be further gains from a more aggressive procedure. Every experiment performed also showed high specificity, defined as the

method	σ_ε	n	ϕ	θ	$\sqrt{\psi}$	γ	FDP	s.d.	FDX	s.e.
base	1	50	4.0	0	1.00	7	0.065	0.112	0.231	0.004
smooth	1	50	4.0	0	1.00	7	0.061	0.120	0.213	0.004
base	2	50	4.0	0	1.00	7	0.040	0.143	0.099	0.003
smooth	2	50	4.0	0	1.00	7	0.043	0.154	0.094	0.003
base	3	50	4.0	0	1.00	7	0.035	0.167	0.054	0.002
smooth	3	50	4.0	0	1.00	7	0.036	0.167	0.054	0.002
base	1	100	4.0	0	1.00	7	0.069	0.097	0.262	0.004
smooth	1	100	4.0	0	1.00	7	0.067	0.107	0.253	0.004
base	2	100	4.0	0	1.00	7	0.060	0.127	0.196	0.004
smooth	2	100	4.0	0	1.00	7	0.057	0.135	0.176	0.004
base	3	100	4.0	0	1.00	7	0.040	0.148	0.092	0.003
smooth	3	100	4.0	0	1.00	7	0.040	0.151	0.086	0.003
base	1	50	2.0	0	1.00	7	0.066	0.117	0.230	0.004
smooth	1	50	2.0	0	1.00	7	0.062	0.128	0.206	0.004
base	1	50	1.3	0	1.00	5	0.074	0.123	0.257	0.004
smooth	1	50	1.3	0	1.00	5	0.067	0.149	0.199	0.004
base	1	50	1.3	0	1.00	7	0.067	0.116	0.236	0.004
smooth	1	50	1.3	0	1.00	7	0.060	0.134	0.191	0.004
base	1	50	0.8	0	1.00	7	0.069	0.121	0.243	0.004
smooth	1	50	0.8	0	1.00	7	0.055	0.145	0.152	0.004
base	1	50	1.3	1	3.08	7	0.046	0.094	0.167	0.004
smooth	1	50	1.3	1	3.08	7	0.043	0.103	0.154	0.004
base	1	50	1.3	1	3.77	7	0.045	0.092	0.164	0.004
smooth	1	50	1.3	1	3.77	7	0.042	0.099	0.149	0.004

Table 2.1: Results for controlling false discoveries in the simulation studies. The FDX gives the empirical probability estimated across the 10,000 simulations. The FDP gives the mean value across the 10,000 simulations. The s.d. column reports the sample standard deviation of the FDP values across the simulations, the s.e. column reports the estimated standard error of the FDX probability.

number of true non-rejections divided by the total number of true null hypotheses. The mean specificity values were all higher than 0.99 and typically higher than 0.996 for both methods, though the base method had marginally higher values. The distributions of t -statistics at null locations were similar for the base and smooth method, but after smoothing the null statistic distribution has slightly heavier right tails. A typical set of histograms for the two null statistic distributions simulated from the baseline scenario simulation is shown in Figure A.1 of Appendix A. The histograms have similar shapes, but the smooth distribution of null

statistics has a 0.95 quantile value of 2.44 compared to 2.38 for the base distribution. The increased heaviness of the right tail after smoothing could cause a small number of additional false discoveries, but the similarity of the mean FDP between the two methods suggests that these false discoveries are not happening at a higher rate than before.

Greater differences between the two methods can be seen in the difference in the number of true rejections. Figure 2.1a shows histograms of the difference in the number of true rejections between the smoothing method and the base method per simulation, for varying levels of the noise standard deviations and sample size. Simulations were excluded from the figure if neither method produced any rejections, and as the noise variance increases it becomes increasingly hard to find any rejections after applying Benjamini-Hochberg. At all three noise levels, most of the distribution’s mass is concentrated on positive values, showing that the smoothing method improved the number of true discoveries. In the simulations across the three noise levels, the mean increase in number of true discoveries using the smoothing method, given that either method made at least one rejection, ranged from 36.4 at $\sigma_\epsilon = 3$ to 53.1 at $\sigma_\epsilon = 1$ with $n = 50$ and similarly from 41.7 to 78.6 for $n = 100$. The numbers are larger than the mean difference of the figures in Table 2.2, since we are conditioning on either method making a rejection. The true rejections for each method are plotted as a scatterplot in Figure A.2 of Appendix A.

Additionally, with increasing noise a more bimodal shape to the distribution can be seen in Figure 2.1a with the strongest peak near 0 and a secondary peak close to the mean value. This occurs due to the increasing proportion of simulations with very little signal relative to the noise, resulting in both methods increasingly make no rejections or a small number of rejections. Despite this, the long right tail of the distribution of the difference in true rejections leads to a positive overall mean increase in this difference for the smoothing method over the base method. This bimodal shape becomes less pronounced when the sample size n was doubled to 100, as seen in Figure 2.1a.

In all experiments, sensitivities are very low because the denominator, the number of false

method	σ_ϵ	n	ϕ	θ	$\sqrt{\psi}$	γ	true rej.	s.d.	sens.	s.d.	% Δ sens.
base	1	50	4.0	0	1.00	7	205.3	100.2	0.074	0.038	-
smooth	1	50	4.0	0	1.00	7	257.8	127.6	0.093	0.047	25.234
base	2	50	4.0	0	1.00	7	45.0	78.3	0.016	0.029	-
smooth	2	50	4.0	0	1.00	7	61.1	105.6	0.022	0.039	35.426
base	3	50	4.0	0	1.00	7	11.2	40.8	0.004	0.015	-
smooth	3	50	4.0	0	1.00	7	16.9	59.2	0.006	0.021	50.464
base	1	100	4.0	0	1.00	7	360.8	107.5	0.130	0.042	-
smooth	1	100	4.0	0	1.00	7	439.4	132.1	0.158	0.050	21.488
base	2	100	4.0	0	1.00	7	154.7	114.1	0.056	0.042	-
smooth	2	100	4.0	0	1.00	7	196.8	146.8	0.071	0.054	26.925
base	3	100	4.0	0	1.00	7	44.9	78.7	0.016	0.029	-
smooth	3	100	4.0	0	1.00	7	61.4	107.4	0.022	0.039	36.537
base	1	50	2.0	0	1.00	7	108.1	59.5	0.039	0.023	-
smooth	1	50	2.0	0	1.00	7	154.5	88.7	0.056	0.033	42.406
base	1	50	1.3	0	1.00	5	42.2	25.1	0.027	0.017	-
smooth	1	50	1.3	0	1.00	5	70.0	50.3	0.045	0.033	65.033
base	1	50	1.3	0	1.00	7	55.2	33.5	0.020	0.013	-
smooth	1	50	1.3	0	1.00	7	89.9	61.1	0.033	0.023	62.524
base	1	50	0.8	0	1.00	7	22.2	14.5	0.008	0.006	-
smooth	1	50	0.8	0	1.00	7	40.6	35.9	0.015	0.013	82.309
base	1	50	1.3	1	3.08	7	207.1	108.0	0.038	0.020	-
smooth	1	50	1.3	1	3.08	7	280.4	150.3	0.052	0.028	35.463
base	1	50	1.3	1	3.77	7	259.2	133.5	0.046	0.024	-
smooth	1	50	1.3	1	3.77	7	342.1	182.6	0.061	0.033	32.090

Table 2.2: Results for making true rejections in the simulation studies. True rejections and sensitivity give the mean values of each quantity across the 10,000 simulations. The sample standard deviations across the simulations are listed after each metric. The percent change in sensitivity of the smooth method from the base method is listed in the column following the s.d. of sensitivity.

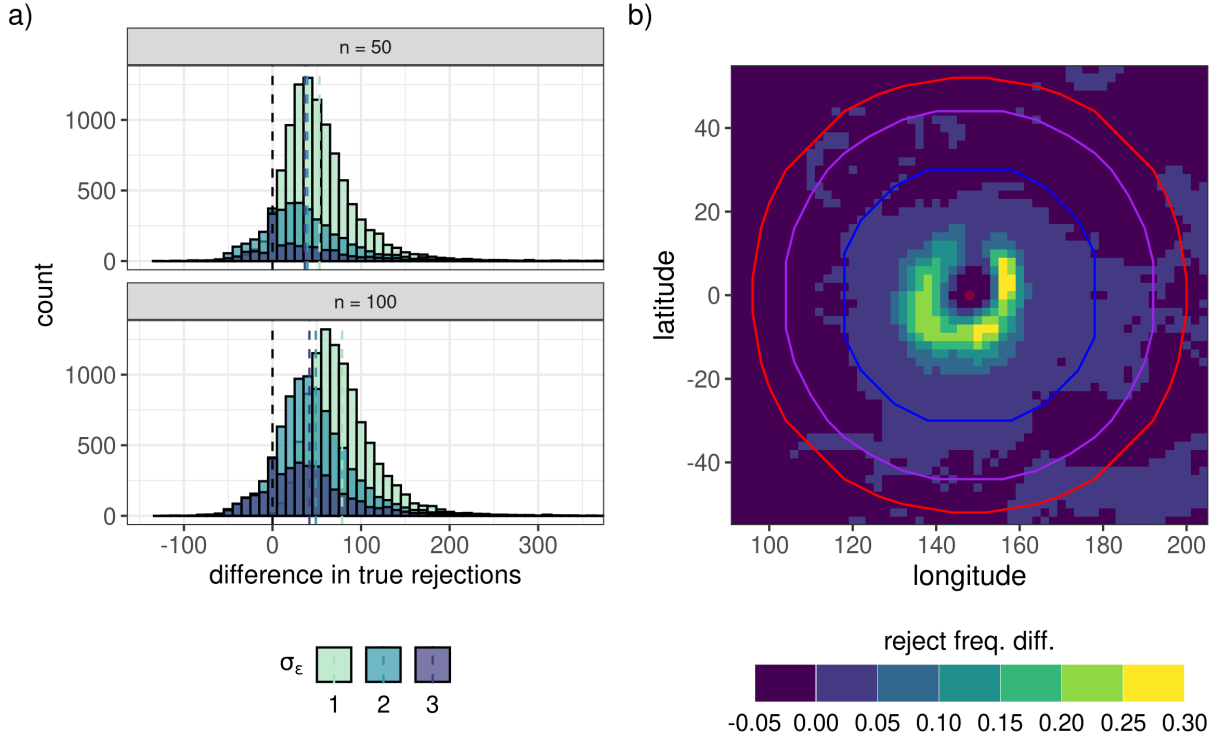


Figure 2.1: a) Histograms of the difference in number of true rejections between the smoothing method and the base method per simulated response, given that either the base or smoothing method resulted in a rejected null hypothesis. Simulation results for the isotropic case are plotted for $\sigma_\varepsilon \in \{1, 2, 3\}$ with separate histograms for $n = 50, 100$. The mean differences are marked by the dashed lines of corresponding colors with zero marked by the black dashed line. For $n = 50$, out of a total of 10,000 simulations, 9,812, 3,916, and 1,507 simulations had at least one rejection for noise levels 1, 2, and 3, respectively. For $n = 100$, the numbers are 10,000, 8,328, and 3,623, respectively. Counts greater than 350 have been truncated from each histogram. b) The difference, smooth minus base, in the relative frequency that each location was rejected across different samplings of the noise. The simulation parameters were $n = 50$, $\sigma_\varepsilon^2 = 1$, $\phi = 1$ using the isotropic case. The covariate field was simulated once, the central point was fixed as l^* for generating y_i and noise was resampled 1,000 times. The red, purple, and blue polygons enclose regions of the spatial process where correlations with the central location are greater than $1 \cdot 10^{-3}$, $1 \cdot 10^{-2}$, and $1 \cdot 10^{-1}$, respectively. The plot has been zoomed in to highlight the central area, but the whole spatial field was generated.

null hypotheses, includes all locations k with $\Sigma_{kl^*} \neq 0$, where l^* is the randomly selected location with nonzero β coefficient. Even with the tapering model, this includes many locations with trivial nonzero correlations. In practice, we probably do not want to reject these locations. Repeating the simulation experiments with the varying noise levels but only

marking locations with $\gamma_k > 0.1$ as true alternatives increased the mean sensitivities of all experiments by factors ranging from 2.9 to 3.2 (see Table A.8 and Table A.9 in the Appendix A). Using these modified true alternatives, mean FDP values also increased between 24 and 32 percent over the original experiment. However, the mean FDP values all were still below the 0.1 target threshold.

Across the original simulations, additional rejections made by the smoothing method are much more likely to occur in regions with significantly higher correlations, as seen in Figure 2.1b. Additionally, the smoothing method results in large percentage increases in the sensitivity over the base method. In every simulation experiment, there was at least a 20% increase in the sensitivity using the smoothing method. In general, as the correlation length scale decreased the increase in sensitivity substantially increased, reaching a peak increase of 82.3% when $\phi = 0.8$ and $\delta = 7$. These experiments demonstrate that empirically the smoothing procedure does successfully increase the power of the hypothesis testing procedure.

Infrequently—in approximately 1.1% of simulations with $n = 50$ and $\sigma_\varepsilon = 1$ —the smoothing method made no rejections while the base method made a positive number of rejections. The opposite situation, with no base rejections but positive smoothing rejections occurred even more rarely in only 0.17% of situations. However, as the noise increased the two methods achieved more parity in these outcomes, with 4.2% of simulations favoring the base method and 3.2% favoring the smoothing method in this regard at $\sigma_\varepsilon = 2$, and 2.3% for both methods at $\sigma_\varepsilon = 3$. One possible way this can favor the base method is when the only rejections are relatively isolated and come from t -statistics that barely exceed the rejection threshold. In these cases, the smoothing method averages the covariate data for locations with t -statistics that were large enough to be rejected with nearby non-rejected covariate locations, causing the originally rejected t -statistics to lose magnitude and consequently fail to be rejected. This suggests that some caution may be warranted by researchers who might wish to employ this method when they suspect that the signal they are investigating is both weak and has a small spatial-scale.

Since we are using the same data to fit the covariance functions and conduct the hypothesis tests, there is the potential that double-dipping could bias our hypothesis testing results. To assess this risk, we ran additional simulations where the smoothing weights were instead fit on a second independent sample of the covariate data. This second sample of size n was drawn from the same multivariate normal covariate distribution as the original sample with the same mean vector, $\mathbf{0}$, and covariance matrix, Σ . The original covariates that were used to generate y_i were standardized and then smoothed using the new weights estimated from the independent data. A typical result comparing the p -values from these two versions of the smoothing method is presented in Figure A.3 of Appendix A. The differences in p -values between the two methods were quite small, typically with magnitudes on the order of 10^{-4} . This suggests that in our simulations the use of the covariate data twice is not significantly biasing the results of our tests. Additionally, we compared the false discovery and evaluation metrics for the two methods and found negligible differences (Tables A.6 and A.7).

To test the sensitivity of these results to the choice of window size, we ran the baseline simulation study, with $n = 50$ and $\sigma_\varepsilon = 1$, using alternative window sizes of $12^\circ \times 12^\circ$ and $16^\circ \times 16^\circ$. The results are presented in Tables A.3 and A.4. The percentage increases in sensitivity were 12.0% and 18.6% for window sizes $12^\circ \times 12^\circ$ and $16^\circ \times 16^\circ$, respectively, compared to the 25.2% increase found with the $20^\circ \times 20^\circ$ window. The decline in sensitivity improvement as the window size shrinks suggest that using the larger window provides a more complete description of the local covariance.

To examine sensitivity of the results to only using a single location with a nonzero β coefficient we ran the baseline scenario again with two nonzero β values. This was done twice: once with equal signals $\beta_1 = \beta_2 = 1$ and once with different signals $\beta_1 = 1.5$, $\beta_2 = 0.5$. These results are presented in Table A.5 of Appendix A. For the base and smoothing methods, the mean FDP declined in the new scenarios between 14% and 23% compared to simulations with the same parameters but a single nonzero β (rows 1 and 2 of Table 2.1). In each case, the two methods had similar percentage declines, but the smoothing method had a smaller

misspec.	method	FDP	s.d.	FDX	s.e.	true rej.	s.d.	sens.	s.d.	% Δ sens.
Matérn-1/2	base	0.065	0.105	0.215	0.004	114.5	94.6	0.041	0.035	-
Matérn-1/2	smooth	0.059	0.124	0.199	0.004	192.6	140.3	0.069	0.052	67.6
Matérn-5/2	base	0.065	0.115	0.236	0.004	187.5	84.4	0.067	0.032	-
Matérn-5/2	smooth	0.063	0.123	0.217	0.004	229.5	108.0	0.082	0.040	22.2
Laplace	base	0.052	0.127	0.160	0.004	122.8	107.5	0.044	0.040	-
Laplace	smooth	0.051	0.136	0.149	0.004	158.1	139.4	0.057	0.051	28.5

Table 2.3: The results of the simulation studies under misspecification. The FDX gives the empirical probability estimated across the 10,000 simulations. The other evaluation metrics give the mean values of each quantity across the 10,000 simulations. The s.d. lists the sample standard deviation across the 10,000 simulations of the preceding column. The s.e. is the estimated standard error for the FDX probability. The percent change in sensitivity of the smooth method from the base method is listed in the column following the sensitivity.

decline in FDP than the base method. The sensitivity also declined for each method by about 5% in the first scenario and about 14% in the second scenario. Again, the declines in sensitivity in both cases were similar, but the smoothing method had smaller declines (6.0% decline for base vs. 3.7% decline for smooth with the same β and 14.2% vs. 14.0% with different betas). The percent increase in sensitivity for the smoothing method over the base method was similar to the situation with a single nonzero β with 28.3% and 25.6% increases in sensitivity in the two new scenarios respectively.

Next, the effect of two types of misspecification on the method were explored. First, the same simulation study scheme was used, but multivariate normal data were generated using tapered Matérn-1/2 and Matérn-5/2 covariance functions instead. The covariance parameter values were held fixed at $(\phi, \theta, \psi, \delta) = (4, 0, 1, 7)$, while n was held fixed at 50 and σ_ε was fixed at 1. Despite the change in the covariance function used to produce the synthetic data, the Matérn-3/2 covariance function was still employed in fitting the local covariance functions and generating smoothing weights. Second, heavier-tailed Laplace noise with scale parameter $\sigma_\varepsilon = 1$ was used in generating $\mathbf{y}$. The misspecification results are collected in Table 2.3. In each case, the smoothing method maintains conservative, empirical FDR control and reduces the estimated FDX while increasing the number of true rejections and the sensitivity, showing

that the method is robust to these types of model misspecification. In the Matérn-1/2 case, there were many more convergence issues when fitting the MLEs and the fitted covariance functions often underestimated the correlation length scale. Despite these issues, in the rougher Matérn-1/2 case smoothing seems especially beneficial, even with the misspecified covariance function.

Finally, we explored the stability of the rejections made with the smoothing method across different samplings of the noise. For this experiment, a sample size of $n = 50$ and an isotropic covariance function with $\phi = 4$ and $\delta = 7$ were used. y was sampled 1,000 times from the model with $\beta_{l^*} = 1$ and $\sigma_\varepsilon^2 = 1$, but, in contrast to the earlier simulations, the location l^* used to generate Y_i was held fixed, rather than randomly resampled. Results from the smoothing method and base method varied across different simulations of the covariates, but a typical case is shown in Figure 2.1b. Compared to the base method, the smoothing method shows a small decline in the rejection frequency for the points that are closest to l^* . However, in return, the smoothing method makes much more frequent rejections for points that are slightly further away.

These simulation studies provide a few key takeaways for applying this method in practice. First, using the primary window size of $20^\circ \times 20^\circ$ resulted in better increases in sensitivity than using smaller window sizes, but these simulations were specifically designed with a covariance structure that mimics the structure of mid-latitude SSTs. For other climate variables, the appropriate window size for covariance fitting and smoothing should be chosen based on domain knowledge about that variable. Second, in the simulations, the smoothing procedure was robust to double-dipping, suggesting that it is not necessary to segment the limited data into separate chunks used for covariance estimation and hypothesis testing. This is crucial because of the short record lengths that are common in climate. Finally, these simulation results showed robustness to mismatches between the covariance function used for local fits and the covariance function that was used to generate the data, so it is not necessary to perfectly match the local covariance structure to see benefits from smoothing.

2.4 Real Covariate Data and Application

2.4.1 SST Covariates with Synthetic Response

To illustrate the smoothing method with real covariate data, the NOAA Extended Reconstructed SST Version 5 dataset of Huang et al. (2017) was used. This dataset contains $2^\circ \times 2^\circ$ latitude-longitude gridded, monthly SST data recorded at over 8,000 locations globally starting in the year 1850. To remove impacts of seasonality, we use only July data, which was then subset to 60°S - 60°N . Locations where the data were constant for more than 25% of the years, which occurs due to seasonal freezing, were removed, and locations in small land-locked seas were also removed as they lacked sufficient contiguous locations for the windowed smoothing method. This filtering process resulted in a total of 170 observations recorded at each of 8,146 locations. From these observations, a 50 year subset from 1959 to 2008 was chosen. This time window overlaps with the availability of high quality re-analysis temperature data from ERA5 (Hersbach et al., 2020, 2023) that is analyzed in Section 2.4.2. The length of the time window is set to 50 years to make the results more comparable to the length of time for which satellite records are available for many climate variables. At every location, location-specific linear trends and means were removed, and the data were divided by location-based standard deviations to generate standardized anomalies. Figure 2.2 shows the fitted MLE values of the model parameters on this data.

In order to illustrate differences between the smoothed and base methods, a gridbox location was fixed and multiple response vectors were generated by adding standard normal noise to the standardized anomalies at that location. Figure 2.3 shows the results of inference for each method with the first 8 samplings of noise after a seed was set. Based on the patterns displayed in Figure 2.3, in the majority of cases the smoothing method increases the size of the rejection regions. In all but the last simulation, the core rejection region located around the selected location is rejected by both the base and smoothing method. In the last simulation, the randomly sampled error weakened the signal to the point that the

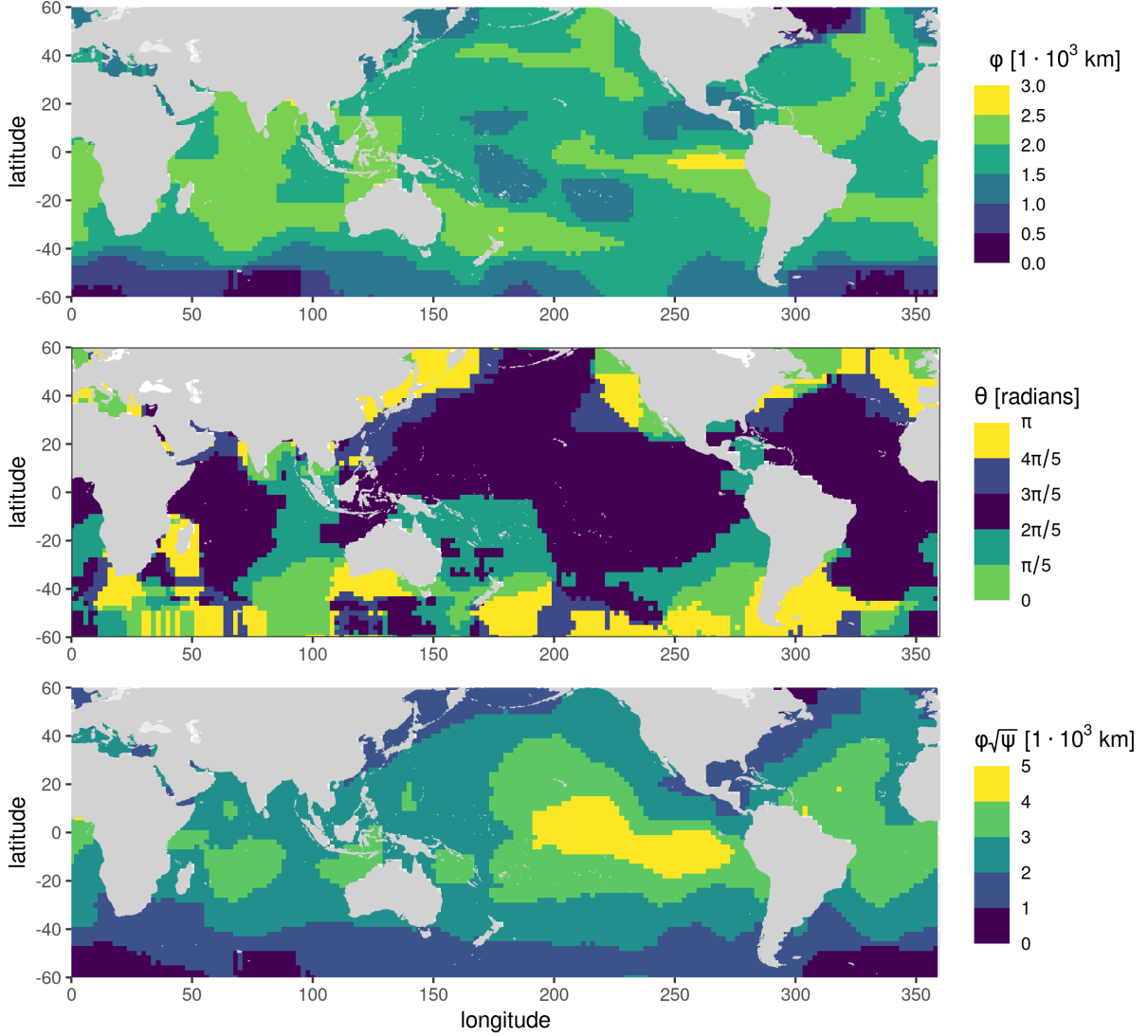


Figure 2.2: The MLE parameters fit on SSTs from the years 1959 to 2008. The parameters were fit using the local model defined in (2.6) with Matérn-3/2 covariance as defined in (2.9). $\phi\sqrt{\psi}$ is plotted rather than ψ , since $\phi\sqrt{\psi}$ is the value of the correlation length scale for the more highly correlated direction.

smoothing method makes no rejections while the base method still rejects this core region. Each method also tends to make additional rejections outside of the core region in spatially coherent patterns, while the base method also tends to make a number of smaller spatially isolated clusters of rejections. These simulations highlight the instability of the rejections

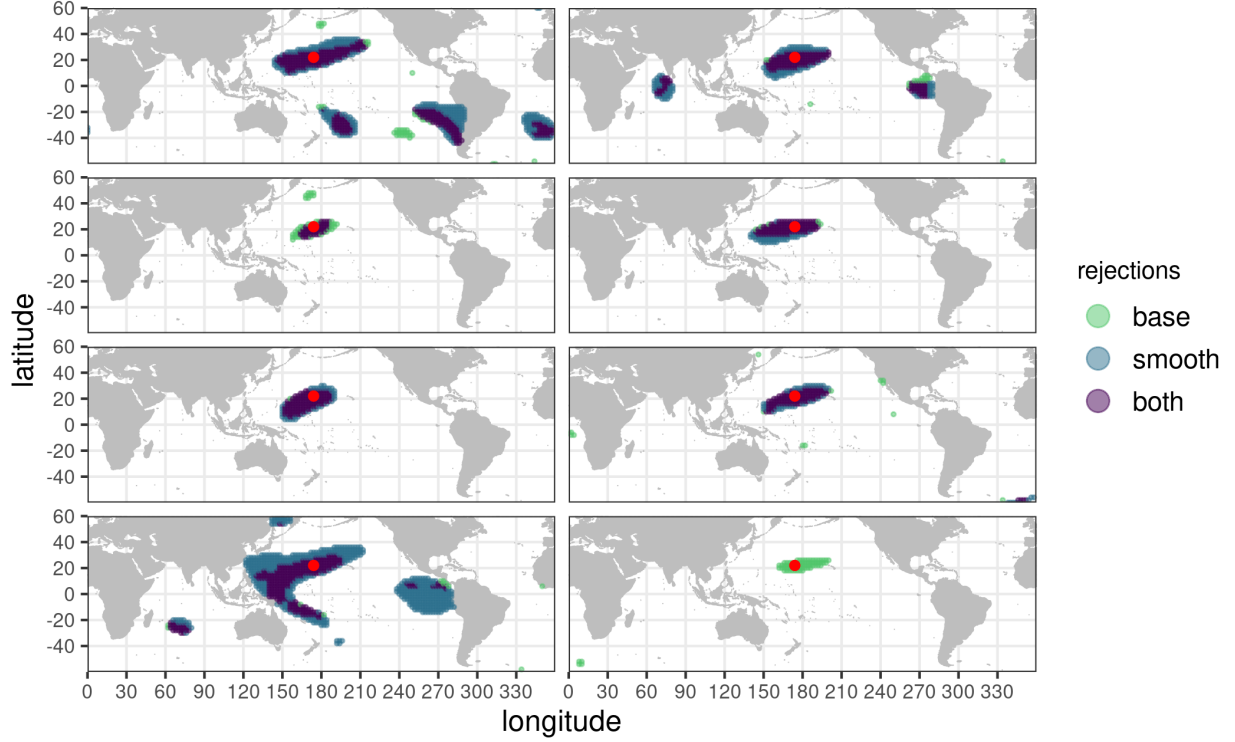


Figure 2.3: Inference performed with 8 independent samplings of the noise. The seed was randomly selected before the 8 generations were made. Rejections are labeled by points, with color depending on whether only one of the two methods rejected that location or both methods did. The location used to generate Y_i is denoted with a red point.

when fitting noisy responses against correlated covariates with only 50 years of data.

2.4.2 SST Covariates and United States Temperatures

Finally, we apply these techniques to a real data case study. Ocean temperatures are an important driver of climate over land through atmospheric teleconnections, wherein convection over the ocean leads to propagation of waves through the atmosphere that affect remote locations. Motivated by prior work that has explored the link between SSTs and Central US heat and drought (e.g., Namias, 1991; McKinnon et al., 2016; Jia et al., 2016), we explored whether there was a significant linear relationships between mean July temperatures in

the Central US (32.5°N-45°N; 105°-125°W) and SSTs. The temperature anomalies were constructed from ERA5 data (Hersbach et al., 2023) which was subsetting to the fifty year period between 1959 and 2008 and area-weighted averaged over the region before a linear trend was removed from the time series. This time series was then regressed against the processed SST covariates from Section 2.4.1. In Figure 2.4, we compare three cases: rejections based on whether individual p-values are less than 0.05 (without correcting for multiple hypothesis testing), rejections based on standard Benjamini-Hochberg with targeted FDR control level $\alpha = 0.1$, and rejections based on Benjamini-Hochberg after applying our smoothing method. Before correcting for multiple hypothesis testing, many locations are rejected. However, given that we are testing 8,146 hypotheses and the data are strongly spatially correlated, this is expected and not directly interpretable. After applying Benjamini-Hochberg without smoothing, no rejections are made, which could suggest that there is no meaningful signal between SSTs and Central US temperatures, in contrast to prior results. However, when the analysis is repeated with our smoothing method applied to the covariates, a cluster of locations in the Pacific Ocean are found to have significant relationships with the temperature time series.

The rejected locations in this cluster all have a negative linear relationship between July SST anomalies and Central US temperatures. This relationship suggests there may be an association between cooler than usual SSTs in this region and hotter July temperatures in the Central US. Cool SSTs in this region have been identified in prior work as a precursor to Central US heatwaves (McKinnon et al., 2016) and low precipitation (Ting and Wang, 1997), although in each case the SST pattern was more widespread than what was identified here. While the regression problem explored here is not identical to the prior work, it raises the possibility that there are more limited regions of SST influence on central US summertime weather than have been identified previously. By pooling the spatial signal at each covariate location, we have increased the strength of the relationship between the temperatures and locations, thus increasing the power of the hypothesis tests. This enables

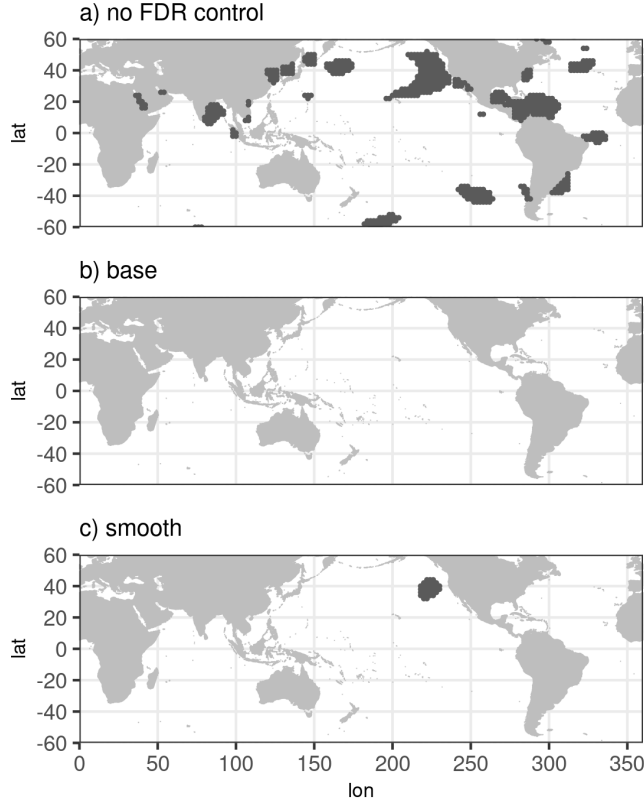


Figure 2.4: Rejections made by each method in the regression analysis of July US temperature anomalies, averaged over the selected region, against July SST anomalies. a) shows rejections made using a 0.05 significance level with no FDR control. b) shows that no rejections are made after applying Benjamini-Hochberg with $\alpha = 0.1$ to the same p -values from the first panel. c) shows the results of combining our covariate smoothing method with Benjamini-Hochberg.

us to detect interesting regions such as this Pacific Ocean region while still protecting against what are likely many spurious rejections when tests were performed without the multiple hypothesis control.

2.5 Summary, Discussion, and Conclusions

In this work we have proposed a local smoothing method, based on fitted local covariance structures, as a pre-processing step for inference. The method is intended for use in climate regression problems, where highly correlated covariates are being tested in parallel at many locations. While we do not present theoretical guarantees, our simulation results demonstrate

that the smoothing method maintains empirical FDR control at the specified level. In addition, the results show increases in sensitivity toward rejecting false nulls by over 20% and high specificity on the true nulls. Similar to the results in Wilks (2016) for the t -tests of means, we find that the average FDP across simulations was conservative compared to the targeted FDR control level, $\alpha = 0.1$.

Despite these advantages, the smoothing method did slightly increase the proportion of simulations for which the procedure made no rejections, suggesting there is a small trade-off between increasing average power at the cost of more frequently making no discoveries. This trade-off became smaller when the noise of the response variable was increased, and was indistinguishable in the high noise setting. If researchers expect that the signal they are investigating is both weak and small-scale, using the original data without smoothing may be more effective in ensuring some discoveries are made.

These simulations also highlight a number of potential issues that climate scientists should keep in mind when performing many hypothesis tests. Repeated simulations where different samplings of the noise were added to the same gridbox while keeping the covariate data fixed produced large variations in the total number of rejections and in their spatial pattern. This phenomena is particularly striking in examples where noise is added to real SST covariates, such as those shown in Figure 2.3. This shows that the outcomes of the testing procedure can acutely depend on the particular observed sampling of the noise. In some simulated covariate cases, despite the clear relationship between the chosen gridbox and the response variable, no gridboxes were selected after performing Benjamini-Hochberg. An examination of these cases revealed that the particular sampling of the noise masked the underlying correlation. Sometimes, when rejections were found, the rejections appeared as a tight core around the true gridbox. Other times, the rejections covered huge swathes of the testing domain. In all of these cases, nothing about the underlying predictor variables or the coefficient β had changed. The changes in the results were simply shifting due to the noise. This should prompt caution when drawing conclusions from the results of hypothesis testing on many

climate phenomena where we recognize that the contribution of internal variability is large and the spatial correlation of covariates is significant, but we only have one observational record to study.

One potential concern with our method is that the same covariate data is used to fit the smoothing model as is used in the hypothesis tests, raising the issue of double-dipping. It is possible that using the data in this way could bias the hypothesis tests and break FDR control, although empirically this did not occur in our simulation studies. One avenue to resolve these concerns would be to fit the local covariance models on climate model output instead. Climate models are based on physical equations rather than observational data, so are independent from our covariate data, but should broadly share the same covariance structures. Further work is needed to determine if the use of an inexact covariance structure in this approach would decrease the power of the smoothing method.

A concern with the FDR control framework is that the false discoveries are only controlled on average, and with any individual sample of data the false discovery proportion is not necessarily controlled. In our simulation studies, we present estimated $\text{FDR}(0.1)$ values that are mostly smaller than 25%, but this suggests that in some cases the FDP might be larger than our target control level about a quarter of the time. As an alternative to FDR control, another line of research (see e.g. Goeman and Solari (2011) and Goeman et al. (2021)) develops methods for finding upper confidence bounds on the FDP of rejection sets. If this confidence bound is large, it signals that there is a high level of uncertainty in the multiple testing problem and the rejections should be treated with skepticism. In the climate setting with highly correlated tests this could be a fruitful avenue for future exploration.

Despite its limitations, the combination of smoothing methods with Benjamini-Hochberg holds promise for improving the reliability and power of large scale hypothesis testing in the climate setting. Applying some correction for multiple hypothesis testing is crucial for performing research that involves many comparisons. As a framework, FDR control finds a useful balance between making discoveries and protecting against false ones. Whether our

spatial smoothing method, another spatial FDR method, or the original Benjamini-Hochberg method is employed, our hope is that correction for multiple hypothesis testing becomes more widespread in climate applications moving forward.

2.6 Data and Code Availability

The NOAA ERSST, Version 5 data (Huang et al., 2017) is publicly available and can be obtained from the NOAA PSL, Boulder, Colorado, USA, at their website: <https://psl.noaa.gov/data/gridded/data.noaa.ersst.v5.html>. The ERA5 data (Hersbach et al., 2020) is also publicly available and can be obtained from the Copernicus Climate Change Service, Climate Data Store, at <https://doi.org/10.24381/cds.adbb2d47>. The code for this project is available at a public github repository https://github.com/kylemcevoy/cov_smooth_climate.

Chapter 3

Quantifying Uncertainty in Extreme Precipitation Quantiles over the Conterminous United States using an Observational Large Ensemble

The observational record of precipitation provides an inherently limited sample of low-likelihood, high-impact extremes, making quantification of uncertainty in these extremes difficult. While climate models are a useful tool, the well-known biases in the precipitation simulated by Earth System Models necessitate alternative approaches. In order to address these challenges, we use statistical resampling techniques to create an Observational Large Ensemble of monthly precipitation, enabling a more robust sampling of the precipitation distribution. Using this large ensemble, we explore how important modes of climate variability combine with unpredictable sources of variability to produce extreme precipitation in the continental United States. Confidence intervals for extreme quantiles of December precipitation show substantial spread in the western and southeastern United States, highlighting the potential that estimates based on our current records might underestimate extremes. To

illustrate the influence of the climate modes on extremes a local case study near Los Angeles is presented. Finally, the method is used to explore variability in precipitation across California, showing the potential that dry years might be drier and dry stretches more frequent than expected based on the observational record due to sampling of internal variability alone.

3.1 Introduction

Effectively managing water resources requires understanding the complex spatiotemporal variability of precipitation, but the short observational records of precipitation and the high variability of the process make estimates of the tails of the precipitation distribution highly uncertain. This uncertainty makes planning for overly wet or extreme drought years difficult. Furthermore, there are a range of processes and timescales relevant for precipitation extremes, ranging from synoptic-scale storms (e.g. Barlow et al., 2019) to decadal-scale variations in sea surface temperatures (SSTs) that can drive persistent atmospheric circulation responses (e.g. Deser et al., 2004). A complete estimation of the distribution of precipitation, and the expected sampling variability from a limited observational record, requires incorporating multiple processes that span timescales.

Here, we focus on the conterminous United States (CONUS), where studies have previously investigated precipitation extremes by fitting generalized extreme value distributions to daily maxima over blocks of time, usually seasons or years (e.g. Zhang et al., 2010; Risser et al., 2021). But, water availability for cities and agriculture also depends on the surpluses or deficits of precipitation that accumulate over much longer time frames, such as months or seasons. Additionally, while the spatial coverage of these studies can be large, the models are generally fit to time series of data individually, without accounting for the complex spatial structure of precipitation. Since water resources are often managed and distributed over much larger regions, it’s important to model the full complexity of the spatiotemporal process.

In contrast, Earth system models (ESMs) produce full spatiotemporal simulations of

precipitation. These models can be run repeatedly with perturbations to the initial conditions to produce large ensembles that can be used to quantify the internal variability in precipitation (Deser et al., 2020). However, while the latest Coupled Model Intercomparison Project (CMIP) Phase 6 models have reduced biases compared to CMIP Phase 5 models (e.g. Danabasoglu et al., 2020), many regional biases persist in various aspects of the precipitation distribution (e.g. Sharif et al., 2025; Rivera and Arnould, 2020; Ossó et al., 2023), suggesting that alternative methods are needed to properly quantify the uncertainty in monthly extremes.

Stochastic weather generators (Wilks and Wilby, 1999) are another tool for generating simulated spatiotemporal fields of precipitation. Typically, these methods have been applied and validated with daily precipitation fields over smaller domains like watersheds (Chen et al., 2010; Verdin et al., 2015; George and P., 2023). Though recent work has applied these ideas to simulating daily rainfall over larger domains, like the state of California (Najibi et al., 2024), these methods generally focus on much smaller spatial domains than CONUS.

Climate modes of variability have been previously linked to shifts in the precipitation distribution over the continental United States in the literature. The positive phase of the El Niño-Southern Oscillation (ENSO) has been linked to increases in both daily maximum winter precipitation (e.g. Zhang et al., 2010; Risser et al., 2021) and seasonal totals of winter precipitation (Risser et al., 2021; Stone et al., 2023; Deser et al., 2018) over the southern US. The positive phase of the Pacific Decadal Oscillation (PDO) has also been linked to increases in daily maximum winter precipitation and winter seasonal totals in the central and eastern US and declines in the Ohio River valley (Zhang et al., 2010). The positive phase of the Pacific North America (PNA) pattern has been associated with large declines in winter daily maxima over the interior west of the US and the Ohio River Valley (Risser et al., 2021), as well as declines in winter precipitation totals over these regions (e.g. Leathers et al., 1991; Risser et al., 2021). The North Atlantic Oscillation (NAO) was found to have only limited effects on winter daily precipitation maxima by Zhang et al. (2010), but Risser et al. (2021) found that the positive phase of the NAO was linked to significant increases in

winter daily maximum return levels and seasonal totals of precipitation over a region centered on Arkansas. Because of the significant impacts of these modes on precipitation variability the effects of these various modes will need to be accounted for in order to properly model precipitation variability over CONUS.

In this work, we expand on the method of McKinnon and Deser (2021) to study extreme monthly and seasonal precipitation over CONUS using an Observational Large Ensemble (Obs-LE). The focus of the analysis is two-fold. The first goal is to quantify the uncertainty in our estimates of extreme monthly precipitation by producing confidence intervals for the quantiles of the distribution. The second is to improve our understanding of how large scale modes of climate variability, such as ENSO, influence these monthly extremes, and how novel combinations of these modes, rarely seen in the observations, can shift the distribution of precipitation and extreme values over CONUS.

3.2 Methods

3.2.1 Data Sources

For observational precipitation data, we use the Global Precipitation Climatology Centre’s (GPCC) 1 degree by 1 degree latitude longitude resolution gridded dataset (Schneider et al., 2017), and for climate model data we use the Community Earth System Model version 2 large ensemble (CESM2-LE) at nominal 1 degree by 1 degree latitude longitude resolution (Rodgers et al., 2021). For the CESM2-LE, we restrict to the 50 members of the large ensemble with the smoothed biomass burning forcing due to concerns about changes in variability induced by changes in the forcing (Heyblom et al., 2022). To calculate the climate modes (discussed below), we use monthly time series of the modes as calculated by the Climate Variability Diagnostics Package (CVDP) version 6 (Maher et al., 2025) for both the observations and the CESM2-LE. For the observations, the monthly Niño-3.4 and Pacific Decadal Oscillation indices are calculated using ERSSTv5 (Huang et al., 2017) sea surface temperature data

from the period 1920-2020, while the PNA and NAO indices are calculated using sea level pressure data from a combination of the ERA-20C reanalysis (Poli et al., 2016) for the years 1920-1939 and the ERA5 reanalysis (Hersbach et al., 2020) for the years 1940-2020.

3.2.2 Statistical Model and Methods

In order to better quantify uncertainty in monthly precipitation extremes given short observational records, we update the observational large ensemble (Obs-LE) method of McKinnon and Deser (2021). In this method, synthetic records of precipitation are generated using a combination of resampling techniques that capture the impact of climate modes of variability and unrelated atmospheric variability on the spatiotemporal precipitation fields.

In order to apply the method, we fit regression models of transformed precipitation (the transformation is discussed below) against climate modes. The choice of modes in this model was based on those that had previously been found to have significant impacts on monthly precipitation totals or on daily precipitation extremes in the literature, as discussed in the introduction. Since different months and locations can have different climate responses to the modes, we fit separate regression models for each month and location. This leads to a collection of regression models

$$\begin{aligned} \tilde{P}_{l,t}(\lambda) = & \beta_{0,l,m} + \beta_{\text{ENSO},l,m}\text{ENSO}_t + \beta_{\text{PDO},l,m}\text{PDO}_t + \beta_{\text{PNA},l,m}\text{PNA}_t \\ & + \beta_{\text{NAO},l,m}\text{NAO}_t + \varepsilon_{l,t}, \end{aligned} \quad (3.1)$$

the subscript l refers to location, the subscript t indexes the months through time, m refers to the month corresponding to time t , and λ is the optimized value of the transformation parameter. The error term $\varepsilon_{l,t}$ is assumed to have mean zero and constant variance. The fitted coefficients from this model are treated as fixed in the rest of this process.

R^2 values for the monthly models of transformed precipitation are shown in Figure 3.1. Similarly to McKinnon and Deser (2021), we find that most R^2 values are low with the majority of values under 0.15, though the values in December, January, and February tend

to be larger. Thus, consistent with prior work (McKinnon and Deser, 2021), the majority of variability in precipitation at the gridbox scale is caused by largely-unpredictable “atmospheric noise”; as we will see, the influence of the modes is greater at larger spatial scales.

As in McKinnon and Deser (2021), we do not fit a forced response in this model. Attempts were made to fit a forced response using atmospheric $\log(\text{CO}_2)$ concentrations and anthropogenic aerosol emissions as regressors. However, when fitting these regressions for individual CESM2 members we found low signal-to-noise ratios for these variables with widely varying fitted forced responses across ensemble members. This suggests that with currently available data it might not be possible to consistently estimate the forced response in precipitation at the gridbox level. Therefore, we dropped the forced response from the model. One consequence of this decision is that the forced response will end up as part of the residual term that is resampled across the record.

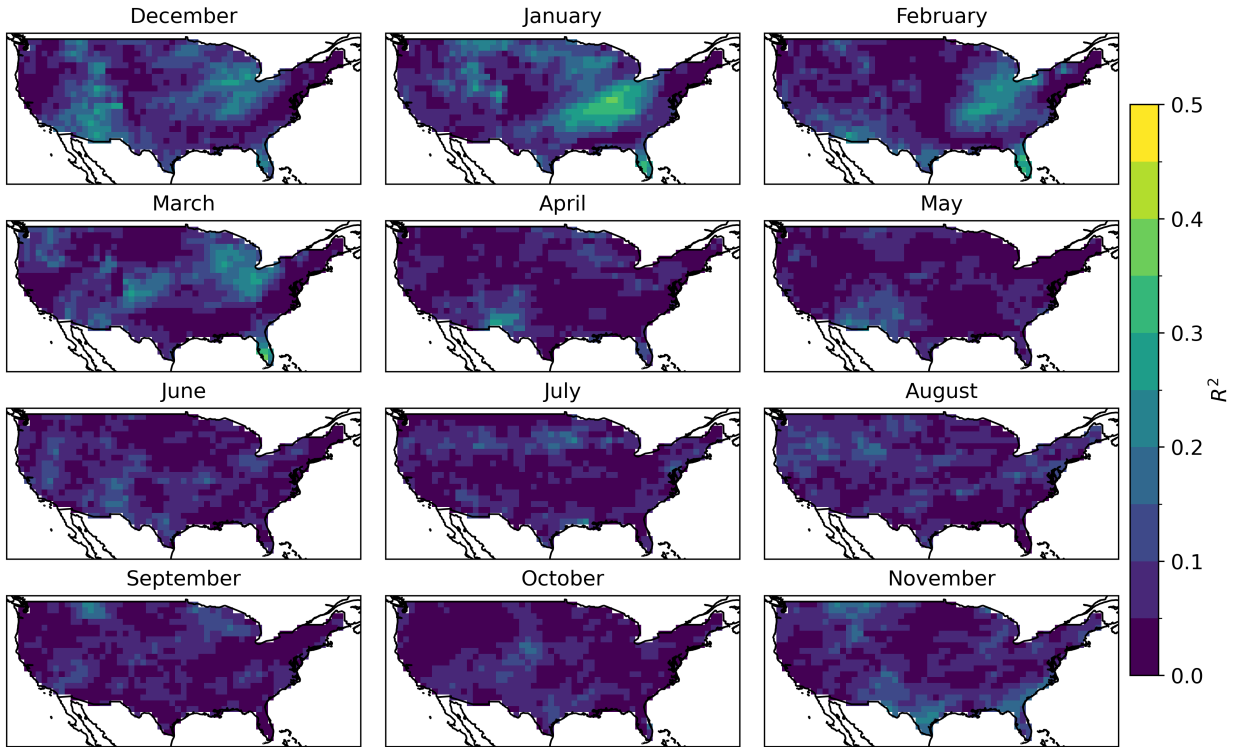


Figure 3.1: Monthly R^2 values for the fits of Eqn. (3.1) on transformed precipitation.

Within our model, we use the Box-Cox transformation to improve the fit by increasing

the symmetry and homoscedasticity of the error terms. The transformation is defined for transform parameter $\lambda > 0$ as,

$$\tilde{P}_{l,t}(\lambda) = \frac{P_{l,t}^\lambda - 1}{\lambda},$$

where $P_{l,t}$ is the monthly precipitation rate in units of mm/day at location index l and time index t . Each month and location is transformed separately, and this transform is optimized separately for each month and location as well. The parameter is selected through maximizing the log likelihood of the residuals across different choices of λ , assuming that the residuals are normally distributed, as in Box and Cox (1964). Since the model assumes homoscedasticity in the errors rather than in the unconditional precipitation, optimizing the transformation using the residuals more directly targets the desired variance stabilization of the residuals. Optimizing λ using the method of Box and Cox (1964) requires that all values of $P_{l,t}$ must be positive to ensure that the Jacobian of the transformation is nonzero. Therefore, we add a value of $1 \cdot 10^{-6}$ to each precipitation value. And, for ease of optimization and the ability to compare across locations, we restricted our optimization of λ to the set $\{1/4, 1/3, 1/2, 2/3, 3/4, 1\}$.

As in McKinnon and Deser (2021), the indices for each of these climate modes are sequentially orthogonalized using the Gram-Schmidt process. This orthogonalization removes any lag-zero correlations between the modes, simplifying the process of generating synthetic modes. The order of the sequential orthogonalization is ENSO, PDO, PNA, and NAO, and further references to these modes within this paper refer to the orthogonalized versions of these mode indices. Therefore, the effect of the PNA, for example, should be interpreted as the effect of the PNA after its correlation to ENSO and PDO has been removed.

In order to simulate new realistic sequences of the climate modes, we employ the Iterative Amplitude Adjusted Fourier Transform (IAAFT) method (Schreiber and Schmitz, 1996, 2000). This method was previously used in McKinnon and Deser (2021) to generate synthetic versions of the ENSO, PDO, and AMV. In that work, the surrogate modes were generated independently from each other; however, while the modes have been sequentially orthogonalized,

this only constrains the lag-zero cross-correlations to be 0. As the lag increases the dampening effect of the orthogonalization on this cross-correlation weakens and the cross-correlations between the ENSO, PDO, and PNA increase in magnitude (see Figure B.1 for an example with the ENSO and PDO modes; the NAO does not significantly correlate with the other modes). Therefore, in order to generate synthetic modes with realistic cross-correlations, we implemented a multivariate version of the IAAFT method as described in Keylock (2012). This multivariate version of the algorithm modifies the synthetic modes so that the phase shifts between the modes are forced to equal those that were observed in the original modes, maintaining the observed cross-correlations (Figure B.2). In order to preserve the monthly pattern of variances in the ENSO, PDO, PNA, the newly generated synthetic modes are rescaled by month, so that the monthly variances match those of the observed index. This process results in synthetic modes that closely match the observed power spectral density of the modes and still retain realistic cross-correlations.

As in McKinnon and Deser (2021), the residuals of the fitted Eqn. (3.1) are resampled using a moving block bootstrap with blocks that include the entire spatial field and multiple, integer years of time. This choice of block helps maintain a realistic spatiotemporal correlation structure in the bootstrapped residuals. To select a sufficiently long block length to preserve the temporal autocorrelation structure of the data, the method of Wilks (1997) was applied to the residual time series of each gridbox. As in McKinnon and Deser (2021), we determined the time block length by taking the 0.97 quantile value of the fitted block lengths across the study domain. This ensures that the blocks are of sufficient length to capture the autocorrelation at most locations, while reducing the influence of outliers. This process resulted in a block length of 24 months for both the GPCC and CESM2 datasets.

Finally, the simulated spatiotemporal fields are made by plugging the synthetic modes

and the bootstrapped residuals into Eqn. (3.1) with the fitted coefficients:

$$\begin{aligned} \tilde{P}_{l,t}^{\text{Synth}}(\lambda) = & \hat{\beta}_{0,l,m} + \hat{\beta}_{\text{ENSO},l,m} \text{ENSO}_t^{\text{Synth}} + \hat{\beta}_{\text{PDO},l,m} \text{PDO}_t^{\text{Synth}} + \hat{\beta}_{\text{PNA},l,m} \text{PNA}_t^{\text{Synth}} \\ & + \hat{\beta}_{\text{NAO},l,m} \text{NAO}_t^{\text{Synth}} + \varepsilon_{l,t}^{\text{Synth}}, \end{aligned} \quad (3.2)$$

where the superscript Synth marks the newly generated components and $\hat{\beta}$ denotes the fitted coefficients.

Conceptually, the fitted values generated by multiplying the fitted coefficients against the synthetic modes model the impact of the climate modes for synthetic ensemble members, and the resampled residual, $\varepsilon_{l,t}^{\text{Synth}}$, can be thought of as containing a random sample of short term atmospheric variability that is uncorrelated with the modes and the features of precipitation that are not captured by the model, including any non-linear effects of the modes, a forced response, or the impact of other modes that were not included in the model. Therefore, this method allows us to explore how the juxtaposition of new sequences of climate modes with different samplings of unrelated internal variability could lead to different frequencies and magnitudes of extreme precipitation. Repeating this process many times enables us to estimate the variability in observed extremes in a way that accounts for alternative sequences of climate modes, including unobserved or rarely-observed combinations of these modes.

Repeating this process many times, we can generate the Obs-LE. For this work, we generated a 1,000 member Obs-LE using the GPCC data. Similarly, we generated 1,000 member ensembles from each of the 50 CESM2-LE members to form 50 Synthetic Large Ensembles (Synth-LEs).

We treat the new synthetic records as independent samples from the climate process, and use these samples to estimate a confidence interval for the τ -quantile, $\tau \in (0, 1)$, of monthly precipitation at each location j , $Q_j(\tau)$. For each synthetic record i and location j , we calculate the τ sample quantile of monthly precipitation, $\hat{Q}_{j,i}(\tau)$, over the desired set of months. The sets of months over which the quantiles are calculated are flexible. For example, we can calculate the quantiles over all months, over all Decembers, or over all months in the

December-January-February (DJF) season. This process gives us 1,000 estimates of the τ quantile at each gridbox, $\{\hat{Q}_{j,i}(\tau)\}_{i=1}^{1000}$. From this set of sample quantiles we estimate the lower and upper bounds of the 95% confidence interval for $Q_j(\tau)$ using the 0.025 and 0.975 quantiles of the estimates, $\{\hat{Q}_{j,i}(\tau)\}_{i=1}^{1000}$, respectively.

3.3 Validation

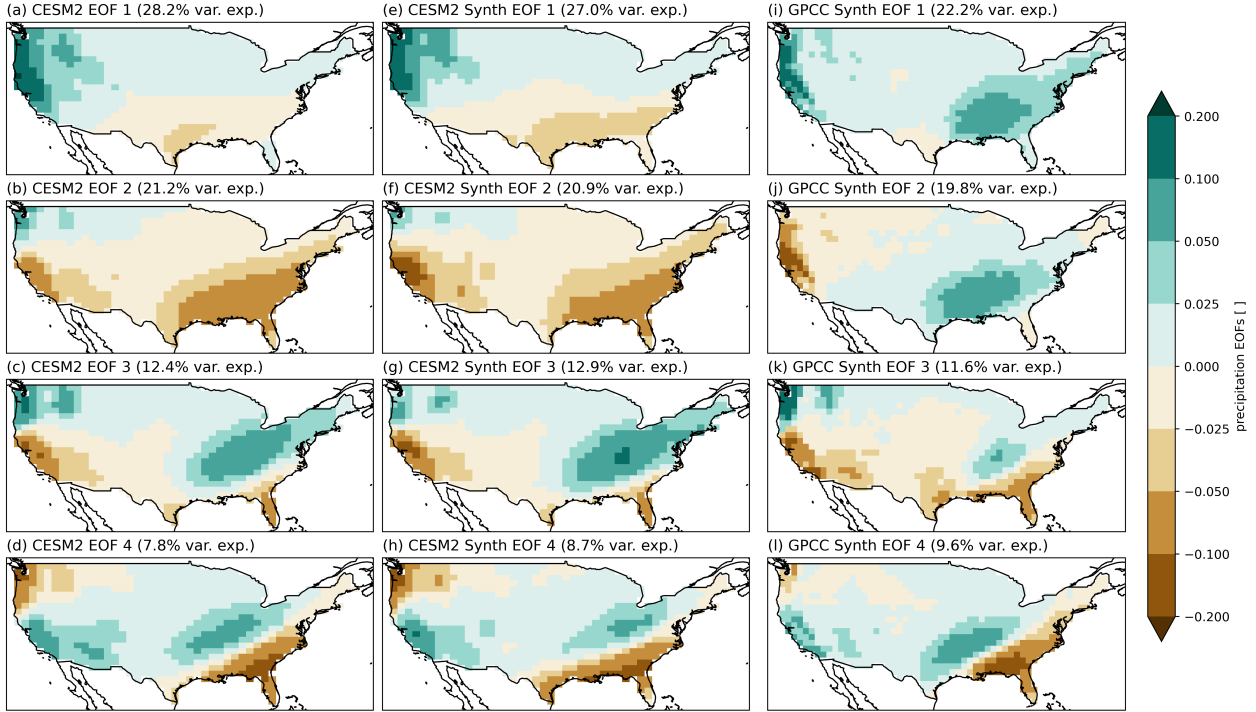


Figure 3.2: EOFs calculated on 100 years of DJF season precipitation totals across 50 ensemble members. For each ensemble the ensemble members are concatenated in the time dimension before the EOF decomposition is performed. For CESM2 the EOFs are found using the 50 members of the large ensemble. For the synthetic ensembles 50 synthetic members were used. The synthetic CESM2 members were generated using the first ensemble member of the large ensemble. The EOFs are standardized to have unit ℓ^2 -norm. Since the signs of the EOFs are arbitrary, some sign flipping has been performed to ease visual comparison. Note that the colorbar is nonlinear with values doubling at each colorbar tick.

To evaluate how well the Obs-LE generated from a single record samples the internal variability in observed extremes, we use the CESM2-LE as a methodological testbed (Figure 4 in Deser et al., 2020). First, to validate the spatiotemporal variability sampled by the method,

we perform an empirical orthogonal function (EOF) analysis on the full 50-member CESM2-LE and compare the results to an EOF analysis performed on the Synth-LE generated from just the first CESM2-LE member. For the Synth-LE the first 50 ensemble members are selected to make the comparison. Data matrices for each ensemble are formed by concatenating the values for the 50 ensemble members along the time dimension. In Figure 3.2 we show the first four modes of the EOF analyses on DJF precipitation totals for the 100 complete DJF seasons over the time span 1920 to 2020. These 100 seasons result in data matrices with $50 \times 100 = 5,000$ rows and 839 spatial columns. The EOF analysis shows that the Synth-LE based on ensemble member 1 broadly matches the spatial patterns and variance contained in each mode from the CESM2-LE. There are some differences in the magnitudes of the patterns, with a larger region of negative patterns in the Southeastern United States in EOF 1 and more intense negative pattern in California in EOF 2. In contrast, the EOFs of the GPCC Obs-LE show much different patterns with a more uniform signal across the US in the first EOF and substantially smaller variance explained by this mode. EOF 2 of the GPCC Obs-LE, with its contrasting pattern of western drying and southeastern wetting, does not have a clear analogue in the first four modes of the CESM2-LE EOFs, though this pattern could potentially be created by a linear combination of other EOFs.

Second, to evaluate how well the impact of the climate modes is reflected in the Obs-LE, we compared DJF composites of the climate modes in GPCC with the Obs-LE. In order to form the composites, we selected all DJF months with mode indices above a value of 1 and took the average of the monthly precipitation values to get a positive phase average. Then we selected all DJF months with mode indices below -1 and took the average to get a negative phase average. The composite was formed by subtracting the negative phase average from the positive phase average. The results are plotted for the GPCC and the Obs-LE in Figure 3.3. The Obs-LE patterns match the GPCC patterns well, though the composite values tend to be somewhat smaller in magnitude for the Obs-LE, especially for the PDO.

Third, we can use the collection of Synth-LEs generated from each CESM2-LE ensemble

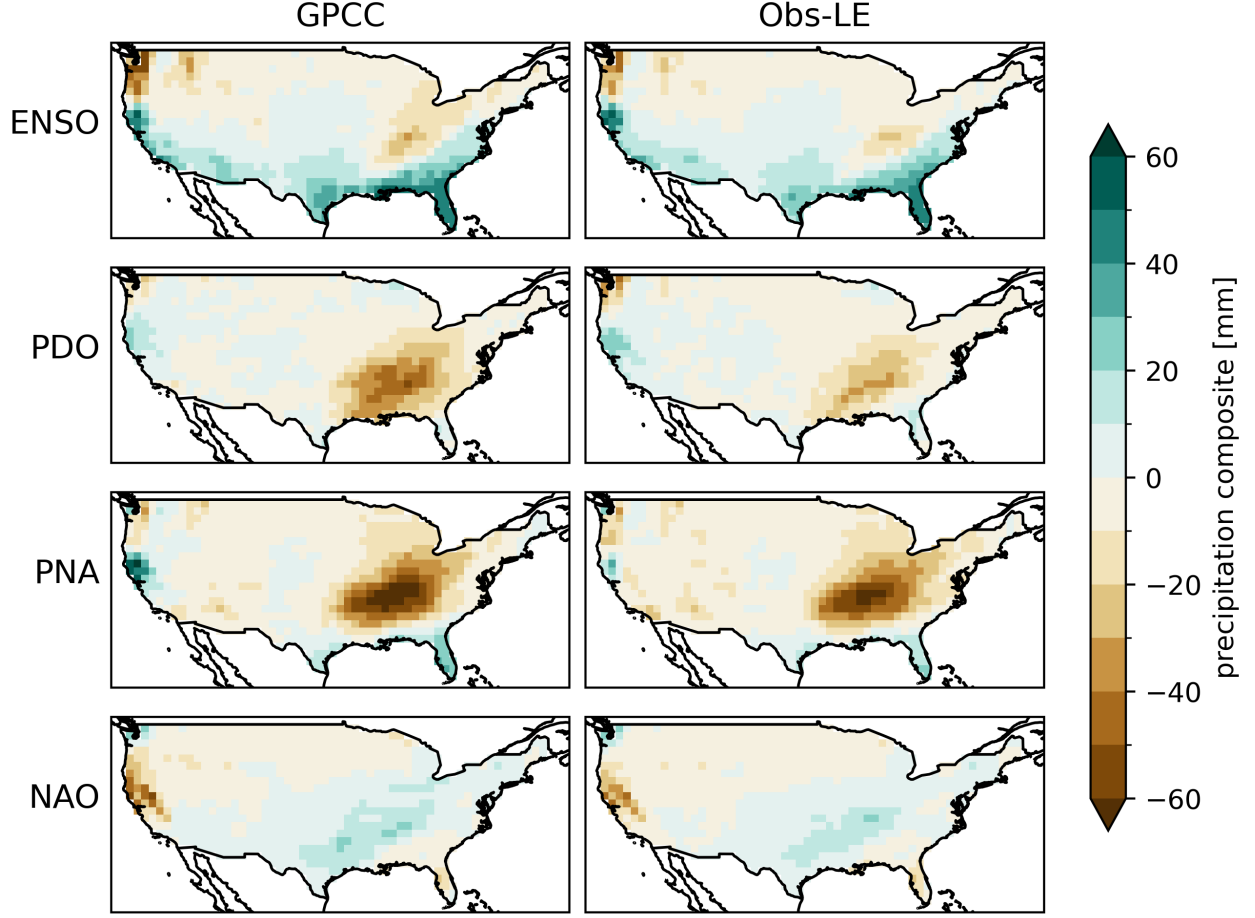


Figure 3.3: Precipitation composites for climate modes of variability in GPCC observations and the GPCC Obs-LE during DJF. For each mode index, averages of monthly precipitation are taken over all months in the DJF season with positive and negative phase, respectively, and the negative phase average is subtracted from the positive phase average to obtain the composite. Positive phase months are defined to be those with index values that were greater than 1 standard deviation above 0, and negative phase months are defined to be those with index values less than 1 standard deviation below 0.

member to understand how well the confidence interval generated from a single record can capture the variability of the full unobserved distribution. For the Synth-LE generated from CESM2-LE member $k \in \{1, 2, \dots, 50\}$, let $\hat{Q}_{k,j,i}(0.95)$ be the sample quantile estimated at location j for Synth-LE member i across the December months, and let $\bar{Q}_{k,j}(0.95) = (1/n) \sum_{i=1}^n \hat{Q}_{k,j,i}(0.95)$ be the associated sample mean of the quantile estimates, where

$n = 1000$. Then,

$$s_{k,j} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n \left(\hat{Q}_{k,j,i}(0.95) - \bar{Q}_{k,j}(0.95) \right)^2},$$

gives the sample standard deviation of the quantiles for Synth-LE k . Let $s_{CESM2,j}$ be the corresponding sample standard deviation taken across the 50 quantile estimates of the CESM2-LE. Then, $s_{CESM,j}$ is an estimate of the variability of this quantile across the climate distribution, and each $s_{k,j}$ corresponds to an estimate of this variability based on a single record.

Figure 3.4a) shows the fractional difference between the mean $s_{k,j}$ value,

$$\bar{s}_j = \frac{1}{50} \sum_{k=1}^{50} s_{k,j},$$

and $s_{CESM2,j}$, normalized by $s_{CESM2,j}$, i.e.

$$\frac{\bar{s}_j - s_{CESM2,j}}{s_{CESM2,j}},$$

for December precipitation. The majority of locations have fractional biases between -0.15 and 0.15 , but a few locations have values below -0.30 and above 0.30 . The average fractional bias value is 0.026 . The variability of these standard deviation estimates is larger on the west coast of the United States and in the eastern part of the country (Figure 3.4b)). Notably, there is a cluster of especially large standard deviations centered in the Sierra Nevada mountains of California. Histograms of $s_{k,j}$ at individual locations show a range of patterns, including large spread and low bias (Figure 3.4c, a gridbox near Los Angeles) and lower spread but high bias (Figure 3.4d, in the eastern US). At most locations there is substantial sampling variability in our estimate of $s_{k,j}$ across the Synthetic-LEs, which should be kept in mind when evaluating the GPCC confidence intervals that are based on our only observational record. Figure B.3 shows analogous results for the 0.05 quantile.

One concern with using the CESM2-LE to test the validity of the quantile variability

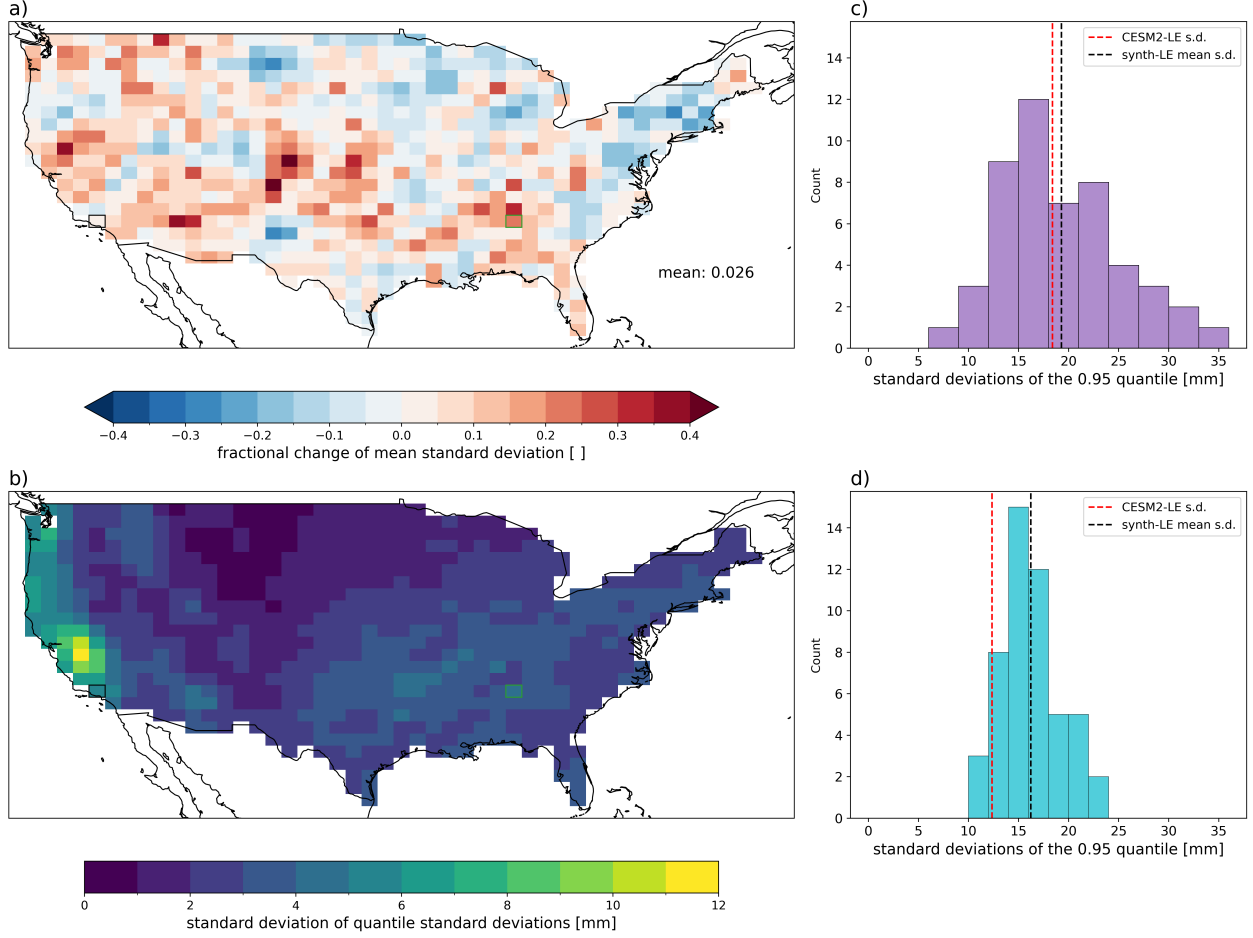


Figure 3.4: a) The fractional change in the mean standard deviation of the 0.95 quantile of December precipitation across the Synth-LEs generated from each CESM2-LE member from the standard deviation of 0.95 quantile estimates of December precipitation across the CESM2-LE member, i.e. $(\bar{s}_{\text{Synth}} - s_{\text{CESM2}})/s_{\text{CESM2}}$. b) The standard deviation of the 50 estimates of the standard deviation of the 0.95 quantile from the 50 Synth-LEs. c) Histogram of the standard deviation estimates of the 0.95 quantile across the 50 Synth-LEs at the gridbox surrounded by the black rectangle in a) and b). d) Histogram of the standard deviation estimates of the 0.95 quantile across the 50 Synth-LEs at the gridbox surrounded by the green rectangle in a) and b). In c) and d) CESM2-LE standard deviation is plotted as a red dashed line, while the mean value across the 50 Synth-LE estimates is plotted as a black dashed line. The bin widths of the two histograms are different.

sampled by the Obs-LE is that the optimization process for λ can produce different values for the GPCC and the CESM2-LE. Figure B.4 shows that the fractional differences of the individual $s_{k,j}$ values depends somewhat on the value of λ with a trend of increasing errors in the right tail as λ decreases. Despite this trend, there is heavy overlap in the interquartile

region of the distributions. Compared to the CESM2-LE, the GPCC optimized values of λ show a lower proportion of 1/4 and increased proportions of 1/3 and 1/2 (Figure B.5).

3.4 Results

Now that we have validated the method, we turn to understanding the range of extreme high and low precipitation that could occur in an 100-year record due to different samples of internal variability. We are particularly interested in cases where there is a large uncertainty in the extremes due to high variability, and in cases where the plausible extremes are meaningfully different from what was observed. To do so, we examine the spread of extreme quantile estimates across the Obs-LE. We first focus on December and use the 1,000 member GPCC Obs-LE to estimate uncertainties in the 0.05 and 0.95 quantiles of precipitation across all December months and produce a 95% confidence interval (CI) for each quantile. These quantiles correspond to left and right tails values that would be exceeded once in every 20 Decembers in the long-run. The confidence intervals for the 0.95 quantile in Figure 3.5 show substantial spread with a handful of gridboxes exceeding 150 mm in spread and many more gridboxes exceeding 100 mm in spread, indicating that there is high uncertainty in our estimates of these quantiles from the current observational records. These regions with the most uncertainty broadly track the regions with the largest 0.95 quantiles, concentrating in the Pacific Northwest and interior South. As shown in Figure 3.5 these regions have especially large upper tail uncertainty with upper bounds of the CI that are at least 50 mm higher than the observed quantiles. One notable region of uncertainty relative to the observed sample quantile is in Southern California concentrated near Los Angeles, where the upper bound is 50 to 100 mm higher than the observed quantile values.

The observed 0.05 quantile value is under 10 mm across much of the interior and southwestern US, leaving little room for uncertainty in the negative direction (Figure 3.6). However, the Pacific Northwest and eastern US show more spread in the confidence interval,

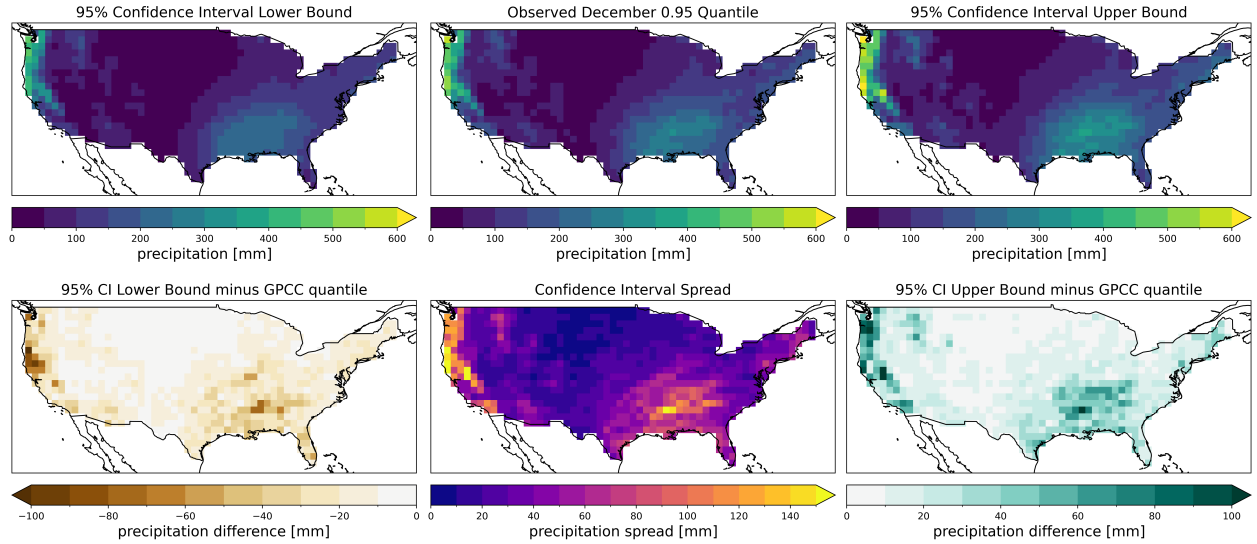


Figure 3.5: Confidence intervals for the 0.95 quantile of December monthly precipitation found using the GPCC Obs-LE. The spread refers to the upper bound of the confidence interval minus the lower bound.

including gaps of over 30 mm between the observed quantile values and the lower bound of the confidence interval at some locations.

Turning our attention to the impacts of the climate modes, some combinations of the climate modes have only rarely, or never, appeared in the records to date, making it difficult to disentangle the impact of the modes from the observed samplings of internal variability. This lack of data obscures how the effects of these modes might combine with each other, potentially leading to more extreme values. To explore how rare combinations of modes could impact December, January, February (DJF) precipitation over CONUS we used the fitted regression coefficients to select sets of modes that maximized precipitation over California and Florida, respectively. For California, this wet-phase maximization occurred when ENSO was in the positive phase and PNA and NAO were in the negative phase. The corresponding dry-phase minimization occurred with these modes in the opposite signs. For the PDO, the signs of the coefficients over California were not consistent across these months, so we did not condition on the PDO. For Florida, the precipitation was maximized with ENSO, PDO, and PNA in the positive phase and NAO in the negative phase.

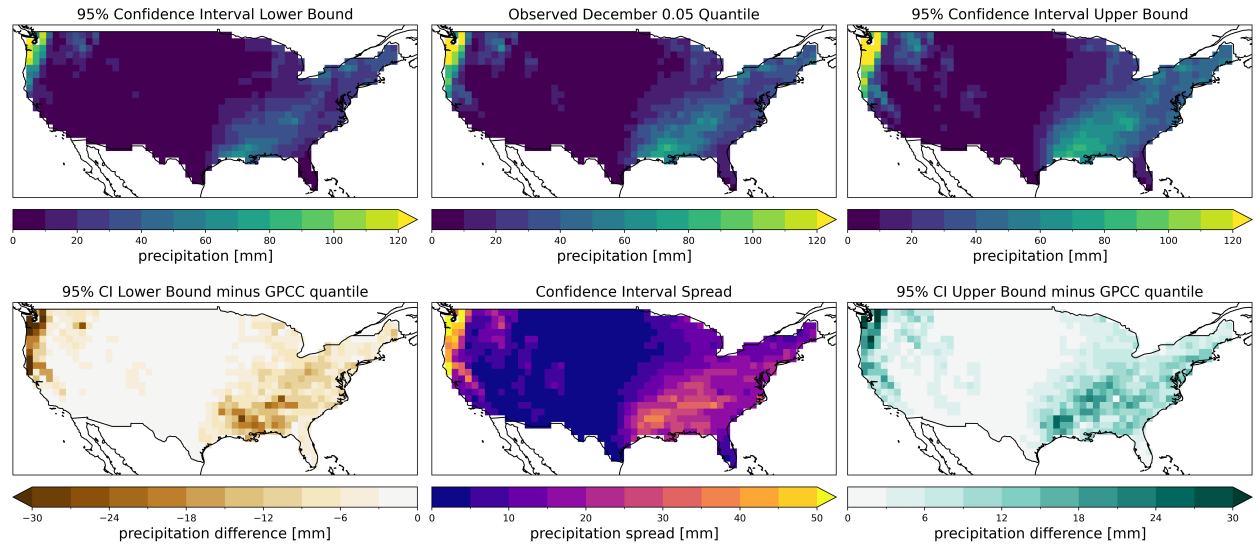


Figure 3.6: Confidence intervals for the 0.05 quantile of December monthly precipitation found using the GPCC Obs-LE. The spread refers to the upper bound of the confidence interval minus the lower bound.

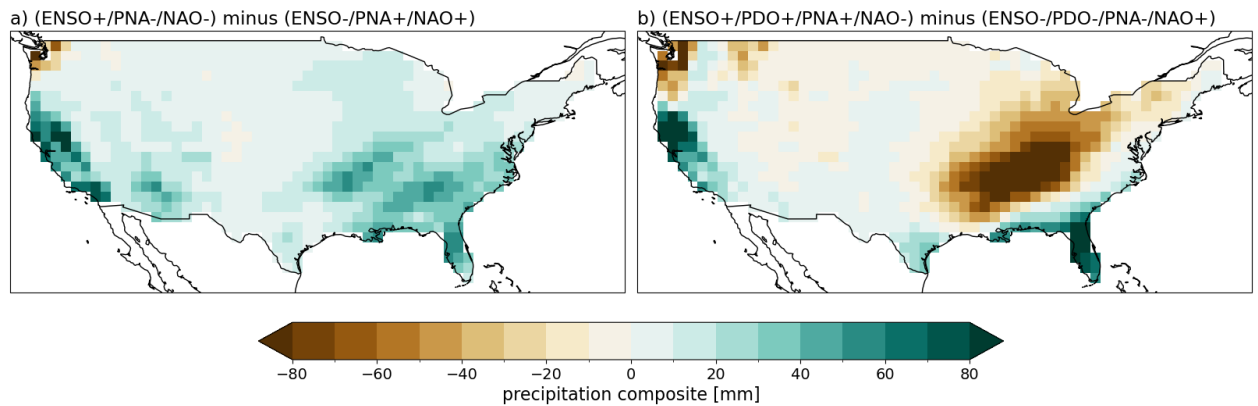


Figure 3.7: Monthly precipitation composites conditioned on multiple modes for December, January, and February (DJF) months from the GPCC-based Observation Large Ensemble. The composites were calculated by averaging over DJF months for which the listed surrogate modes were above (+) or below (-) one standard deviation from 0, and subtracting the average over the subset with modes above or below 1 standard deviation with the opposite signs. The modes in a) were selected to maximize precipitation in California, while the modes in b) were chosen to maximize Florida precipitation.

To quantify the impact of these combinations, we took the total of 303,000 December, January, and February months in the Obs-LE (1000 ensemble members times 101 years times 3 months per year) and found the subsets of the months where the standard deviation of each index was simultaneously above a magnitude of 1 in the direction of either the wet or dry phases. The precipitation in these months was then averaged over the wet and dry phases and the dry phase average was subtracted from the wet phase average. The results in Figure 3.7 show much stronger effects than the individual mode composites of Figure 3.3, illustrating the increased potential for extreme precipitation with favorable combinations of modes. While the spatial pattern for the California-based composite shows above average precipitation across nearly all of CONUS, with the exception of western Washington, the Florida-based composite shows that the wettest conditions in Florida arise during the driest conditions in the southern Mid-west and interior South and vice versa.

While the composites emphasize the average effects of the modes, internal variability is still a dominant component of the precipitation fields. To illustrate the influence of internal variability with a case study, Figure 3.8a compares histograms of monthly precipitation during the DJF season in the Obs-LE at the gridbox centered at $34.5^{\circ}N$, $118.5^{\circ}W$ latitude-longitude (this gridbox lies just north of the city of Los Angeles). The histograms are plotted under dry, neutral, and wet conditions defined by negative ENSO and positive PNA, Neutral ENSO and PNA, and positive ENSO and negative PNA, respectively. The neutral conditions are all DJF months for which the mode index is within 1 standard deviation of 0 and the positive (resp. negative) condition are defined as months for which the index is 1 standard deviation above (below) 0. Under dry-mode conditions there is a notable increase in the concentration of monthly totals under 50 mm and a shortening of the right tail values compared to the neutral conditions, while under the wet-mode conditions the distribution shifts to the right and becomes more skewed. The impact on the extremes of the distribution can be seen in the shifting of the 0.95 quantile value from a value of 144 mm in dry conditions, to 201 mm in neutral conditions, and 282 mm in the wet conditions. Despite this strong impact of ENSO

and PNA on the shape of the distribution, the role of internal variability is key as shown by the spread of the histograms under all three conditions. Even under the dry conditions, 32.7% of months have precipitation amounts above the overall median and even under wet conditions we see 8.8% of months under the the overall 0.2 quantile.

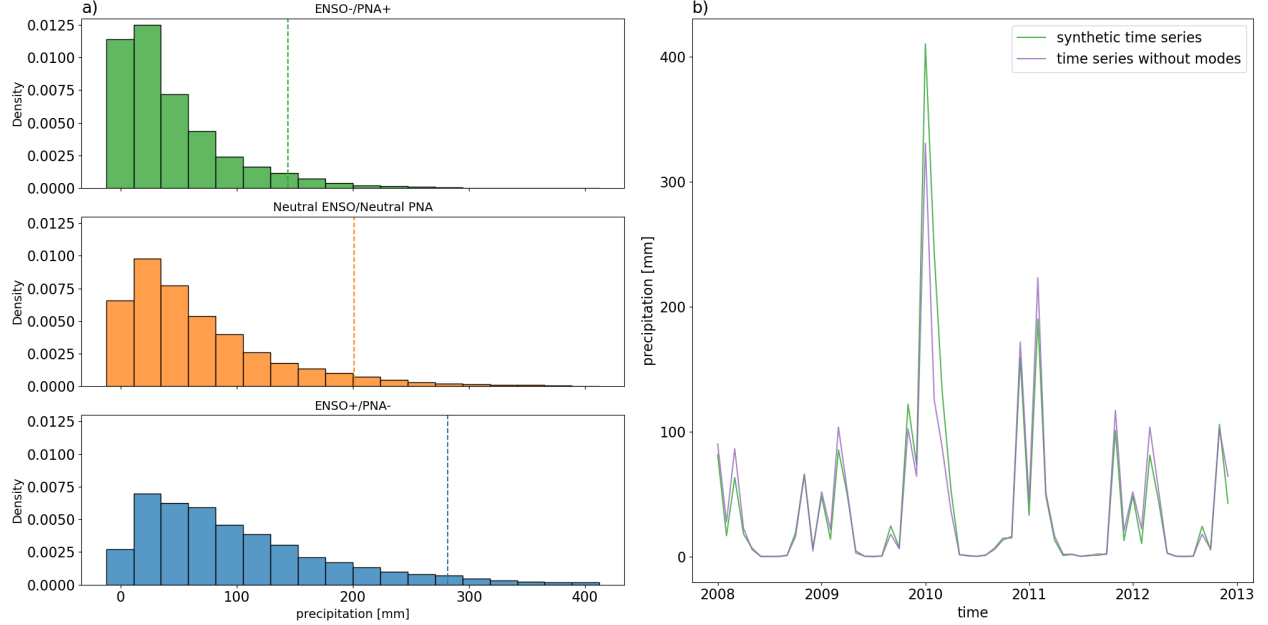


Figure 3.8: a) Histograms of DJF season monthly precipitation values from the GPCC Obs-LE at the $34.5^{\circ}N \times -118.5^{\circ}W$ latitude-longitude gridbox. The histograms are conditioned on values of the ENSO and PNA indices, with ENSO/PNA+ and ENSO/PNA- referring to the positive and negative phases of those modes, respectively. The mode is in the positive phase when the standardized index is greater than 1, and the mode is in the negative phase when the index is less than -1 . The neutral phase of each mode refers to when the index is between -1 and 1 standard deviations from 0. Dashed lines indicate the estimated conditional 0.95 quantile of the DJF monthly precipitation distribution. b) An example synthetic time series highlighting an extreme event simulated by the Obs-LE at the same gridbox. An alternate realization of the model is plotted where the same sequence of residuals was input to the model, but with all modes set equal to 0.

The largest extreme values are most likely to arise during the juxtaposition of modes and atmospheric variability that both favor precipitation. In the context of our model, this effect can be demonstrated by comparing very extreme events simulated in the Obs-LE with an alternate realization of those same events simulated with the same residual sequence but with all of the modes in the model set to 0. Figure 3.8b shows an example of such a

comparison at the same gridbox near Los Angeles. At this location, the selected extreme month had a simulated precipitation value of 410.0 mm which is above the simulated 0.998 quantile of monthly precipitation during DJF at this location in the Obs-LE, and is larger than the maximum observed value of 342.3 mm in the GPCC dataset from January 1969. The event simulated with the same residual but with all of the modes set to 0, had a monthly precipitation value of just 330.5 mm. In this simulated month, the surrogate ENSO index had a value of 2.73 while the other indices were neutral. The combination of a strong ENSO with a large residual component led to a record breaking value.

3.4.1 California Water Resources

California experiences a high degree of interannual variability in precipitation with important consequences for municipal water supplies, ecosystem health, and agriculture. Here we use the Obs-LE to better understand uncertainty in key aspects of California water year precipitation. In order to define California precipitation, we take the GPCC gridboxes whose centers lie within the state borders and take the sums of the monthly totals from October to September of the following year. Then, we take area-weighted means of the gridbox time series to get time series of mean California precipitation. Applying this process to each Obs-LE member produces an ensemble of water year time series, and from this ensemble we can estimate quantile distributions and other features of California precipitation. These analyses are validated in Figures B.6, B.7, and B.8.

First, we analyze the 0.2 quantile of California precipitation. This quantile has previously been used as a threshold for “very dry years” (Persad et al., 2020). The observed 0.2 quantile over the years 1920 to 2020 was 394.3 mm, but the Obs-LE distribution for the quantile, shown in Figure 3.9a, shows substantial spread around this observed value. The 95% CI for this quantile calculated from the Obs-LE is 369.6 mm to 425.4 mm, showing the potential for much drier values of this quantile. We can repeat this process conditional on the value of the ENSO index. Since most precipitation in California falls during the winter, DJF, season, we

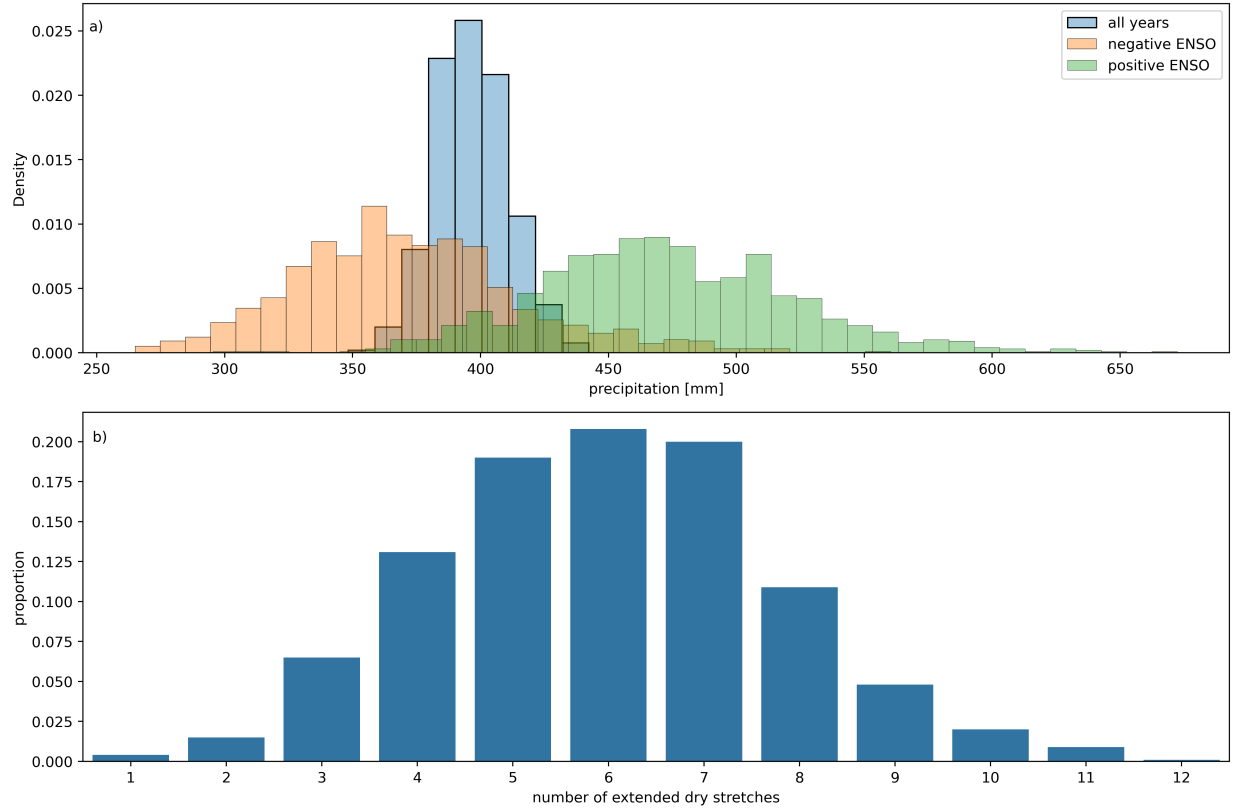


Figure 3.9: a) Histograms of 0.2 quantile estimates across the Obs-LE for California water year (September to October) mean precipitation. Negative and positive ENSO are defined by averaging the ENSO index over the DJF season, and selecting corresponding water years to include based on whether the index was above 1 or below -1 . b) The proportion of Obs-LE members with the specified number of extended dry stretches over the 100 year synthetic record. Extended dry stretches are defined as sequences of at least three consecutive years of below median California average precipitation. Sequences longer than three years are only counted once.

use the mean DJF index for each water year to find positive ENSO (mean index values greater than 1) and negative ENSO (mean index values less than 1) water years. Finding the 0.2 quantiles on these subsets of water years, gives us distributions of the quantile for each ENSO phase. As expected from the substantially smaller samples of years that fall into each phase, the variability of the conditional quantiles is much higher than the unconditional quantile estimates (Figure 3.9a). While the observed conditional 0.2 quantiles do not differ much from the unconditional quantile (401.0 mm for negative ENSO and 397.9 mm for positive ENSO), these estimates are based on only 18 and 19 water years of data for negative and positive

ENSO, respectively. In the Obs-LE, the conditional 95% CIs are 384.2 mm to 580.4 mm for positive ENSO and 292.0 mm to 475.5 mm for negative ENSO. These large uncertainties suggest that 100 years is not a sufficient sample to estimate the conditional quantiles with precision, and this uncertainty should be accounted for in water resources planning.

In addition to dry year precipitation values, the sequencing of dry years can dramatically impact the water resources picture for the state. In the observed record, there are 6 distinct sequences of at least three below average precipitation years in a row between 1920 and 2020. Many of these are associated with notable droughts including the 2011-2016 drought and the 1986-1992 drought. However, the number of such sequences over a 100 year span is not fixed but will itself vary. By counting the number of such sequences across different Obs-LE members, we can understand the variability in the number of extended dry sequences. The distribution of the number of extended dry stretches across the Obs-LE members, shown in Figure 3.9b, peaks at the observed value, and about 60% of the members had counts of 5, 6, or 7 dry stretches. About 95% of members had counts greater than equal to 3 or less than equal to 9. These results suggest that even without factoring in any changes to the climate it's plausible that we could see significantly more multi-year dry spells in California than we observed over the 100 years between 1920 and 2020.

3.5 Discussion

In this work, we have expanded upon the Obs-LE precipitation method of McKinnon and Deser (2021) and employed it to estimate uncertainty in extreme precipitation sample quantiles and explore the impacts of modes on the behavior of extreme monthly precipitation. We also used the Obs-LE to explore uncertainty in California precipitation with important implications for water resources planning.

The key takeaway of this work is that uncertainties in the quantiles of the observed precipitation are large even over 100 year time spans. This is especially true of quantiles

that lie deeper in the tails of the distribution, but, as demonstrated in the California results, even at larger spatial scales and less extreme quantiles there is significant uncertainty in our estimates. This uncertainty is exacerbated when we consider the quantiles conditional on climate modes.

One limitation of the current method is that we have assumed a relatively simple model for how precipitation depends on climate modes of variability. In reality, even in transformed space the precipitation might not depend linearly on the climate modes. We might also expect interactions between the modes and residual variability. Another possibility is that the tails of the distribution have a different dependence structure on the modes than the center of the distribution. As in McKinnon and Deser (2021), we did not include a forced response in this model because we found that it could not be consistently estimated across CESM2-LE members. This is not to say that there has been no forced response in precipitation over this time period, but that it cannot be consistently estimated over this time frame. While a more complex model may better capture the true precipitation response to the climate modes, it is unlikely that we could estimate such a complicated dependence structure well given our limited observational record.

While this work has considered uncertainty relative to new samplings of variability and the modes, we have not addressed other sources of uncertainty in the data. The coefficients and the precipitation transform are fixed before the resampling procedure is performed, so we do not propagate uncertainty from estimating these quantities through to our final confidence intervals. We also not consider uncertainty in our choice of observational dataset. Despite these limits in scope, we have still found substantial uncertainty in the tail quantiles of precipitation.

Another limitation of our work is that the method has been applied to relatively coarse data in the spatial domain. While understanding precipitation extremes at this coarseness is important for managing water supplies, for infrastructure planning and disaster resilience it would be preferable to have data at a higher spatial resolution. Nothing precludes this method

from being applied to higher resolution data spatially, however it’s possible that noisier high resolution data could cause more instability in the coefficient estimates for the climate modes, leading to more unstable synthetic ensembles. This instability could be investigated and possibly stabilized through a spatial model or a smoothing approach.

While we have only focused on uncertainty over the observed record, other research has highlighted potential changes in the precipitation distribution through the end of the century, including more interannual variability globally (Pendergrass et al., 2017) and increases in both “very wet” and “very dry years” in California (Persad et al., 2020). These results, in conjunction with our findings of high uncertainty over the current historical record, suggest that stakeholders need to plan for a broad range of water resources scenarios in the future.

In future work, we could incorporate estimated changes in interannual or monthly variability from climate models into our synthetic records. This would allow us to explore how climate change might impact future variability. Additionally, we could use an Obs-LE approach to generating synthetic precipitation records at the daily time scale. Applying the method to daily data is more challenging as it would require generating a realistic pattern of wet and dry days as well as realistic samples from very skewed distributions. However, daily data would enable interesting analyses of risks over shorter time frames.

A deeper understanding of the uncertainties in precipitation extremes will enable stakeholders to better prepare for scenarios involving the tails of the distribution. Of particular note were the high upper tail uncertainties that we found for estimates of the 0.95 quantile in the Pacific Northwest and southeastern United States. These regions should account for this uncertainty in their planning for the future. Additionally, the potential for California to have faced a much drier climate with more extended dry stretches over the last century should be considered when discussing future water scenarios in the state. In each case, understanding the limitations of our current estimates is key to avoid overconfidence when planning for extreme events over the coming years.

Chapter 4

Quantifying Interannual Variability in Colorado River Flows at Lees Ferry

The Colorado River provides an important supply of water for seven US states, but declining streamflows in the 21st century have led to challenging cuts in water supplies and difficult negotiations over how to divide the remaining water. Understanding the variability in streamflows is key to making more informed water management decisions regarding the river, but this is made more complex by the high interannual variability of the river flows. In the literature, naturalized streamflows at the Lees Ferry Gaging Station have been modeled as a linear function of temperature and precipitation in the Upper Colorado River Basin. In this work, we use Observational Large Ensembles of temperature and precipitation in the Upper Colorado River Basin to generate synthetic temperature and precipitation records. The spatial influence of the climate modes in the Observational Large Ensemble is analyzed. We find that the El Niño-Southern Oscillation produces substantial changes in the monthly temperature and precipitation distributions across the year, though other modes also make substantial differences to monthly distributions. These records are combined with a linear model to produce a synthetic ensemble of streamflow records. This synthetic ensemble is used to estimate variability in streamflows over time and produce quantile estimates for each

year. We find that mean streamflows declined 11% between 1920 and 2018. These records are then combined with estimates of future warming to model the impacts under climate change. We find a 29.9% decline in mean streamflow by 2050 relative to 1921, along with 23.9% declines in the estimated 0.975 quantile of streamflows and 50.1% declines in the estimated 0.025 quantile for streamflows. These declines suggest significant future challenges for the management of the Colorado River.

4.1 Introduction

The Colorado River provides a critical supply of water for seven US states, numerous Native American tribes, and Mexico, but reduced water availability due to the ongoing 21st century drought has brought severe challenges for managing this resource (United States Bureau of Reclamation, 2026b). River water contributes to key municipal supplies of water for cities including Los Angeles, Phoenix, and Denver (Cohen, Michael J., 2011), and is key to ecosystem health, including fish populations (Richter et al., 2020). Additionally, it is estimated that 79% of basin water is used for agricultural purposes (Richter et al., 2020), supporting significant economic activity in the seven US states that share the Colorado River water. Currently, two key reservoirs on the river, Lake Powell and Lake Mead, are again near the historic lows that were recently set in 2022 and 2023 (United States Bureau of Reclamation, 2026b), and the Bureau of Reclamation has invoked the 2019 Colorado River Drought Contingency Plan for the third time since 2020 (United States Bureau of Reclamation, 2026a). Without significant actions the Glen Canyon Dam at Lake Powell and the Hoover Dam at Lake Mead could reach the minimum power pool level, when the dams could no longer produce hydropower, or the dead pool level, when water could no longer pass out of the dams (United States Bureau of Reclamation, 2026b). In this context, as plans are being negotiated to decide on river management for the future, it is crucial to understand interannual variability in Colorado River streamflows over the historical record and into the

future.

The Upper Colorado River Basin (UCRB) typically contributes about 92% of total natural flows, the estimates of what river flows would have been with water diversions removed, across the river basin (United States Bureau of Reclamation, 2026b), making it especially important to understand the variability of Colorado River flows coming from the upper basin. These flows are measured at the Lees Ferry Gaging Station which lies just below the upper basin and plays an important role in allocating water between upper and lower basin states under the Colorado River Compact of 1922 (United States Bureau of Reclamation, 2026b).

A number of works in the literature have investigated natural flows at Lees Ferry, studying sensitivity of the flows to temperature and precipitation in the UCRB (e.g. Vano et al., 2012; Udall and Overpeck, 2017; Lehner et al., 2019; Hoerling et al., 2019, 2024). Over decadal time scales UCRB precipitation has higher correlations than temperature with Lees Ferry flows in the observational record (Hoerling et al., 2024), but many studies have shown that both variables play significant roles (Vano et al., 2012; Lehner et al., 2019; Hoerling et al., 2019). Physically, some portion of precipitation directly feeds into runoff, which drains into the river over the basin, so we'd expect a positive relationship between increased precipitation and flows. In contrast, increased temperatures are associated with higher levels of evapotranspiration, reducing the amount of runoff that makes it into the river (Vano et al., 2012). These sensitivities have been estimated through hydrologic models applied to climate model output (Vano et al., 2012) and regression models of streamflows against temperature and precipitation (Lehner et al., 2019; Hoerling et al., 2024).

Many of these studies have assessed forced responses of temperature and precipitation in climate models over the next few decades or until the end of the century and used these estimates of the forced response to assess changes in streamflows based on sensitivity estimates (Vano et al., 2012; Lehner et al., 2019; Hoerling et al., 2019). Models consistently show that the forced response in temperature increases across the basin (Lehner et al., 2019; Hoerling et al., 2024), but precipitation changes are less certain with significant differences across

different models (Hoerling et al., 2019). However, the ensemble means of the Coupled Model Intercomparison Project phase 6 (CMIP6) multi-model ensemble and the Community Earth System Model version 2 (CESM2) initial condition large ensemble (Rodgers et al., 2021) do show precipitation increases across the first half of the 21st century (Hoerling et al., 2024). Due to the negative relationship between streamflows and temperatures, these forced changes are expected to result in decreased Colorado River flows over the next few decades, though it is possible that forced precipitation increases could balance out the forced temperature changes (Lehner et al., 2019; Hoerling et al., 2024).

In addition to understanding the forced response, understanding internal variability of Colorado River streamflows is critical to assessing potential impacts in the near future due to the large contribution of internal climate variability to precipitation over the UCRB (McKinnon and Deser, 2021) and precipitation’s key role in streamflows. Works that have considered internal variability have often done so through initial condition large ensembles of climate models (e.g. Lehner et al., 2019; Hoerling et al., 2019, 2024). However, while climate models have improved over time, there are still significant biases in precipitation in the latest generation of climate models over the UCRB (Pierce et al., 2022). This suggests the importance of considering observationally based estimates of internal variability in precipitation and temperature, such as those derived using the Observational Large Ensemble (Obs-LE) methods of McKinnon et al. (2017), McKinnon and Deser (2018), and McKinnon and Deser (2021).

The Obs-LE method uses statistical resampling techniques to generate spatiotemporal records of climate variables, and directly incorporates variability associated with important climate modes of variability, such as the El Niño-Southern Oscillation (ENSO) into the synthetic records. Properly accounting for the influence of the modes is crucial to generating realistic variability as prior work has indicated key roles for climate modes in influencing temperature and precipitation in the southwestern US, including drying associated with negative phase of the Pacific Decadal Oscillation (PDO) (Lehner et al., 2018; Seager et al.,

2023) and cooling associated with ENSO (Deser et al., 2018). The evaluation of the influence of these modes in our model distinguishes our work from stochastic streamflow generation approaches such as Gold et al. (2024), which incorporates dry and wet states for rivers in the basin in a Hidden Markov Model but does not link those conditions to the climate modes.

In this work, we evaluate the variability and trends in Colorado River streamflows at Lees Ferry by combining regression models of Lees Ferry streamflows as a function of temperature and precipitation with spatiotemporal fields of temperature and precipitation over the UCRB generated using the Obs-LE methodology. This produces ensembles of synthetic streamflows that can be used to estimate variability in streamflows over time. Using projections of forced changes in temperature from climate models, we project how the mean and variability of the streamflows at Lees Ferry might change due to a warming climate. This methodology produces a novel estimate of streamflow variability based on the observational record with important implications for water policy on the Colorado River.

4.2 Methods

4.2.1 Data Sources

For observational precipitation we use the Pierce et al. (2021) dataset. This dataset contains daily precipitation observations at $1/16^\circ$ by $1/16^\circ$ latitude longitude resolution covering the conterminous United States (CONUS) over the years 1915 to 2018. The Livneh et al. (2013) dataset was used for daily temperature maxima. This dataset contains daily temperature maxima recorded at $1/16^\circ$ by $1/16^\circ$ latitude longitude resolution over CONUS. The dataset originally covered the years 1915 to 2013, but was extended to 2018 by Lu Su. These datasets were downloaded from David W. Pierce’s website (https://cirrus.ucsd.edu/~pierce/nonsplit_precip/ accessed on 11/10/2025). From these datasets, monthly mean temperature maxima and monthly precipitation totals were calculated for all gridboxes falling within the Upper Colorado River Basin. These gridbox time series were processed into

basin-wide time series by taking area-weighted means of precipitation totals, P_i , and mean daily maximum temperatures, T_i . Then, water year values were calculated by summing precipitation and averaging temperature over the water year, which spans the months from October to the following year’s September.

In order to estimate the forced response in monthly mean daily temperature maxima, we used the Community Earth System Model version 2 large ensemble (CESM2-LE) climate model at nominal 1 degree by 1 degree latitude longitude resolution (Rodgers et al., 2021). For the large ensemble, we use the 50 members of the large ensemble with smoothed biomass burning forcing to avoid any increases in variability of the estimated forced response over the years 1997 to 2014 (Heyblom et al., 2022).

To calculate the climate modes over the time period 1920 to 2018 we use monthly time series of the modes as calculated by the Climate Variability Diagnostics Package (CVDP) version 6 (Maher et al., 2025). For observational data, the monthly Niño-3.4 and PDO indices are calculated using ERSSTv5 (Huang et al., 2017) sea surface temperature data from the period 1920-2018, while the Pacific North America pattern (PNA) PNA and North Atlantic Oscillation (NAO) indices are calculated using sea level pressure data from a combination of the ERA-20C reanalysis (Poli et al., 2016) for the years 1920-1939 and the ERA5 reanalysis (Hersbach et al., 2020) for the years 1940-2018.

Natural flow estimates of the Colorado River at Lees Ferry were obtained from Table 8 of the 76th Annual Report by the Upper Colorado River Commission (<http://www.ucrccommission.com/wp-content/uploads/2025/07/UCRC-WY2024-Annual-Report.pdf>, accessed on 03/25/2026).

4.2.2 Flow Model

First, we model the conditional mean of Colorado River naturalized streamflows at Lees Ferry as a function of UCRB precipitation and temperature. The natural flows have a more direct relationship with the temperature and precipitation, but in reality the actual stream flows

are smaller due to upstream water diversions.

As in Lehner et al. (2019) and Hoerling et al. (2024) we use linear regression to model water year totals of naturalized flows of the Colorado River at Lees Ferry as a function of basin temperature and precipitation over the water year. Since the high resolution Livneh dataset only contains daily maximum temperatures rather than daily mean temperatures, we use the monthly mean of daily maximum temperatures in our models instead. This in contrast to Lehner et al. (2019) and Hoerling et al. (2024) where average temperatures were used, but our fitted model produces similar results (4.3.1). Let Q_i be the natural flow at Lees Ferry for water year i , $T_{i,\text{Basin}}^*$ be the mean water year temperature anomalies, and $P_{i,\text{Basin}}^*$ be basin water year precipitation total anomalies. Then, we fit the following regression model

$$Q_i = \alpha + \beta_P P_{i,\text{Basin}}^* + \beta_T T_{i,\text{Basin}}^* + \varepsilon_i, \quad (4.1)$$

for some $\alpha, \beta_P, \beta_T \in \mathbb{R}$, where i indexes the water year, and ε_i is assumed to be i.i.d. with mean 0 and constant variance for the purposes of fitting the model. Following standard practice in the literature (e.g. Lehner et al., 2019; Hoerling et al., 2024), the anomalies for temperature are calculated by subtracting the mean, $T_{i,\text{Basin}}^* = T_{i,\text{Basin}} - \bar{T}_{\text{Basin}}$, while the anomalies for precipitation are further normalized by dividing by the mean value, $P_{i,\text{Basin}}^* = (P_{i,\text{Basin}} - \bar{P}_{\text{Basin}})/\bar{P}_{\text{Basin}}$. As in Lehner et al. (2019), we tested the inclusion of an interaction term between temperature and precipitation in the model but found it to have a negligible impact, so the interaction term was dropped from the model.

4.2.3 Observational Large Ensembles

In order to generate synthetic flow records, we first apply the Observational Large Ensemble methodology to monthly precipitation totals and mean daily maximum temperature in the UCRB. Using the high resolution Livneh and Pierce datasets as the basis for the synthetic ensembles enables the synthetic precipitation to capture the significant impacts of orography

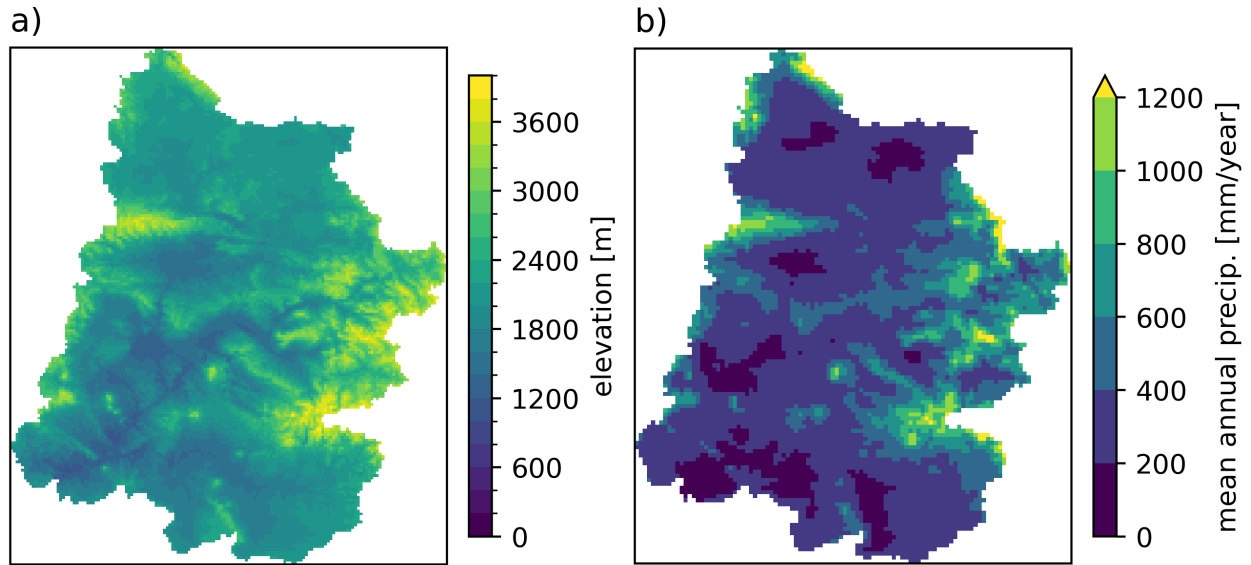


Figure 4.1: a) UCRB elevation from the PRISM 4 km dataset (Daly et al., 1997, , accessed on 05/08/2026 from <https://prism.oregonstate.edu/data/supporting.php>) and b) Mean annual precipitation in the UCRB from the Livneh dataset.

on precipitation in the basin (Figure 4.1).

For precipitation, we follow the methodology outlined in Chapter 3. However, the monthly means of daily temperature maxima are significantly less skewed than those of precipitation, making a Box-Cox transformation unnecessary for the temperature records. Also, in contrast to precipitation, where a forced response is difficult to estimate from a single record (McKinnon and Deser, 2021), the forced response in mean monthly daily max temperatures has a much higher signal-to-noise ratio (Hoerling et al., 2019). Therefore, we estimate a forced response for daily max temperatures by fitting linear regressions for the time series of monthly mean daily max temperature at each gridbox, $T_{l,t}$ where t indexes time and l indexes location,

against a time series of the log of global atmospheric CO₂ concentrations. The log(CO₂) concentrations measure accumulating greenhouse emissions over time, allowing us to estimate the forced response. As before, since the forced response could be different for each month and location, we fit a separate regression model for the forced response for each location and month pair

$$T_{l,t} = \gamma_m + \beta_{F,m} \log(\text{CO}_2)_t + \nu_t, \quad (4.2)$$

where m indexes the month associated with t , $\gamma_m \in \mathbb{R}$ is an intercept term, and ν_t is i.i.d. mean zero noise with constant variance. We constrain $\beta_{F,m}$ to be positive, to avoid unrealistic negative trends in temperature due to sampling variability over the observational record. Subtracting the estimated forced response, $T_{l,t}^{\text{Forced}} = \hat{\gamma}_m + \hat{\beta}_{F,m} \log(\text{CO}_2)_t$, from $T_{l,t}$ gives an approximately stationary residual series, $\tilde{T}_{l,t}$. We apply the observational large ensemble to this residual time series, then add back in the estimated trend, $T_{l,t}^{\text{Forced}}$ to the synthetic time series to get a realistic temperature record.

This gives the following regression model for the temperature residual time series

$$\begin{aligned} \tilde{T}_{l,t} = & \beta_{0,l,m} + \beta_{\text{ENSO},l,m} \text{ENSO}_t + \beta_{\text{PDO},l,m} \text{PDO}_t + \beta_{\text{PNA},l,m} \text{PNA}_t \\ & + \beta_{\text{NAO},l,m} \text{NAO}_t + \varepsilon_{l,t}, \end{aligned} \quad (4.3)$$

the subscript l indexes the location, the subscript t indexes the months through time, and m is the month corresponding to time t . The error term $\varepsilon_{l,t}$ is assumed to be mean zero with constant variance. Following the prior methodology this gives us the corresponding synthetic model

$$\begin{aligned} T_{l,t}^{\text{Synth}} = & \hat{\beta}_{0,l,m} + T_{l,t}^{\text{Forced}} + \hat{\beta}_{\text{ENSO},l,m} \text{ENSO}_t^{\text{Synth}} + \hat{\beta}_{\text{PDO},l,m} \text{PDO}_t^{\text{Synth}} \\ & + \hat{\beta}_{\text{PNA},l,m} \text{PNA}_t^{\text{Synth}} + \hat{\beta}_{\text{NAO},l,m} \text{NAO}_t^{\text{Synth}} + \varepsilon_{l,t}^{\text{Synth}}, \end{aligned} \quad (4.4)$$

The surrogate modes are generated using the same method as in Chapter 3. Applying the method of Wilks (1997) to calculate a block size for each month and location and taking the

0.97 quantile across all locations and months, we find that a four year block size is sufficient to capture the autocorrelation structure for precipitation and a three year block size is sufficient for temperature.

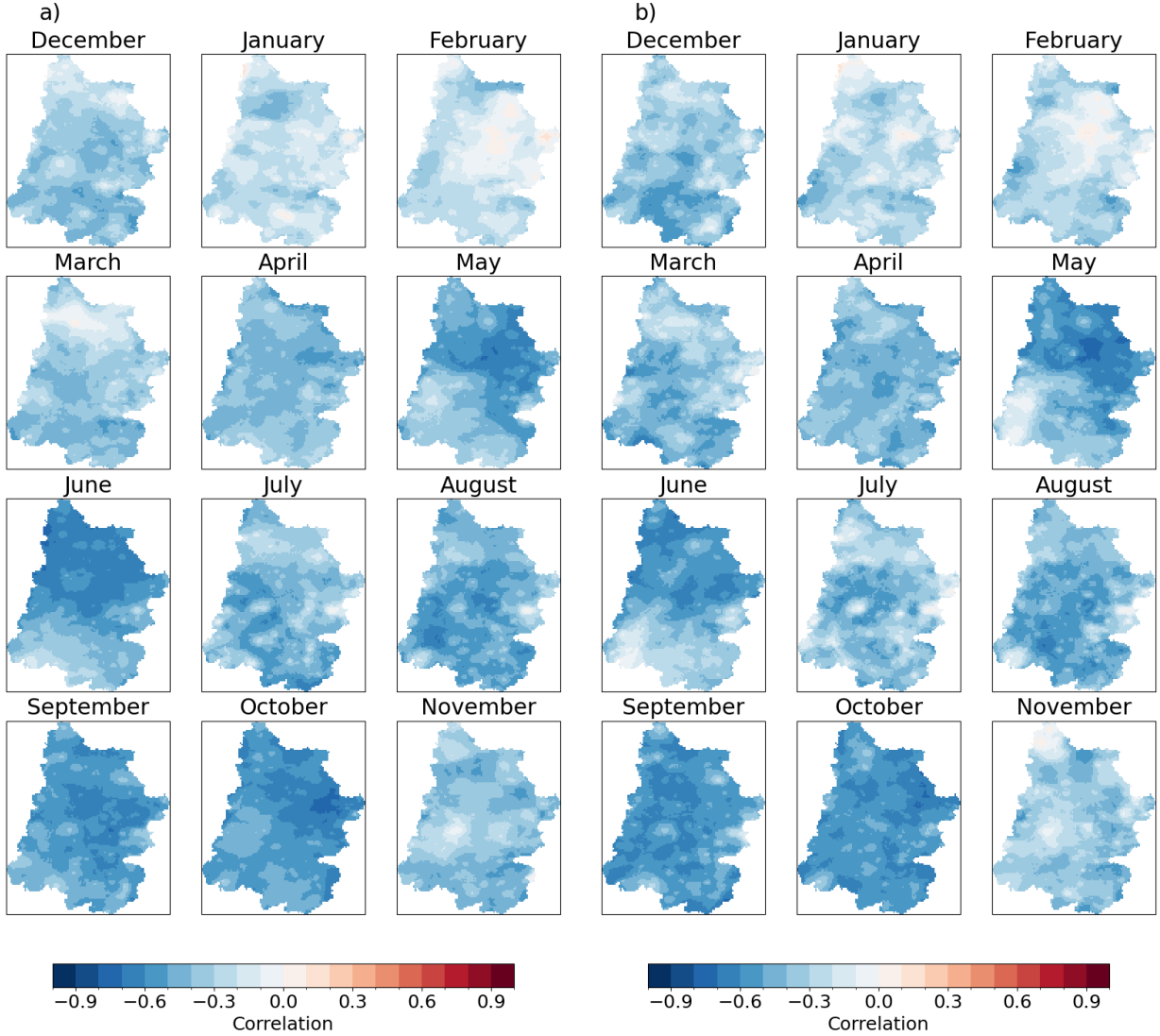


Figure 4.2: a) Gridbox correlations between monthly mean daily maximum temperature and monthly precipitation totals for each month in observations b) The same as a) but for a randomly selected Obs-LE member.

In order for the synthetic records to adequately represent the joint variability of temperature and precipitation, the synthetic ensemble members should match the significant correlations between $T_{l,t}$ and $P_{l,t}$ in the observational record (Figure 4.2a). To capture this joint spatiotemporal structure between temperature and precipitation we use the same surro-

gate modes and the same time blocks of residuals for each corresponding synthetic ensemble member of precipitation and temperature. Therefore, we choose the longer block size of four years for both temperature and precipitation and ensure that the same time blocks are chosen for each variable in the bootstrap samples. While there is some variability between members, following this process does typically produce reasonable correlations between precipitation and temperature (Figure 4.2b). Repeating this process 1,000 times gives us a synthetic ensemble of spatiotemporal temperature and precipitation records that we can use to generate synthetic natural flows at Lees Ferry.

From the large ensemble, we generate composites to analyze the spatial patterns of the impacts of different modes over the UCRB. To create these composites for a particular climate mode, we take all months with corresponding surrogate mode values greater than 1, and take the average of the climate variable over those months to find the positive phase average. We follow the same process for surrogate mode values less than -1 to find the negative phase average. The composite is found by subtracting the negative phase average from the positive phase average. We also generate trend maps by taking the fitted forced response, $T_{l,t}^{\text{Forced}}$, from 2018 and subtracting the value in 1920 from it.

4.2.4 Synthetic Flows

Following the same averaging procedure used for the original observations, we process the synthetic monthly precipitation and temperature records into basin-wide monthly means for temperature and totals for precipitation. Then, we convert these monthly time series into water year mean time series of daily max temperatures and total time series of precipitation, and calculate anomalies, $T_{i,\text{Basin}}^{*,\text{Synth}}$ and $P_{i,\text{Basin}}^{*,\text{Synth}}$, using the mean values from the observations. Plugging these synthetic time series of anomalies into the fitted model of Eqn. (4.1) produces time series of predicted natural flows at Lees Ferry for each ensemble member:

$$\hat{Q}_i^{\text{Synth}} = \hat{\alpha} + \hat{\beta}_P P_{i,\text{Basin}}^{*,\text{Synth}} + \hat{\beta}_T T_{i,\text{Basin}}^{*,\text{Synth}} \quad (4.5)$$

These predicted flows give estimated conditional mean values for given temperature and precipitation anomalies, but in order to fully represent the year-to-year variability of the flows we also need to sample the variability that does not project onto temperature and precipitation.

We can sample this variability by bootstrapping the leverage-weighted residuals of the flow model and adding bootstrapped samples of these residuals to the predicted flow values. We use the leverage-weighted residuals since the original residuals tend to underestimate variability in the errors (Davison and Hinkley, 1997), but most of the weighted residuals are very similar with an average inflation of 1.6% in magnitude and a maximum inflation of 7.5% in magnitude. Applying the block selection procedure of Wilks (1997) to the residuals, the estimated block size is just one year, indicating that there is limited autocorrelation in the residuals. Therefore, we sample each residual independently. However, since there some evidence of heteroscedasticity (Figure 4.4b), we perform a localized bootstrap rather than bootstrapping the full sample of residuals.

To produce the local bootstrap, we first truncate the predicted values $\hat{Q}_i^{\text{Synth}}$ to the range of the fitted values from the original model. Then for each truncated value, $\tilde{Q}_i^{\text{Synth}}$, we construct an interval $(\tilde{Q}_i^{\text{Synth}} - 4, \tilde{Q}_i^{\text{Synth}} + 4)$ (an example window is shown in Figure 4.4b). Residuals from the original model whose associated fitted value falls within this interval are candidates to be selected, and from these residuals we sample one at random with uniform probabilities. Combining the time series of these residuals, e_i^{Synth} with the time series of predicted mean values, we get the time series of synthetic flows Q_i^{Synth} ,

$$Q_i^{\text{Synth}} = \hat{Q}_i^{\text{Synth}} + e_i^{\text{Synth}}. \quad (4.6)$$

Repeating this process 1,000 times, we generate an ensemble of synthetic flows.

From the synthetic flow ensemble, we can generate estimates for distributional quantiles for each water year. Through the Obs-LE, these estimated quantiles integrate the influence of

interannual variability in the temperature and precipitation over the UCRB on streamflows, and through sampling the residuals the quantiles integrate unpredictable variability that is not captured by the regression model. We estimate the quantiles for each water year with the 0.025 and 0.975 sample quantiles of the set of 1,000 synthetic flow values. These quantiles give an estimate of the variability of the distribution for each water year, allowing us to understand how the range of plausible flows changes over time.

These estimates do not account for uncertainty in the fitted coefficients, $\hat{\beta}_P$ and $\hat{\beta}_T$, but this additional source of uncertainty can be accounted for by adding an additional bootstrapping step when fitting model (4.1) and generating synthetic flows that account also account for model uncertainty (see Algorithm 6.4 of Davison and Hinkley (1997)). These bootstrap distributions can be used to find prediction intervals for particular synthetic trajectories as well as confidence intervals incorporating coefficient uncertainty for the ensemble mean and quantile estimates. The bootstrap 90% confidence intervals generated in this way show a mean spread across the water years of 0.66 maf for the mean, 1.9 maf for the 0.975 quantile, and 1.1 maf for the 0.025 quantile (Figures C.1, C.3, C.2). The bootstrap prediction interval generated for a single set of synthetic precipitation and temperature values shows much less spread than the gap between the estimated quantiles (Figure C.4).

4.2.5 Future Projections

Trends in the flows of the Colorado River over the last century and the negative correlation between temperatures and flows suggest the possibility that the flows of the river could decline more substantially under a changing climate, but the large contribution of internal variability to flows could still lead to large interannual variability in streamflows. Given the potential biases of earth system models over this region, we combine observationally derived estimates of variability with future warming trends derived from climate models in order to explore future variability.

In order to use these observationally derived estimates of variability over this future period,

we assume that the variability in temperature, precipitation, and natural flows holds constant over the period 1920 to 2050. And, given model disagreements over precipitation trends in the UCRB (Lehner et al., 2019), we also assume that there is no forced response in precipitation over this time period. In contrast to precipitation, models show consistent warming of temperatures in the UCRB. Therefore, we estimate a forced response in temperatures over 2018 to 2050 using the ensemble mean of the CESM2-LE. The ensemble mean is calculated by subsetting to gridboxes with centers that fall within the borders of the UCRB, taking an area-weighted average of those gridboxes for each time point, and then averaging across the 50 ensemble members. While the CESM2-LE ensemble mean shows significant biases relative to the Livneh dataset (Figure 4.3), we assume that the amount of warming in the ensemble mean over the time period 2018 to 2050 is a reasonable proxy for the amount of warming we would expect to see in the observations. We find this warming by subtracting the ensemble mean value in 2019 from the ensemble mean for each year from 2019 to 2050 and taking a centered, 5 year rolling average of the warming to reduce the interannual variability of the estimate. Finally, for each year of the period 2019 to 2050 the estimate of warming is added to the ensemble mean of the Obs-LE in 2018 to generate a new estimate of the mean under the warming trend (Figure 4.3). To generate the new temperature Obs-LE under warming, a 32 year temperature ensemble is generated with the new mean values under warming substituted for the original estimate of the forced trend. A 32 year ensemble of precipitation is generated using the same process as before, and these ensembles are used to construct synthetic flow ensembles and quantile estimates over the period 2019-2050.

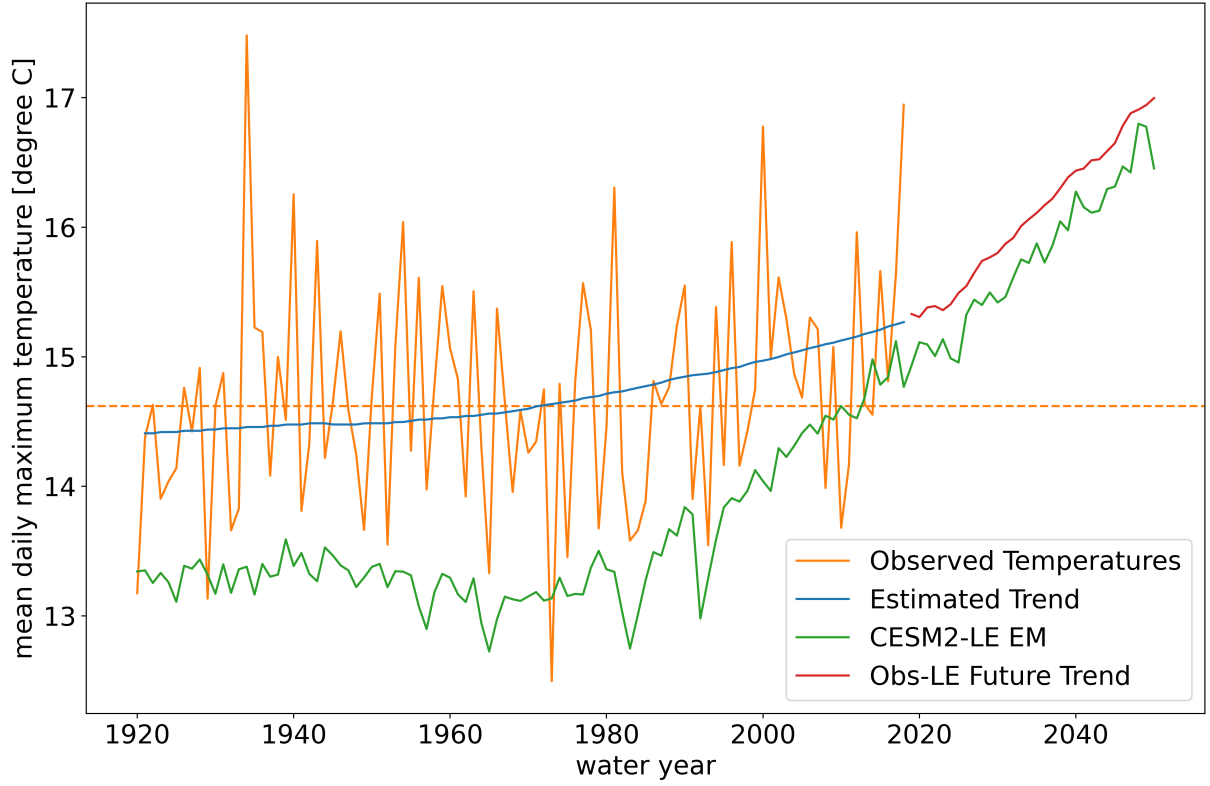


Figure 4.3: Trends in the monthly mean of daily maximum temperatures averaged over the UCRB for the observations and the CESM2-LE ensemble mean. The dashed line gives the mean of the observed temperatures over the time period 1920 to 1970. The estimated trend is the area-weighted and water year averaged mean of the gridbox estimates from Model (4.2). The future trend gives the mean value of the 1,000 ensemble members with the trend extrapolated from the warming in the CESM2-LE ensemble mean.

4.3 Results

4.3.1 Flow Model

The fitted regression model gives the equation

$$\hat{Q}_i = 14.3 - 1.7T_i^* + 14.3P_i^*, \quad (4.7)$$

where Q is in units of million acre-feet (maf), T_i^* is anomalies in units of degree Celsius, and P_i^* is unitless since it is normalized by its mean. The correlation between Q_i and P_i^* is

0.81, while the correlation between Q_i and T_i^* is -0.69 . This correlation indicates a much higher influence of temperature on Lees Ferry streamflows at the annual scale than the -0.26 correlation found in Hoerling et al. (2024) for decadal averages. The values of P_i^* range from -0.40 to 0.44 and P_i^* has a standard deviation of 0.19 . The values of T_i^* range from -2.2 to $2.8^\circ C$ with a standard deviation of $0.86^\circ C$. Setting P_i^* to its minimum and maximum observed values while holding T_i^* constant at the mean would give predicted flows of 8.6 and 20.6 maf, respectively. Doing the same for T_i^* gives values of 9.6 and 18.0 maf. However, since T_i^* and P_i^* themselves have a correlation of -0.59 and have coefficients of opposite signs in the model the interannual variability of the predicted flows tends to be large. The R^2 of the model is 0.72 , and the predictions show good agreement with the observed values. However, the residuals show some evidence of heteroscedasticity with larger residual variability for higher fitted values (Figure 4.4b).

Converting the model values to percentage scale, we find that a $1^\circ C$ increase in daily maximum temperatures, while holding P_i^* at 0 , is associated with an 11.8% decline in natural flows. For a 1% increase in precipitation, while holding T_i^* at 0 , the model predicts a 1% increase in natural flows. These values are similar to the $-11.1\%/^\circ C$ and $1.2\%/ \%$ values for the basin over 1950 to 2008 reported in Lehner et al. (2019).

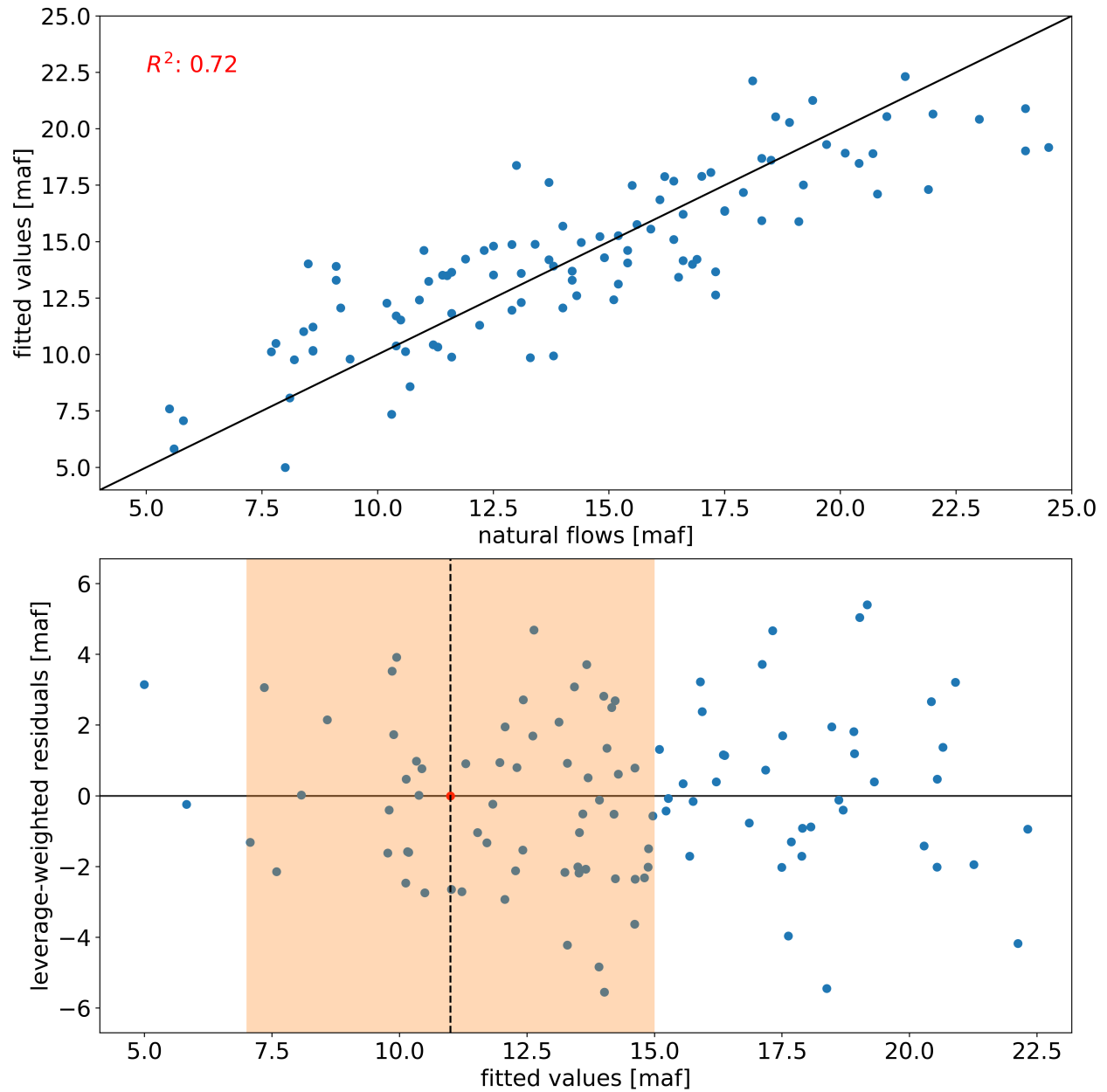


Figure 4.4: a) Fitted values vs. actual values for Model (4.1) fitted to the observational data. b) Leverage-weighted residuals vs. fitted values for Model (4.1). The red point shows a hypothetical fitted value of 11.0 and the colored band surrounding it shows the corresponding window for the localized bootstrapping procedure.

4.3.2 Forced Response and the Influence of Climate Modes

The forced response (Figure 4.5) shows very distinct month-to-month variation with March showing very strong warming of over 2°C nearly uniformly across the basin, while October shows little warming outside of a few tiny pockets. December, May, September, and November all show more limited geographic spread of warming patterns, while June, July, August, January, February show more warming. Without the positivity constraint October, December, and May would primarily show a cooling trend over the basin with more mixed results in other months. The pattern of warming shows geographic coherence across months with similar spatial patterns of warming. Some areas with higher elevations (see Figure 4.1a)) appear to show less of a warming trend than lower lying areas, though the pattern is not apparent everywhere.

Monthly ENSO composites show that for most months there are predominantly wetter conditions across the basin (Figure 4.6a), though the December-January-February (DJF) season shows some drying in the north, with larger negative values concentrated in regions with higher elevation. The temperature composites for ENSO show mostly cooler conditions outside of February and May with the strongest cooling concentrated in the south of the basin and in October, November, March and April (Figure 4.6b).

Monthly PDO composites show weaker wetting signals than ENSO with July, August, October, and November months showing a combination of weak wetting and weak drying, though February shows strong wetting at higher elevations. In daily maximum temperature, the PDO composites show strong seasonal differences with DJF and March-April-May (MAM) seasons show very strong warming signals with large regions of over 1.5°C warmer temperatures in the positive phase of PDO. June-July-August (JJA) show mostly cooler temperatures during the positive phase of PDO. September-October-November (SON) show milder differences with a mixture of warming and cooling.

Monthly PNA precipitation composites (Figure 4.8) show strong drying signals in December and January, wetting signals in June and September with moderate signals in the

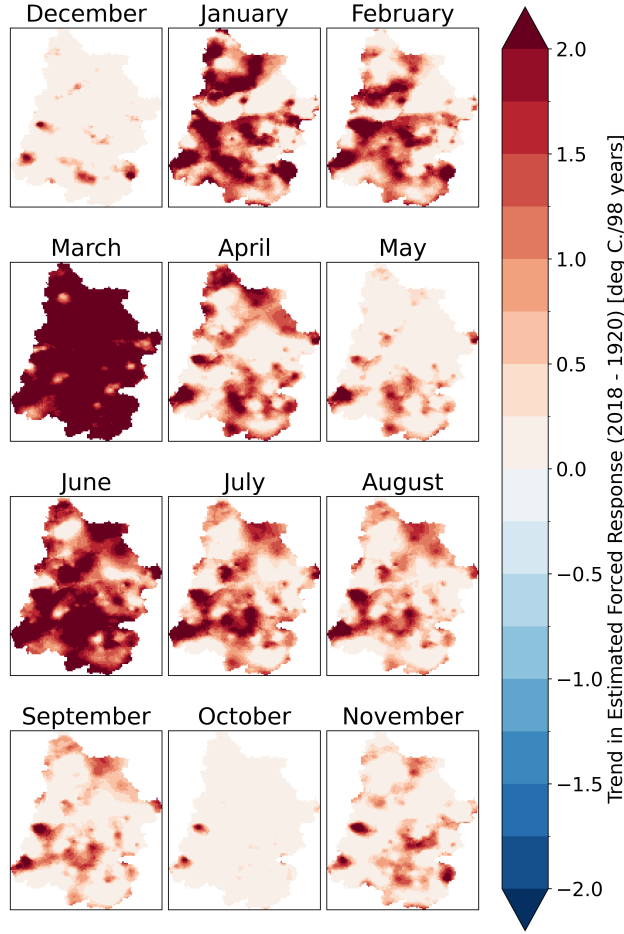


Figure 4.5: Trends in the estimated forced response in mean monthly daily maximum temperature from 2018 to 1920. The forced response is estimated as a linear regression of mean monthly daily maximum temperature against $\log(\text{CO}_2)$ with the constraint that the coefficients must be positive.

remaining months. PNA mean daily max temperature composites show strong warming in DJF and cooling in JJA. NAO composites show weak precipitation signals, but show very strong warming in MAM and June and July.

These composites show the spatial effects of the modes on the precipitation and temperature fields for each month, but the modeled synthetic flows depend on basin-wide spatial averages and water year temporal averages of these fields. Figure 4.10a shows that the generally positive precipitation impacts of ENSO combined with the significant cooling effect over most months leads to higher mean modeled flows during positive ENSO phase years, though the mean difference of 2.8 maf between the positive phase and negative phase is

smaller than the standard deviations of 4.5 and 4.0 maf associated with the positive and negative phases. ENSO leads to the largest gap in mean values between the positive and negative phases of the modes, though the PNA, which is 2.4 maf wetter on average during the negative phase compared to the positive phase, is a close second. Each of the modes results in a gap of at least 1.2 maf between the positive and negative phases with the PDO resulting in the smallest gap between the two phases. Flipping the signs of the composites in Figure 4.8 to match the negative phase of the PNA shows that the strong wetting and cooling signals during DJF during the negative phase outweighs the compensating effects in other months to produce the wetter negative phase PNA conditions.

4.3.3 Mean Flows and Quantile Estimates

The quantile estimates of flows show a declining trend over the observational time period, 1921 to 2018, that accelerates as warming intensifies (Figure 4.11). The mean flow of the ensemble of synthetic flows drops from 14.8 maf in water year 1921 to 13.2 maf by 2018, an 11% decline. The 0.975 quantile estimate declines from approximately 23.8 maf to 21.8 maf by 2018 and the 0.025 quantile estimate declines from 6.9 maf to 5.4 maf. While there is significant interannual variability in the quantile estimates, five year windowed averages of the estimates show a small but steady decline over the observational time period associated with the warming trend.

Projecting flows from 2019 to 2050 with the estimated warming trend we find much more substantial declines due to warming with the mean synthetic flow dropping from 12.98 maf in 2019 to 10.4 maf in 2050, a decline of 19.9%. The 0.975 quantile also declines further, dropping to 18.1 maf a 16.6% decline relative to 2019, while the 0.025 quantile declines to just over 3.4 maf. Comparing the values in water year 2050 to the values in 1921, we find a 29.9% decline in the mean synthetic flows, along with a 23.9% decline in the 0.975 quantile and a 50.1% decline in the 0.025 quantile.

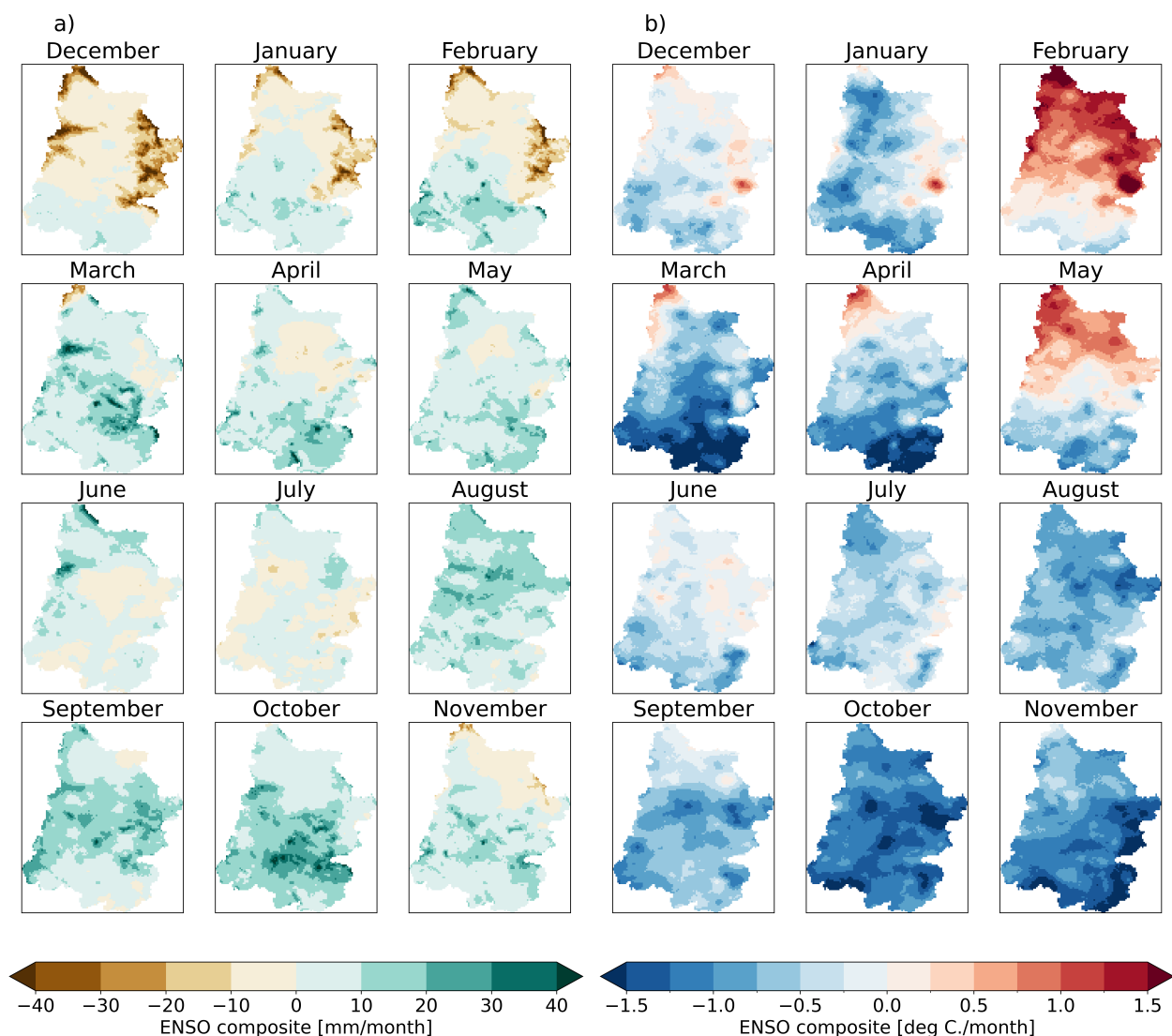


Figure 4.6: a) Monthly ENSO composites of precipitation in the Obs-LE. b) Monthly ENSO composites for temperature in the Obs-LE. For each month, all Obs-LE months where the corresponding surrogate ENSO mode was greater than 1 were selected and averages were taken of these to get the positive phase ENSO average. The same was done with months with surrogate ENSO modes that were less than 1 to produce negative phase ENSO averages. The composite is formed by subtracting the negative phase average from the positive phase average.

4.3.4 Three-year Totals

In addition to the annual variability of flows, multi-year accumulations of streamflows play a key role in managing water resources on the Colorado River. Stretches of multiple consecutive below-average years can result in dropping levels at the Lake Powell and Lake Mead reservoirs,

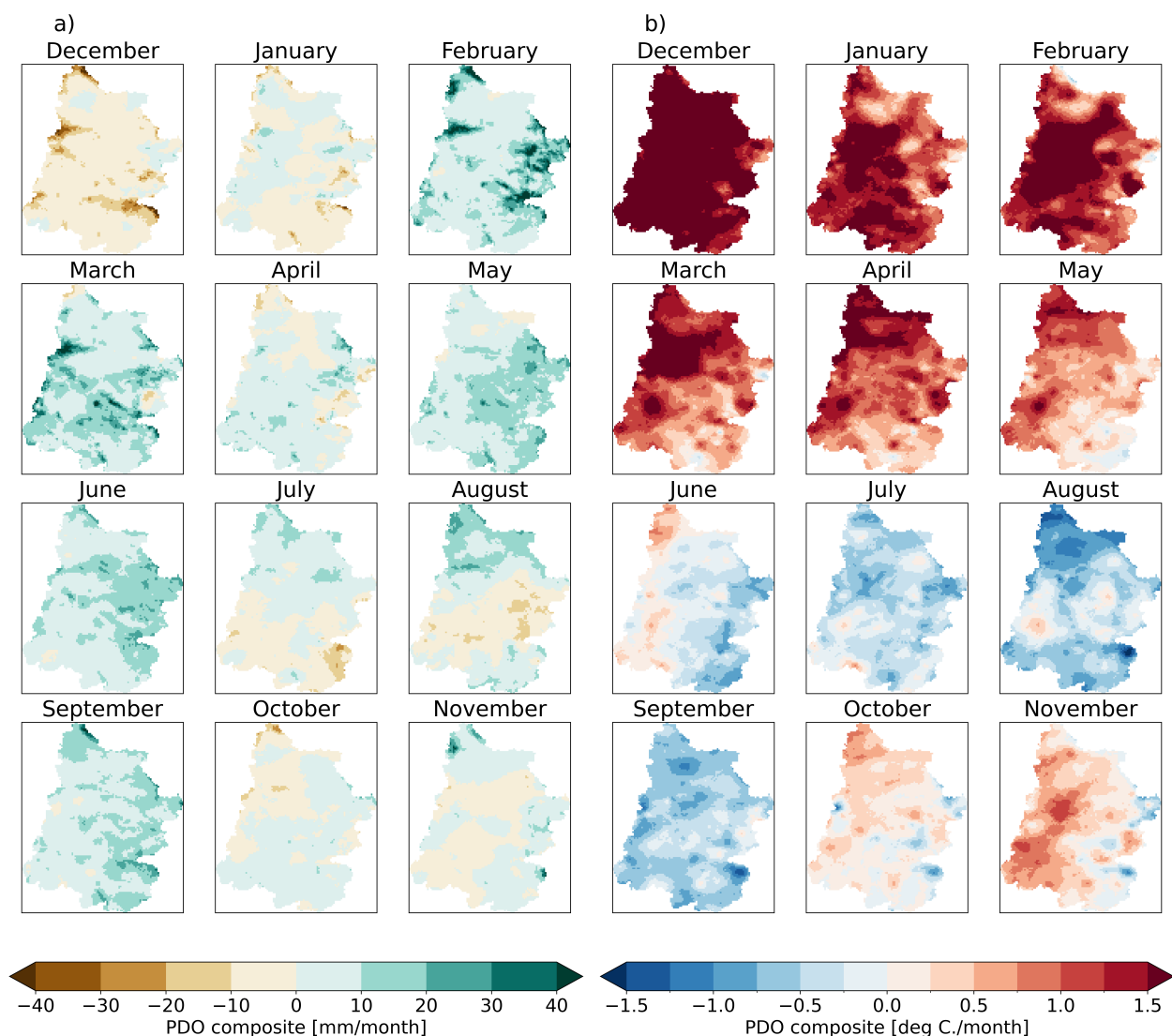


Figure 4.7: a) Monthly PDO composites of precipitation in the Obs-LE. b) Monthly PDO composites for temperature in the Obs-LE. For each month, all Obs-LE months where the corresponding surrogate PDO mode was greater than 1 were selected and averages were taken of these to get the positive phase PDO average. The same was done with months with surrogate PDO modes that were less than 1 to produce negative phase PDO averages. The composite is formed by subtracting the negative phase average from the positive phase average.

requiring more drastic cuts to water deliveries. The recent stretch of flows between water years 2020 and 2022 was the lowest on record with a three-year total of just 24.9 maf, though the three-year stretch ending in 2004 was a close second at 25.1 maf. These flows were well below the 42.68 maf average three-year sum between 1920 and 2018.

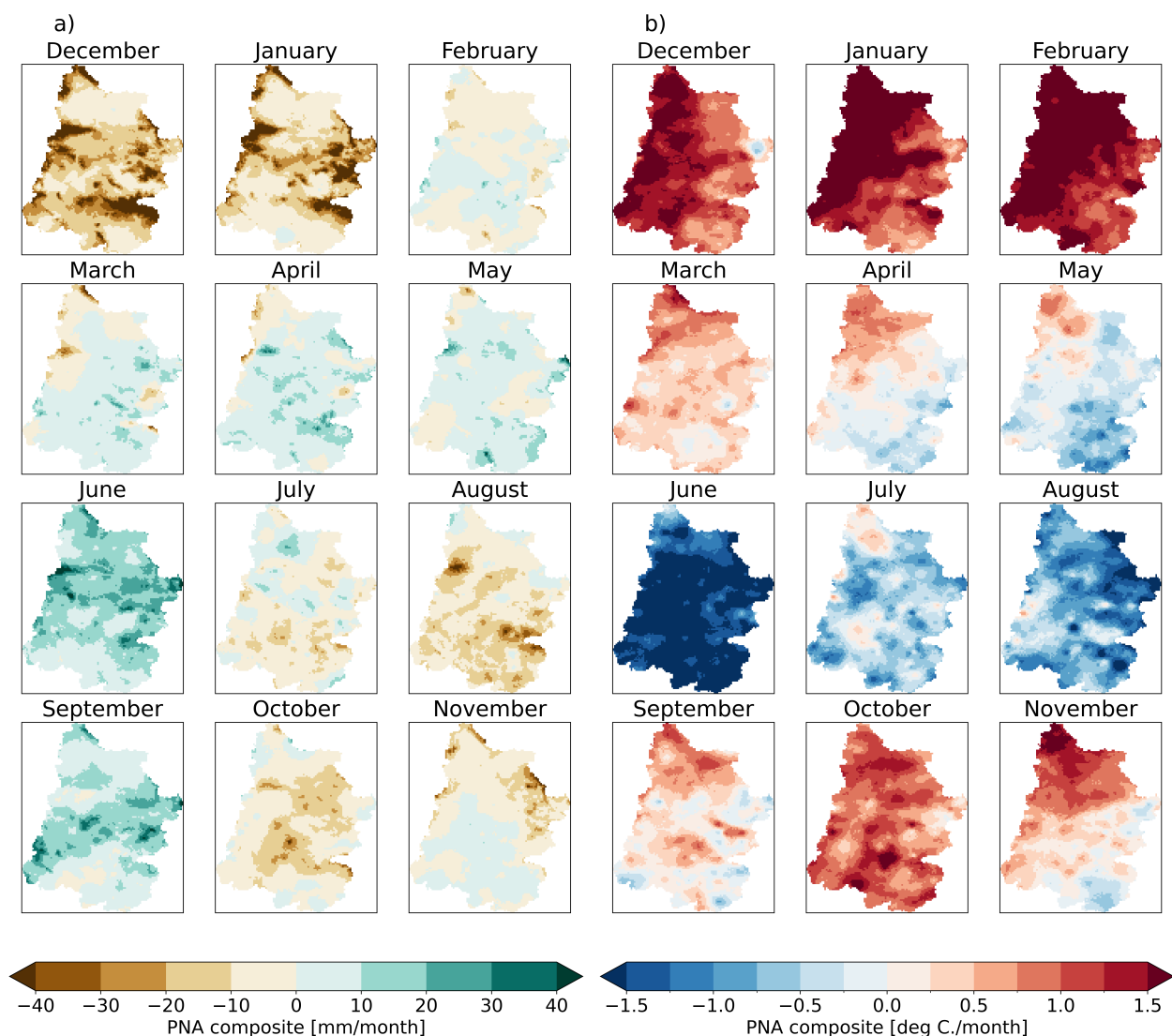


Figure 4.8: a) Monthly PNA composites of precipitation in the Obs-LE. b) Monthly PNA composites for temperature in the Obs-LE. For each month, all Obs-LE months where the corresponding surrogate PNA mode was greater than 1 were selected and averages were taken of these to get the positive phase PNA average. The same was done with months with surrogate PNA modes that were less than 1 to produce negative phase PNA averages. The composite is formed by subtracting the negative phase average from the positive phase average.

To investigate how anomalous this three-year total was we compare its value to the distribution of three-year totals simulated in the synthetic streamflows (Figure 4.12). For three-year synthetic totals ending in 2022, we find that only 2.3% of simulated totals are smaller than this value, and the 0.1 quantile of the synthetic three-year totals ending in 2022

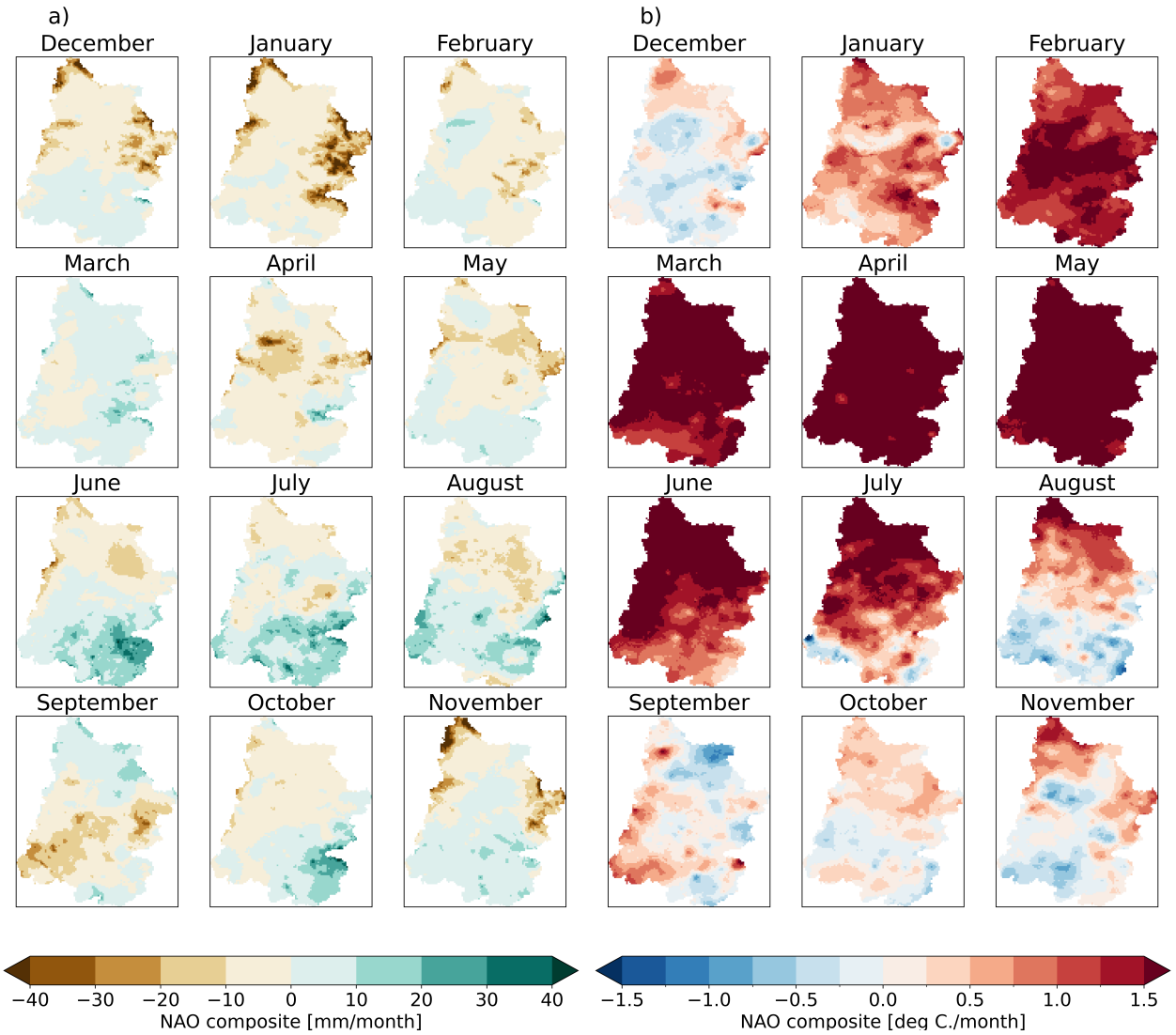


Figure 4.9: a) Monthly NAO composites of precipitation in the Obs-LE. b) Monthly NAO composites for temperature in the Obs-LE. For each month, all Obs-LE months where the corresponding surrogate NAO mode was greater than 1 were selected and averages were taken of these to get the positive phase NAO average. The same was done with months with surrogate NAO modes that were less than 1 to produce negative phase NAO averages. The composite is formed by subtracting the negative phase average from the positive phase average.

is 29.7 maf. However, under future warming we expect that smaller three-year totals become more common-place with a significant shift in the distribution by 2050 (Figure 4.12). In 2050, 16.9% of simulated three-year totals are smaller than the observed 2022 value, and the 0.1 quantile value is 22.9 maf. While large multi-year accumulations are still possible in 2050,

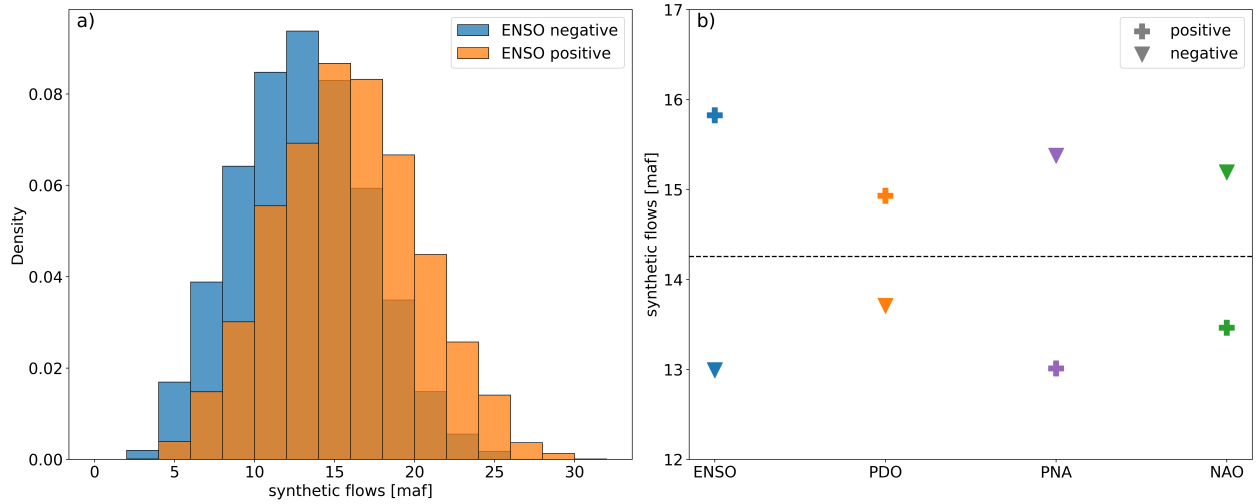


Figure 4.10: a) Histograms of synthetic flows for water years with ENSO in positive phase vs. negative phase. Where positive phase is defined via the water-year average ENSO index being above 1 and negative phase is defined via the average being below -1 . b) Mean synthetic flows for the positive and negative phases for each mode. The dashed line shows the overall mean synthetic flow.

they become much rarer in the simulated distribution, putting further stress on the ability of the reservoirs to recover after dry stretches.

4.4 Discussion and Conclusion

The primary goals of this work were to quantify internal variability in the natural flows of the Colorado River at Lees Ferry through annual quantile estimates and to investigate how warming over the basin could be responsible for declining flows over the observational record and into the future. Based on our model, we find that warming over the time period 1921-2018 may have resulted in significant declines in mean flows of 11%. This is comparable to the linear trend fit in Hoerling et al. (2019), where a 20% decline was estimated over the longer observational 1896-2018 but the fit included the wetter period at the beginning of the 20th century.

Projecting out to 2050 we find much more substantial declines in river flows with an additional decline of 16.6% in mean natural flow in 2050 compared to the estimated mean

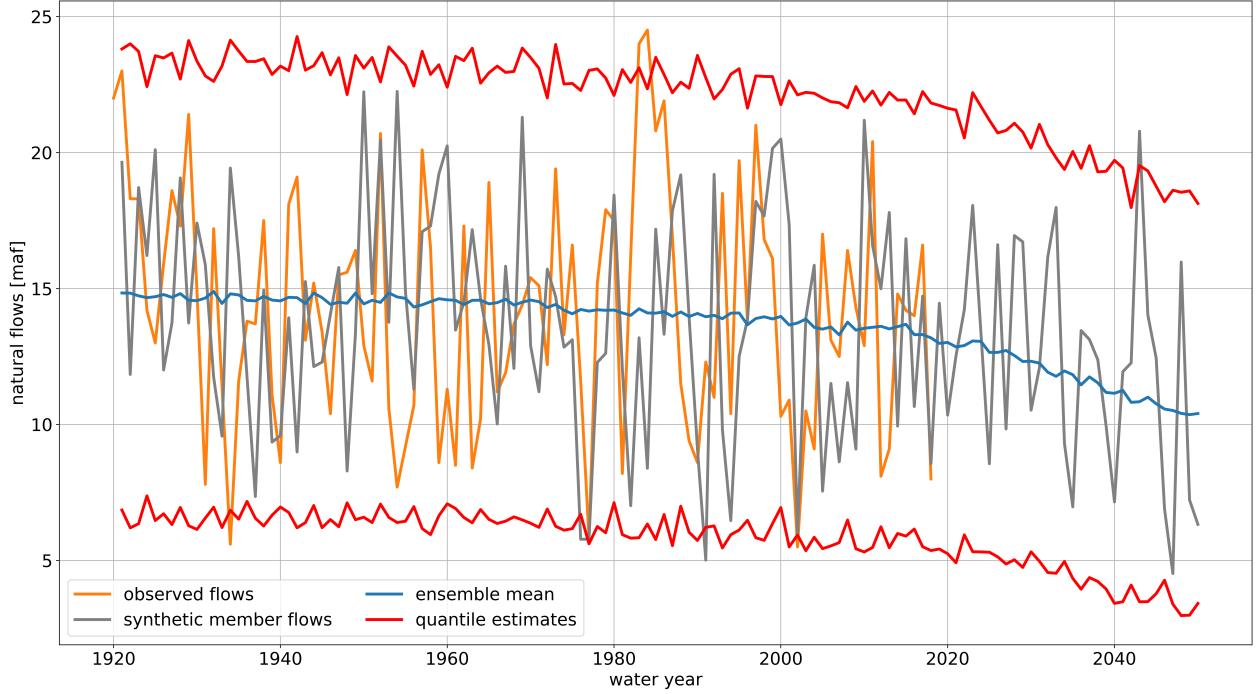


Figure 4.11: The observed natural flows at Lees Ferry from 1920 to 2018 are plotted along with a member of the synthetic flow ensemble, the mean of the synthetic flow ensemble and estimates of the 0.975 and 0.025 quantiles derived from synthetic flows.

in 2019. Comparing the average synthetic flows over 2000 to 2025 to the flows over 2026 to 2050, we find a 14.5% decline in flows, comparable to results from climate models under middle temperature sensitivity conditions with the precipitation sensitivity set to 0 (Figure 6b of Hoerling et al. (2024)).

There are a number of limitations of the current work. First, we did not consider uncertainty in the estimation of the coefficients for the Obs-LE. This uncertainty and its effects on the results could be estimated by bootstrapping the observational record of precipitation and temperature and pushing those values through the rest of the analysis. However, performing this uncertainty analysis would be computationally intensive. Another alternative would be to fit a spatial model for the coefficients and use that to estimate coefficient uncertainty.

Another limitation is that we have not considered a potential forced response in precipitation. The signal-to-noise ratio of the forced response in precipitation is often low at the

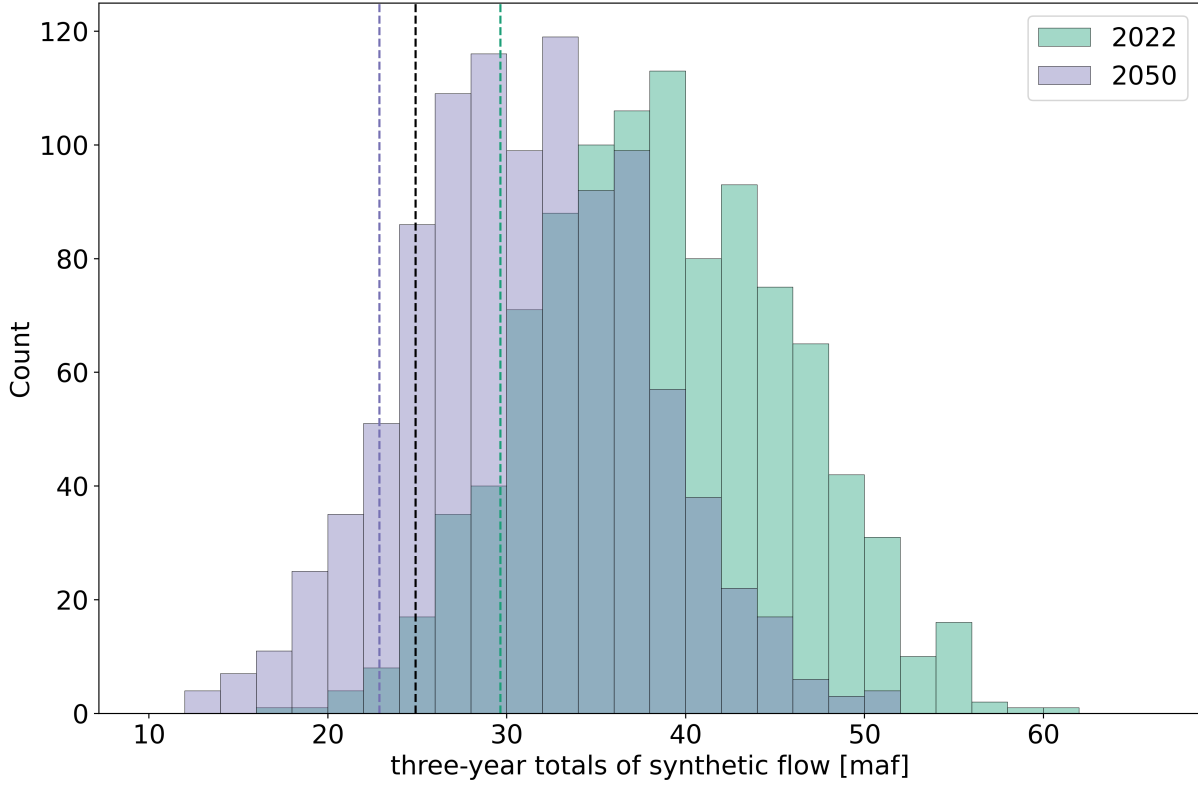


Figure 4.12: Histograms of three-year totals of synthetic flow compare the distribution of totals ending in 2022 with the distribution in 2050. The black dashed line shows the observed three-year total ending in 2022, while the colored dashed lines show the 0.1 sample quantiles of the respective distributions.

gridbox level even at coarser resolution than considered here (McKinnon and Deser, 2021), making it difficult to accurately fit the response over the span of our current observational records. In contrast to temperature, climate model initial condition large ensembles and multi-model ensembles show limited forced response in the ensemble mean of precipitation over the 20th century in the basin, but future projections in these ensembles show increases in precipitation between 2000 and 2050 (Fig. 3 of Hoerling et al. (2024)). However, biases remain in model representations of precipitation over the basin (Pierce et al., 2022).

As another caveat, when we generate the synthetic modes we assume that the observations are sampled from a stationary process, which implicitly assumes that there is no forced signal in the climate modes. Some research with climate model ensembles has suggested that under

strong warming scenarios there are forced changes in ENSO (Maher et al., 2018), so this assumption may not be reasonable, especially for the period from 2019 to 2050. In addition, Kuo et al. (2025) found that radiative forcing from anthropogenic aerosols was interacting with shifts in SSTs to reduce precipitation over the southwestern US during positive phase ENSO. This finding runs contrary to our assumptions that there is no forced response in precipitation and that the impacts of the modes are constant through time.

In future work, the impact of potential shifts in the precipitation mean and variability on projected flows could be explored further. We could take a range of estimates of the forced response in precipitation from climate models, and explore how those estimated forced responses could alter the results found here. This may be especially important for projecting a range of scenarios out into the future.

Another consideration for future work would be to increase the complexity of Model (4.1). Different parts of the basin could have different run-off efficiencies and different evapotranspiration responses. By accounting for these regional differences we might be able to find a better and more interpretable model for streamflows, and this model could be used to better understand how spatial variability in the basin affects streamflows.

Given the fraught negotiations between upper and lower Colorado River basin states over allocations of water, these estimated declines in mean streamflows could pose significant further challenges to the ecosystems, cities, and farmers that rely on the Colorado River. Despite the significant declines in the mean and the quantiles, there is still a large amount of spread in the distribution of streamflows in 2050. This reflects the large interannual variability in UCRB precipitation and the strong contribution of precipitation to streamflows. However, under a warmer climate we project that extremely dry years, below the 5.5 maf minimum of the observational record, become much more likely. Both the concerning declines projected here and in other studies and the substantial interannual variability we've estimated should be accounted for in future plans for Colorado River water usage to avoid more devastating cuts without sufficient preparation.

Chapter 5

Discussion

The methods presented in this thesis approach climate variability and the spatiotemporal structure of climate data from different angles, but each involves a careful consideration of how that structure can be accounted for or leveraged to improve the statistical methods. Effectively doing so requires an interdisciplinary approach that combines statistical expertise with domain knowledge of climate science. In Chapter 2, the knowledge that the regression coefficients in the hypothesis testing problem often vary smoothly in space for climate problems made a smoothing approach a natural one. Understanding that the spatial covariance of climate data is not stationary across space, led to a local approach, allowing the method to adapt to different covariance structures in different regions. This prevented excessively smoothing regions where climate variables operate on smaller spatial scales, preserving important features of the data in the final hypothesis tests. In Chapters 3 and 4, understanding that the climate modes are important drivers of variability in temperature and precipitation made it key that the methods incorporated variability in the sequence and combinations of climate modes in order to properly simulate the internal variability of the climate system. Failing to do so would have risked creating synthetic records that did not properly represent the full of variability of the climate on the timescales of each modes. In each case, knowledge from climate science informed the methods, enabling them to better adapt to the problem at hand.

Another commonality between the chapters is that the existing sample of climate data is small compared to the amount of variability in the data, leading to challenges in inference and uncertainty quantification. In Chapter 2, this tension was resolved by sharing spatial information across locations, reducing variability without increasing the sample size. In Chapters 3 and 4, synthetic samples of data were generated using resampling methods to enable alternative samples of the climate. In each case, a larger sample of data would resolve many of the challenges, improving the quality of regression inferences or the generation of synthetic data. With a large enough sample these methods might be unnecessary, though, unfortunately, collecting additional observational climate data will require a bit of a wait.

However, there are a few alternatives to observational data that are used in climate. One alternative is paleoclimate reconstruction, which uses the relationship between proxy variables, such as tree ring widths or the ratios of chemical isotopes found in shells, and variables of interest to reconstruct past climate records. Unfortunately, these reconstructions have their own biases and the proxies are often sparse in space and time. Another possibility is the use of climate models. While the current biases of climate models make them an insufficient stand-in for observational data, it's possible that as models continue to develop there will be a point in the future where model output will be an appropriate substitute for observational data. In another direction, as neural net architectures continue to improve, it's also possible that they will lead to a breakthrough in production of synthetic climate data. In the meantime, we will have to find creative ways to take advantage of the observational records that we have.

Resampling methods, which form the basis for the methods of Chapter 3 and 4, are a core methodology in modern statistics, used in procedures from bootstrap confidence intervals, to permutation tests, to Monte Carlo simulations. In these chapters, the resampling approach enabled the retention of complex spatiotemporal structure in the synthetic ensembles. While another model-based spatiotemporal method, such as Gaussian processes, could have been employed, it is difficult to estimate the complex structure of climate data with a model. And

these models could have led to bias or unrealistic variability in the synthetic output. By employing resampling, we are able to side-step the challenges of producing a model for the residual variability, and instead retain the structures that were in our original sample of data. In Chapter 4, multiple layers of resampling were performed: one in the observational large ensemble and one to generate the synthetic flows. This highlights the flexibility of these methods in practice, and their ability to leverage the limited sample of data that we have.

In future work, the Observational Large Ensemble methods of Chapter 3 could be extended to work with daily precipitation data, but applying these methods to daily data would require resolving some challenges. First, daily data contains correlated fields of zero and nonzero precipitation that have varying spatial structures. Modeling this pattern of zeros or attempting to resample it would likely lead to interesting complications. Second, daily data has a much more highly skewed structure than monthly data, making the transformed regression approach more risky. The nonzero precipitation could instead be modeled using a Gamma regression or another generalized linear model, but resampling residuals is not as straightforward in this setting. If these challenges can be resolved, producing synthetic spatiotemporal fields of precipitation at the daily scale would be highly useful for climate scientists and stakeholders. For Chapter 3 and Chapter 4, future work could include estimating changes in temperature and precipitation variability in climate models and incorporating those variability changes into the synthetic ensembles. This would provide important additional insights into how extremes and Colorado River variability could change with the changing climate.

A major theme of this work is that the internal variability of the climate could have led to vastly different observational records at scales both large and small. This knowledge should give us caution when interpreting the results of inferences based on a single record, even when we try to account for that variability in our methodology. In order to sufficiently prepare for precipitation extremes or the possibility of ongoing drought in the Colorado River Basin we need to meaningfully engage with uncertainty in our estimates and the variability of these processes. We should also acknowledge the limitations of our analyses and consider the

possibility of events that have not previously been observed or appear unlikely in our models. As the climate continues to change the ability of the observational record to meaningfully inform our inferences about the future will only decrease and our caution should therefore increase. While there are significant risks ahead, collaborations between statisticians and climate scientists enable us to more meaningfully understand the inherent uncertainties that we will face.

Appendices

Appendix A

Supplemental Materials Chapter 2

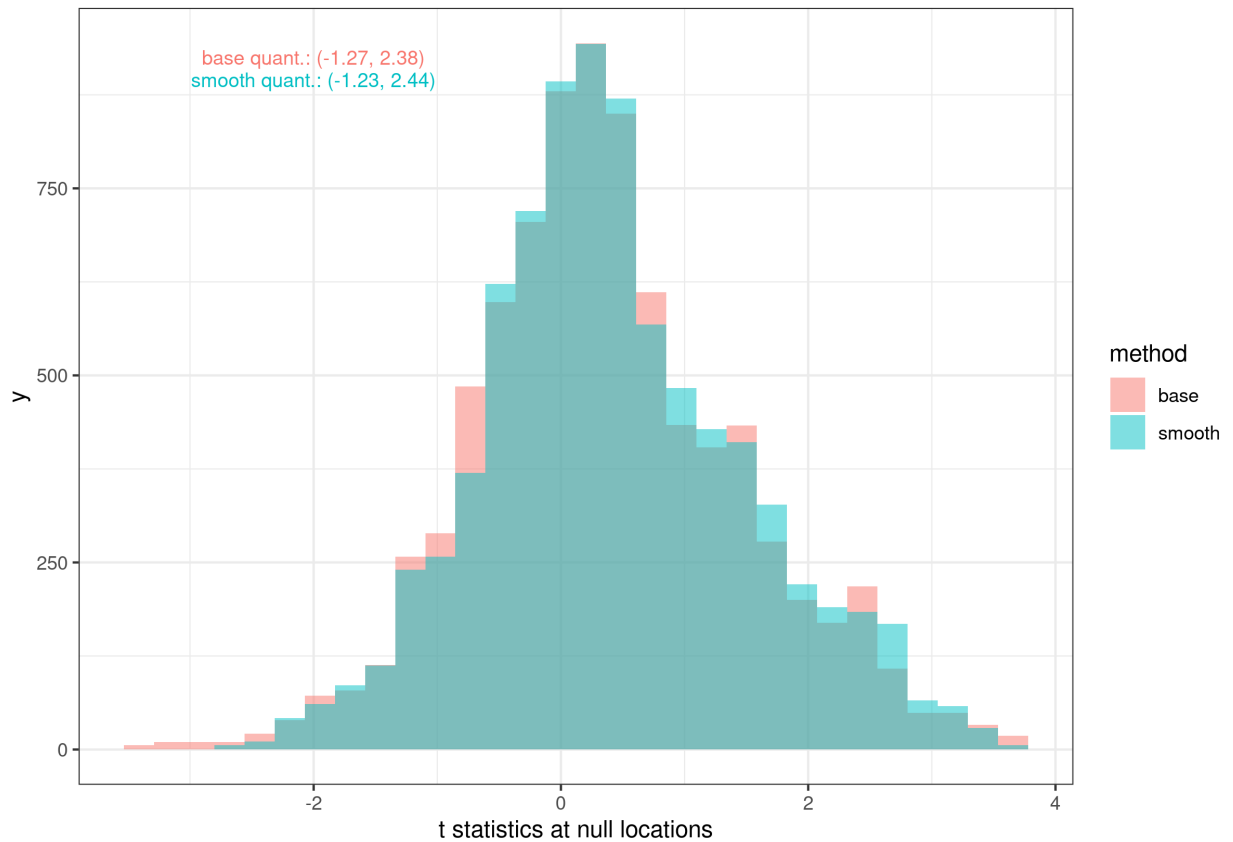


Figure A.1: Histograms of t -statistics at null locations for the base and smooth method. The t -statistics are from a single iteration of the simulations run with the baseline parameter choices. A single iteration is plotted, but the pattern is typical of other iterations. The 0.05 and 0.95 quantiles of each distribution are labeled in the top left corner.

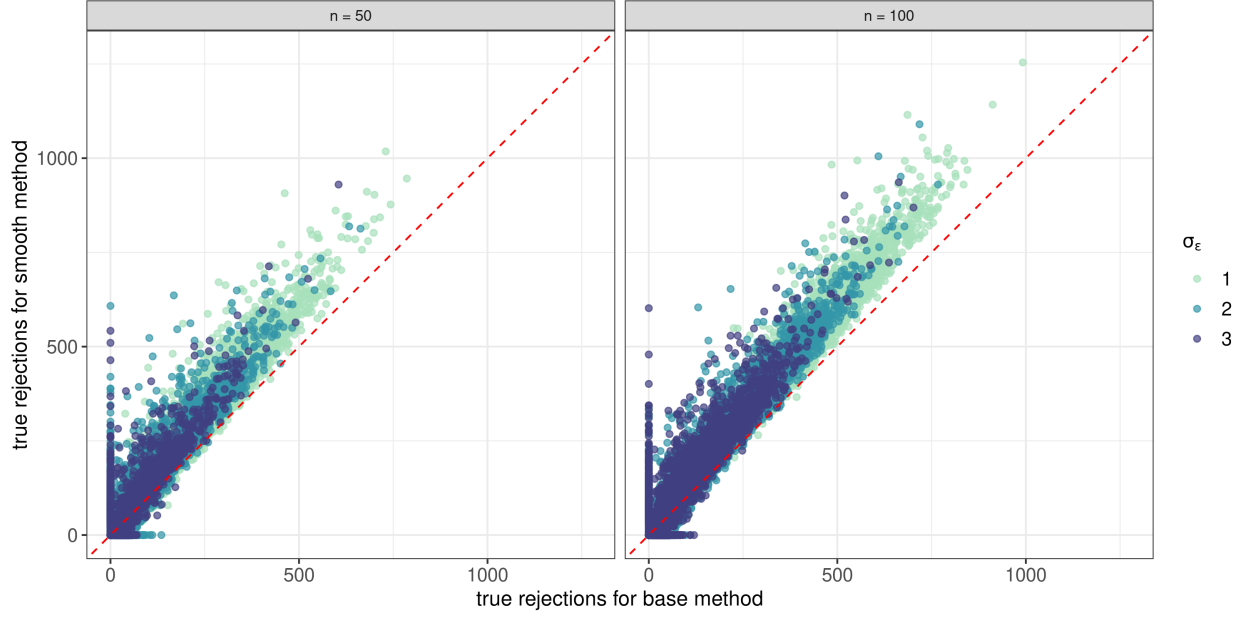


Figure A.2: True rejections made by each method in the simulation studies varying σ_ε and n . The dashed line shows $y = x$.

method	true rej.	FDP	sens.
base	204.5	0.064	0.073
YB	207.2	0.069	0.074
smooth	253.2	0.062	0.091

Table A.1: Results applying the method of Yekutieli and Benjamini (1999), labeled as YB in the table, to the baseline simulation. Values are mean values over 200 iterations generated from 10 spatial fields and 20 random y samples per field.

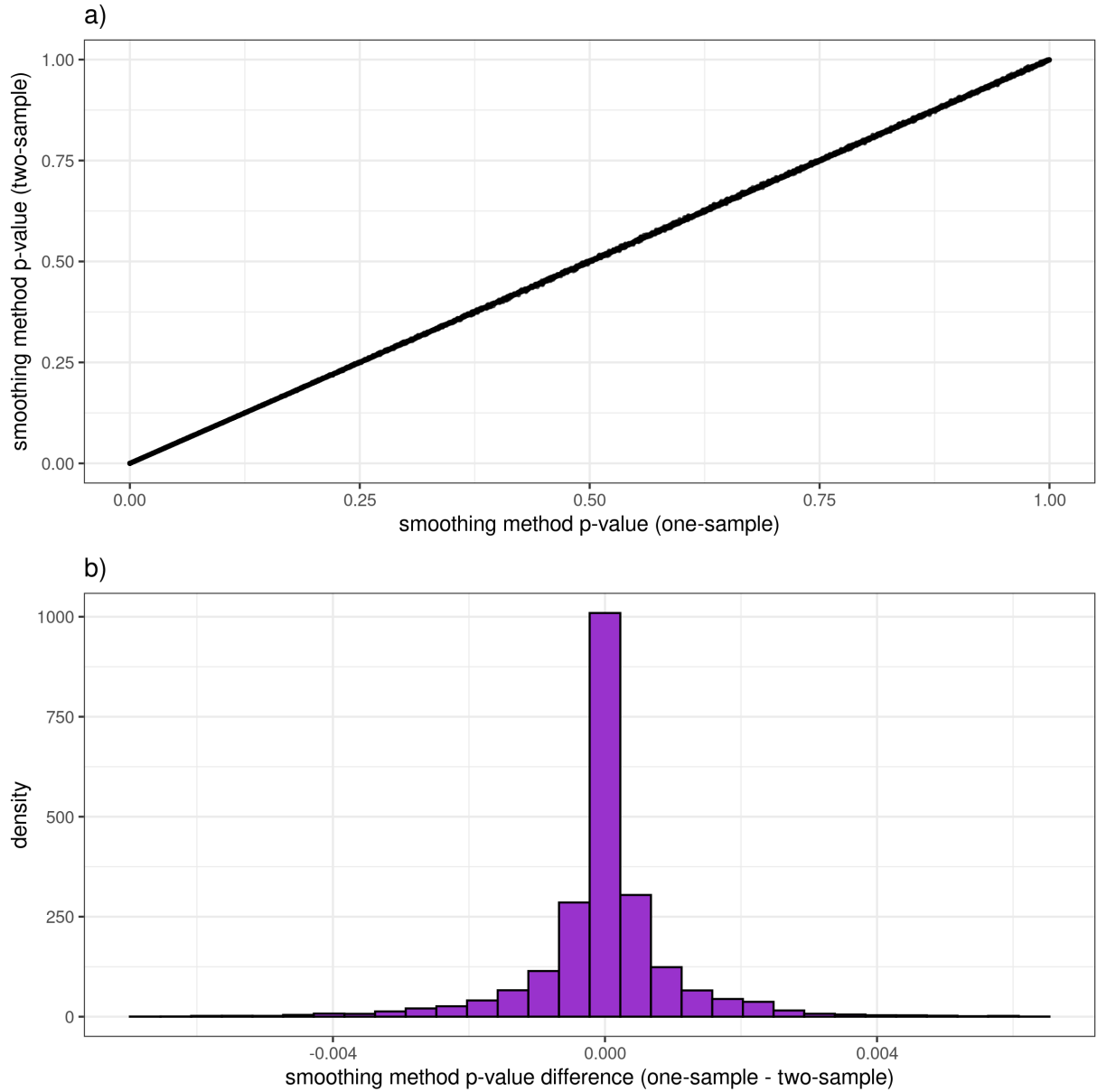


Figure A.3: Plots comparing two sets of smoothing method p -values from a single iteration of the simulations run with the baseline parameter choices. The first set was generated using smoothing weights that were fit as described in the paper, using the same sample of covariate data for fitting the smoothing as was used in the hypothesis tests. The second set used a second independent sample of the covariate data for fitting the smoothing weights. A single iteration is plotted, but the pattern is typical of other iterations. a) Scatterplot comparing the one-sample and two-sample p -values. b) Histogram of the differences in each location's p -values between two versions of the smoothing method.

window	σ_ε	FDP	true rej.	sens.
4	1	0.341	400.7	0.143
4	2	0.439	227.6	0.081
4	3	0.541	133.4	0.048
20	1	0.501	662.8	0.237
20	2	0.572	492.2	0.176
20	3	0.638	379.5	0.136

Table A.2: Results applying the FDR_L method of Zhang et al. (2011) to the baseline simulation study varying σ_ε . Neighborhood sets are chosen analogously to how the local windows for smoothing were selected in our method. The window column gives the length of neighborhood window edges in degrees. The median filter was used and the λ parameter of their method was set to 0.1. The values are means over 10,000 iterations.

method	window	σ_ε	n	ϕ	θ	$\sqrt{\psi}$	γ	FDP	s.d.	FDX	s.e.
base	12	1	50	4	0	1	7	0.071	0.119	0.255	0.004
smooth	12	1	50	4	0	1	7	0.071	0.124	0.249	0.004
base	16	1	50	4	0	1	7	0.064	0.111	0.231	0.004
smooth	16	1	50	4	0	1	7	0.062	0.119	0.220	0.004

Table A.3: Simulation results showing the false rejection metrics of the proposed smoothing method in the baseline simulation study with smoothing windows of $12^\circ \times 12^\circ$ and $16^\circ \times 16^\circ$ rather than $20^\circ \times 20^\circ$. The window column gives the length of the smoothing window edges in degrees longitude and latitude. The metric values are means (or standard deviations) over 10,000 iterations of the simulation study run with the baseline parameters. For the FDX the standard error of the empirical probability is shown rather than the standard deviation.

method	window	true rej.	s.d.	sens.	s.d.	% Δ sens.
base	12	212.4	104.26	0.076	0.039	-
smooth	12	238.3	116.06	0.086	0.044	12.0
base	16	211.8	105.27	0.076	0.040	-
smooth	16	251.5	125.28	0.090	0.047	18.6

Table A.4: Simulation results showing the power metrics of the proposed smoothing method in the baseline simulation study with with smoothing windows of $12^\circ \times 12^\circ$ and $16^\circ \times 16^\circ$ rather than $20^\circ \times 20^\circ$. The window column gives the length of the smoothing window edges in degrees longitude and latitude. The parameters for the simulations are given in Table S3. The metric values are means (or standard deviations) over 10,000 iterations of the simulation study run with the baseline parameters.

β_1	β_2	method	FDP	s.d.	FDX	s.e.	true rej.	s.d.	sens.	s.d.	% Δ sens.
1.0	1.0	base	0.050	0.087	0.181	0.004	321.7	171.3	0.069	0.041	-
1.0	1.0	smooth	0.049	0.094	0.184	0.004	413.4	221.9	0.089	0.053	28.3
1.5	0.5	base	0.053	0.090	0.202	0.004	300.1	122.8	0.063	0.030	-
1.5	0.5	smooth	0.052	0.099	0.199	0.004	377.2	160.1	0.080	0.039	25.6

Table A.5: Simulation results showing the evaluation metrics of the proposed smoothing method in the baseline simulation study with two randomly selected locations for generating y with coefficients β_1, β_2 , respectively. The values are means (or standard deviations) over 10,000 iterations of the simulation study run with the baseline parameters. For the FDX the standard error of the empirical probability is shown rather than the standard deviation.

method	σ_ε	n	ϕ	θ	$\sqrt{\psi}$	γ	FDP	s.d.	FDX	s.e.
base	1	50	4	0	1	7	0.061	0.108	0.222	0.042
smooth	1	50	4	0	1	7	0.058	0.118	0.204	0.040
two-sample smooth	1	50	4	0	1	7	0.058	0.118	0.205	0.040

Table A.6: Simulation results comparing the false discovery control of the proposed smoothing method as presented in the paper with the scenario where the smoothing weights are fit on a second independent sample of the covariate data. The values are means (or standard deviations) over 10,000 iterations of the simulation study run with the baseline parameters. For the FDX the standard error of the empirical probability is shown rather than the standard deviation.

method	σ_ε	n	ϕ	θ	$\sqrt{\psi}$	γ	true rej.	s.d.	sens.	s.d.
base	1	50	4	0	1	7	211.1	104.2	0.076	0.040
smooth	1	50	4	0	1	7	264.0	131.4	0.095	0.049
doublesmooth	1	50	4	0	1	7	264.0	131.3	0.095	0.049

Table A.7: Simulation results comparing the evaluation metrics of the proposed smoothing method as presented in the paper with the scenario where the smoothing weights are fit on a second independent sample of the covariate data. The values are means (or standard deviations) over 10,000 iterations of the simulation study run with the baseline parameters.

method	σ_ε	n	ϕ	θ	$\sqrt{\psi}$	γ	FDP	s.d.	FDX	s.e.
base	1	50	4	0	1	7	0.083	0.127	0.295	0.046
smooth	1	50	4	0	1	7	0.080	0.137	0.274	0.045
base	2	50	4	0	1	7	0.051	0.163	0.120	0.032
smooth	2	50	4	0	1	7	0.055	0.176	0.117	0.032
base	3	50	4	0	1	7	0.044	0.187	0.064	0.024
smooth	3	50	4	0	1	7	0.046	0.190	0.065	0.025

Table A.8: Results for controlling false discoveries when the noise level simulation study is performed (with the same seed as the original simulation study with these simulation parameters), but with only locations where $\gamma_j > 0.1$ marked as true discoveries. The FDX gives the empirical probability estimated across the 10,000 simulations. The FDP gives the mean value across the 10,000 simulations. The s.d. column reports the sample standard deviation of the FDP values across the simulations, the s.e. column reports the estimated standard error of the FDX probability.

method	σ_ε	n	ϕ	θ	$\sqrt{\psi}$	γ	true rej.	s.d.	sens.	s.d.	% Δ sens.
base	1	50	4	0	1	7	199.3	93.5	0.237	0.110	-
smooth	1	50	4	0	1	7	249.0	115.9	0.296	0.135	24.669
base	2	50	4	0	1	7	42.9	73.7	0.051	0.088	-
smooth	2	50	4	0	1	7	57.4	97.2	0.068	0.116	33.348
base	3	50	4	0	1	7	10.2	37.8	0.012	0.045	-
smooth	3	50	4	0	1	7	15.1	53.2	0.018	0.063	47.390

Table A.9: Results for making true rejections when the same simulation study is performed (with the same seed as the original simulation study with these simulation parameters), but with only locations where $\gamma_j > 0.1$ marked as true discoveries. True rejections and sensitivity give the mean values of each quantity across the 10,000 simulations. The sample standard deviations across the simulations are listed after each metric. The percent change in sensitivity of the smooth method from the base method is listed in the column following the s.d. of sensitivity.

Appendix B

Supplemental Materials Chapter 3

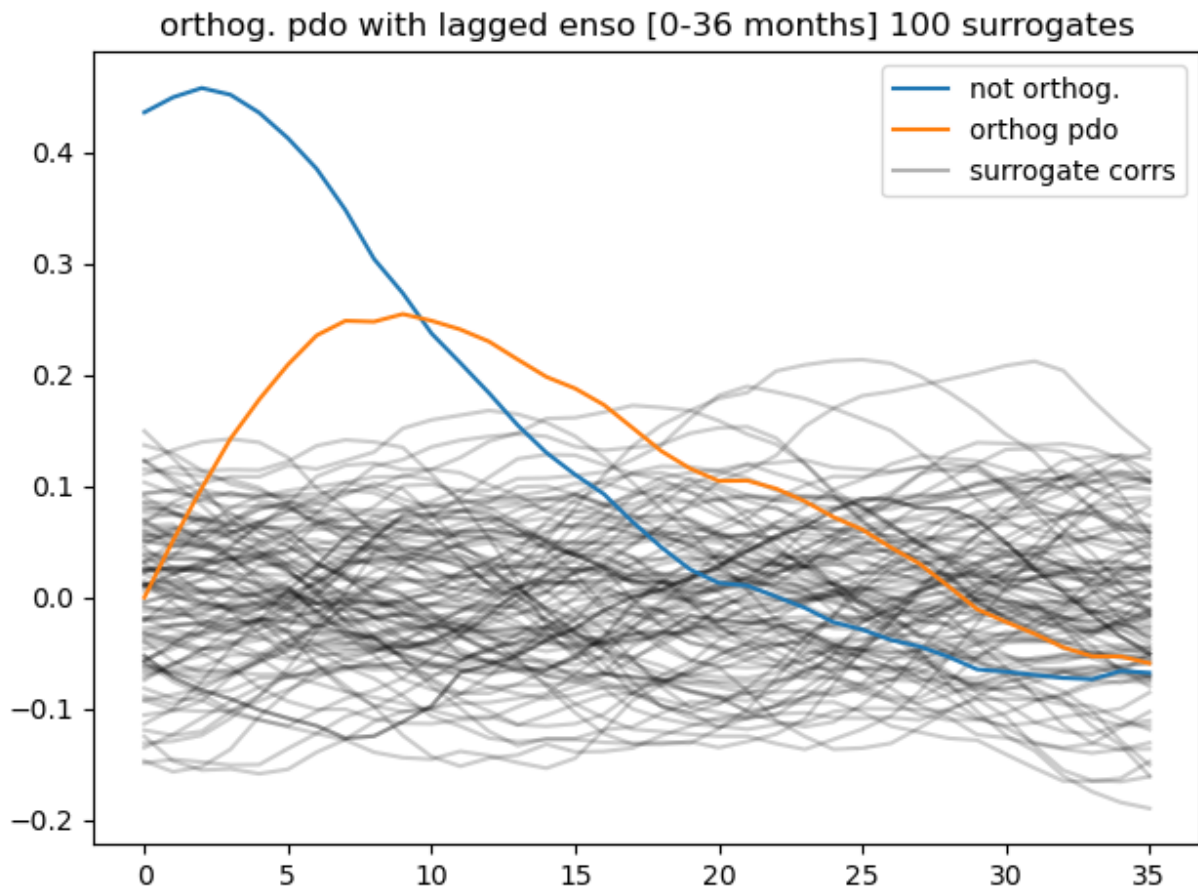


Figure B.1: Lagged correlations between ENSO and PDO Surrogate Modes generated with the univariate IAAFT method compared to lagged correlations found in the observed modes and in the orthogonalized observed modes. ENSO is lagged relative to the PDO time series.

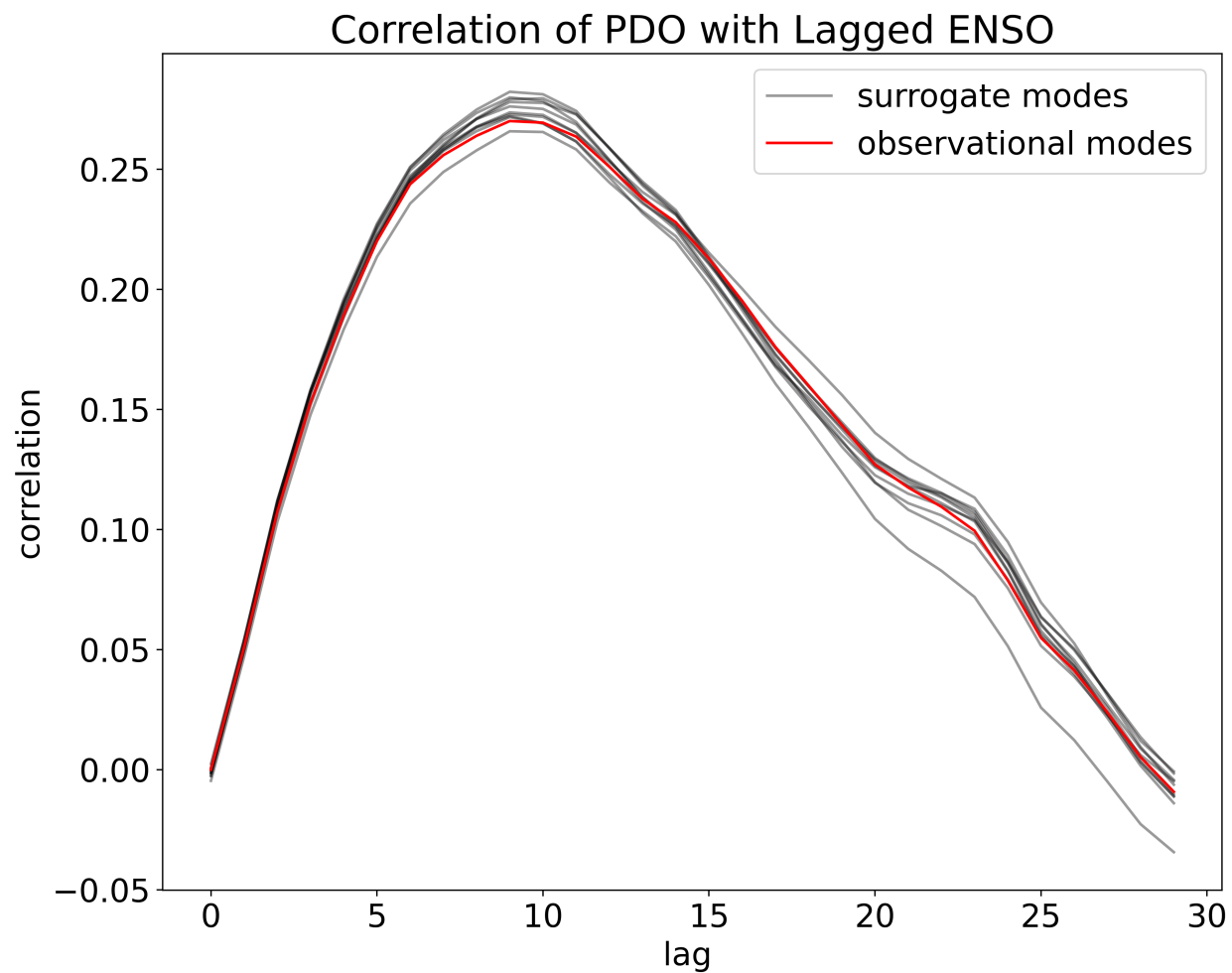


Figure B.2: Lagged correlations between ENSO and PDO Surrogate Modes generated with the multivariate IAAFT method compared to lagged correlations found in the orthogonalized modes. ENSO is lagged relative to the PDO time series.

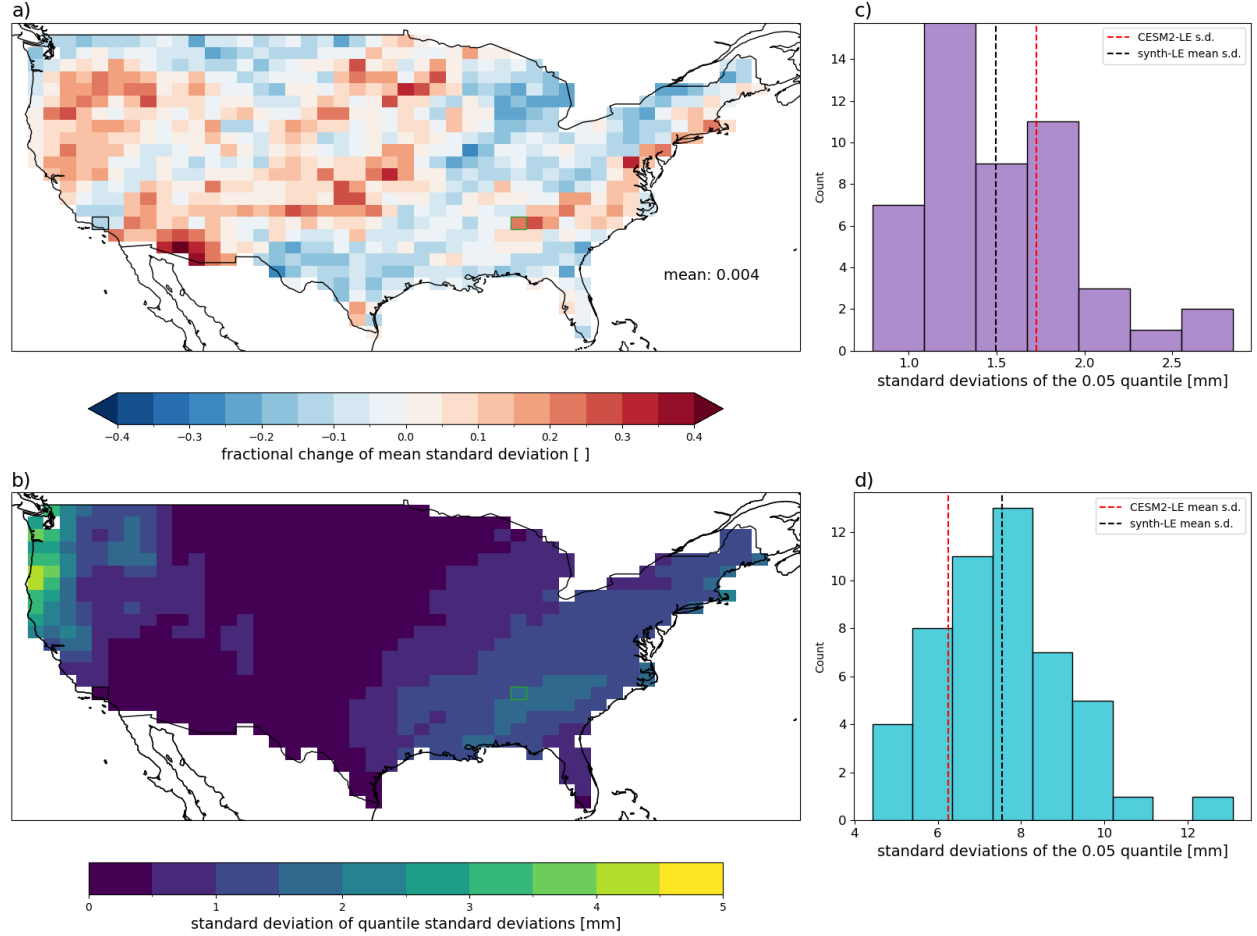


Figure B.3: a) The fractional change in the mean standard deviation of the 0.05 quantile of December precipitation across the Synth-LEs generated from each CESM2-LE member from the standard deviation of 0.05 quantile estimates of December precipitation across the CESM2-LE member, i.e. $(\bar{s}_{\text{Synth}} - s_{\text{CESM2}})/s_{\text{CESM2}}$. b) The standard deviation of the 50 estimates of the standard deviation of the 0.05 quantile from the 50 Synth-LEs. c) Histogram of the standard deviation estimates of the 0.95 quantile across the 50 Synth-LEs at the gridbox surrounded by the black rectangle in a) and b). d) Histogram of the standard deviation estimates of the 0.05 quantile across the 50 Synth-LEs at the gridbox surrounded by the green rectangle in a) and b). In c) and d) CESM2-LE standard deviation is plotted as a red dashed line, while the mean value across the 50 Synth-LE estimates is plotted as a black dashed line. The bin widths of the two histograms are different.

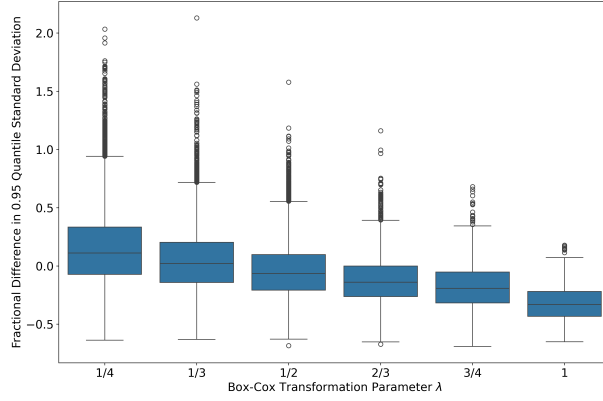


Figure B.4: Boxplots comparing the fractional difference in December 0.95 quantile standard deviations over the Synth-LEs generated from each ensemble member and the standard deviation over the CESM2-LE, categorized by the optimized λ value for each CESM2-LE member. The boxplots include values over all CONUS locations. Maps of the mean fractional difference values were plotted in Figure 4a)

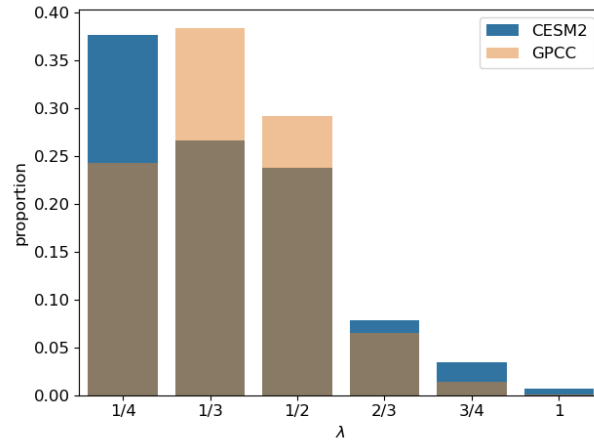


Figure B.5: Comparison of the proportion of each optimized λ value selected over CESM2-LE and GPCC data for December months.

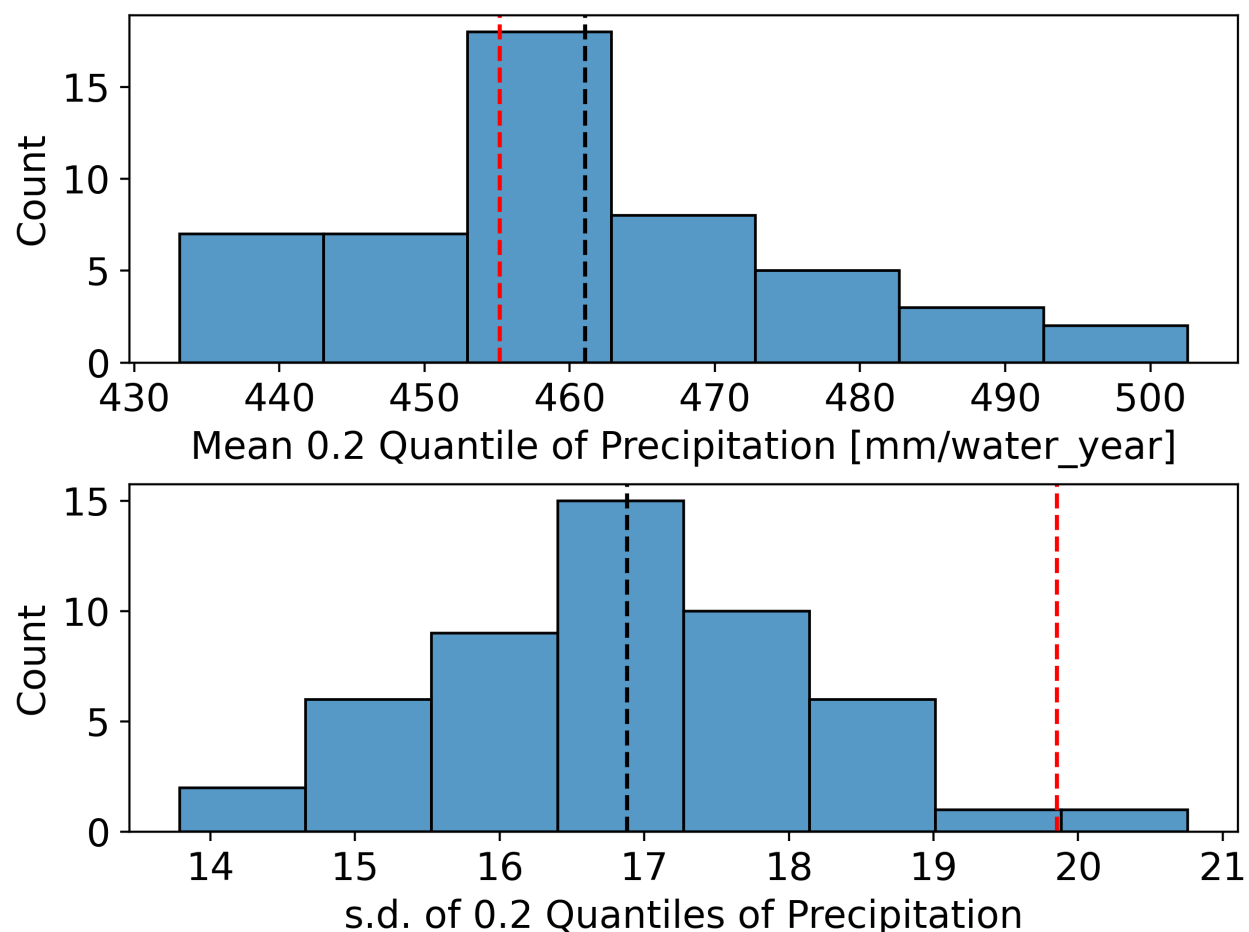


Figure B.6: a) Histograms of mean 0.2 quantile estimates across the Synth-LEs generated from each CESM2-LE member for California water year (September to October) mean precipitation. b) Histograms of the standard deviation of 0.2 quantile estimates across the Synth-LEs generated from each CESM2-LE member for California water year (September to October) mean precipitation. Mean values for each distribution are plotted as the black dashed line, while the CESM2-LE mean and standard deviation are plotted as the red dashed line.

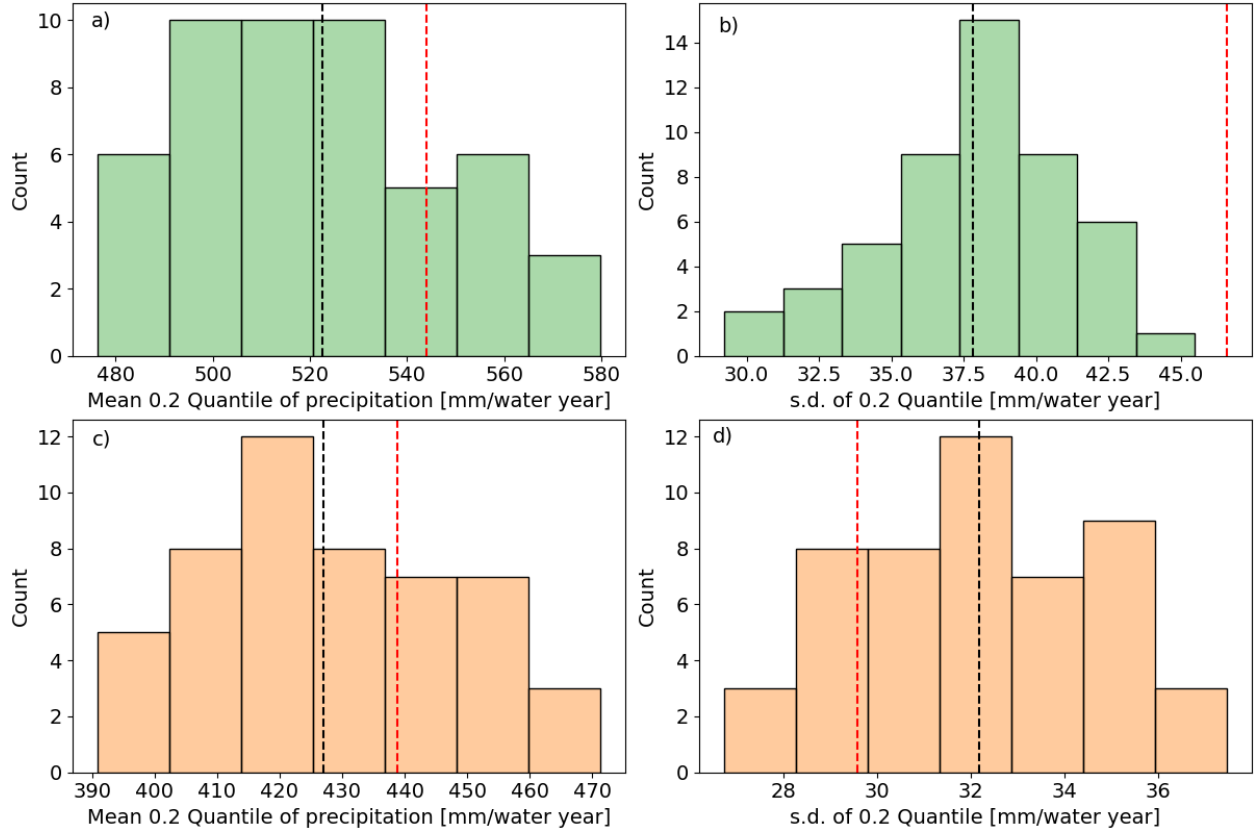


Figure B.7: a) Histograms of mean 0.2 quantile estimates across the Synth-LEs generated from each CESM2-LE member for California water year (September to October) mean precipitation during water years with positive ENSO. b) Histograms of the standard deviation of 0.2 quantile estimates across the Synth-LEs generated from each CESM2-LE member for California water year (September to October) mean precipitation during water years with positive ENSO. c) and d) as in a) and b), but for negative ENSO. Positive and Negative ENSO are defined by averaging the ENSO index over the DJF season, and selecting corresponding water years to include based on whether the index was above 1 or below -1 . Mean values for each distribution are plotted as the black dashed line, while the CESM mean and standard deviation are plotted as the red dashed line.

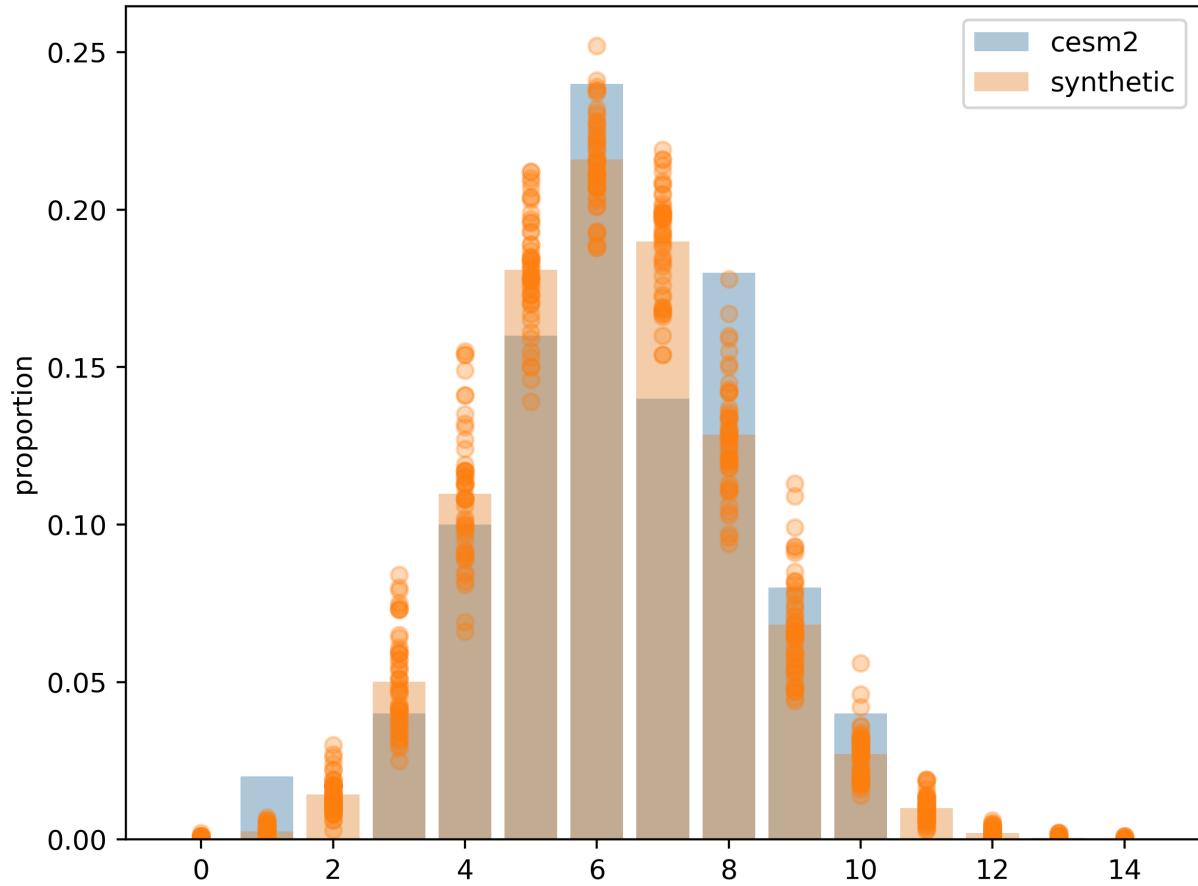


Figure B.8: The proportions of ensemble members with the specified number of extended dry stretches over the 100 year synthetic record. Extended dry stretches are defined as sequences of at least three consecutive years of below median California average precipitation. Sequences longer than three years are only counted once. The synthetic ensemble bars show the proportion across the 50 synthetic ensembles generated from each CESM2-LE member. The points show the proportions for each synthetic ensemble separately.

Appendix C

Supplemental Materials Chapter 4

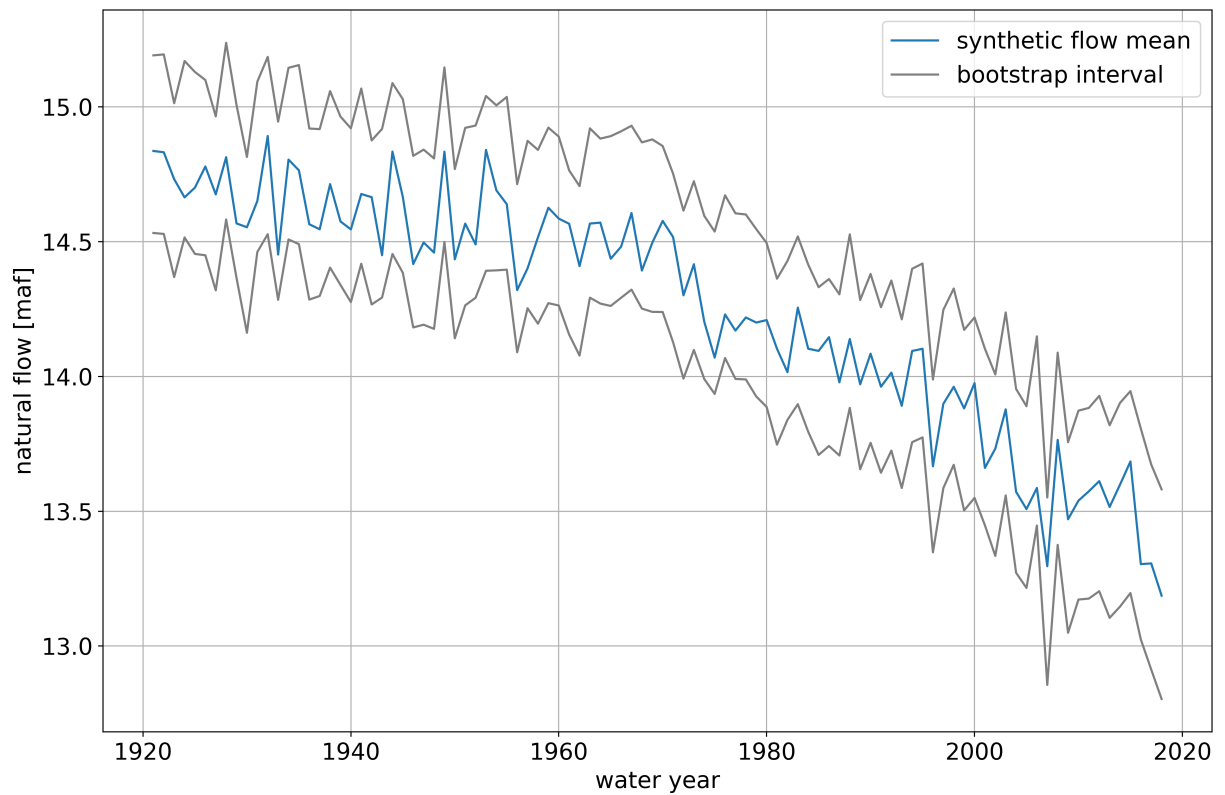


Figure C.1: Bootstrap 90% confidence interval for the mean of the synthetic flows, where the bootstrap step is applied while fitting Model (4.1).

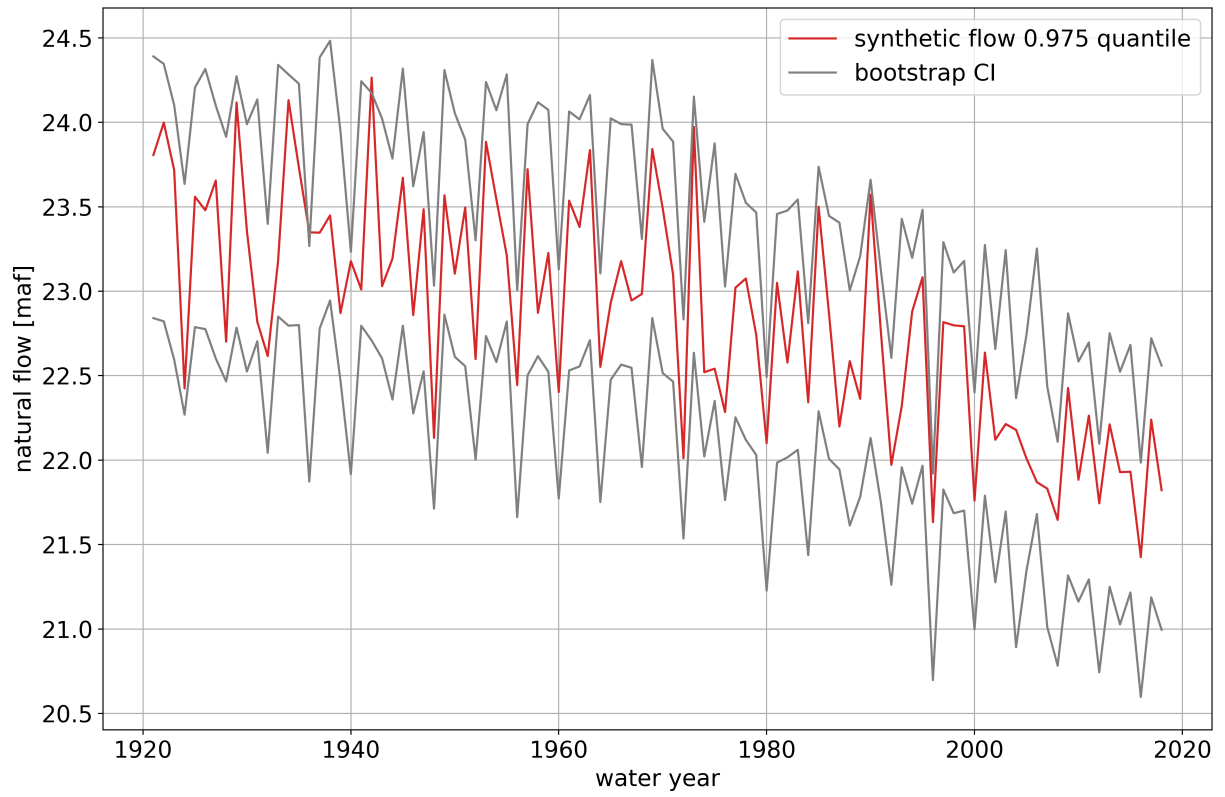


Figure C.2: Bootstrap 90% confidence interval for the 0.975 quantile of synthetic flows, where the bootstrap step is applied while fitting Model (4.1).

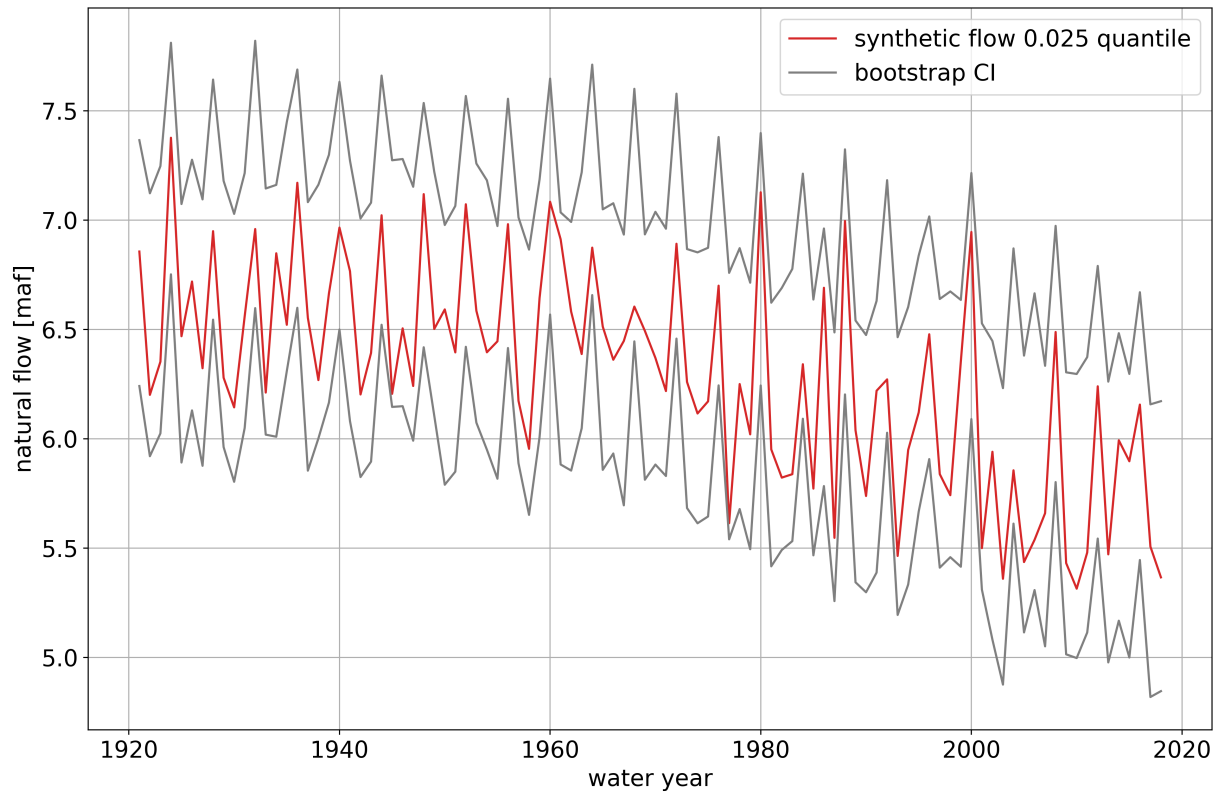


Figure C.3: Bootstrap 90% confidence interval for the 0.025 quantile of synthetic flows, where the bootstrap step is applied while fitting Model (4.1).

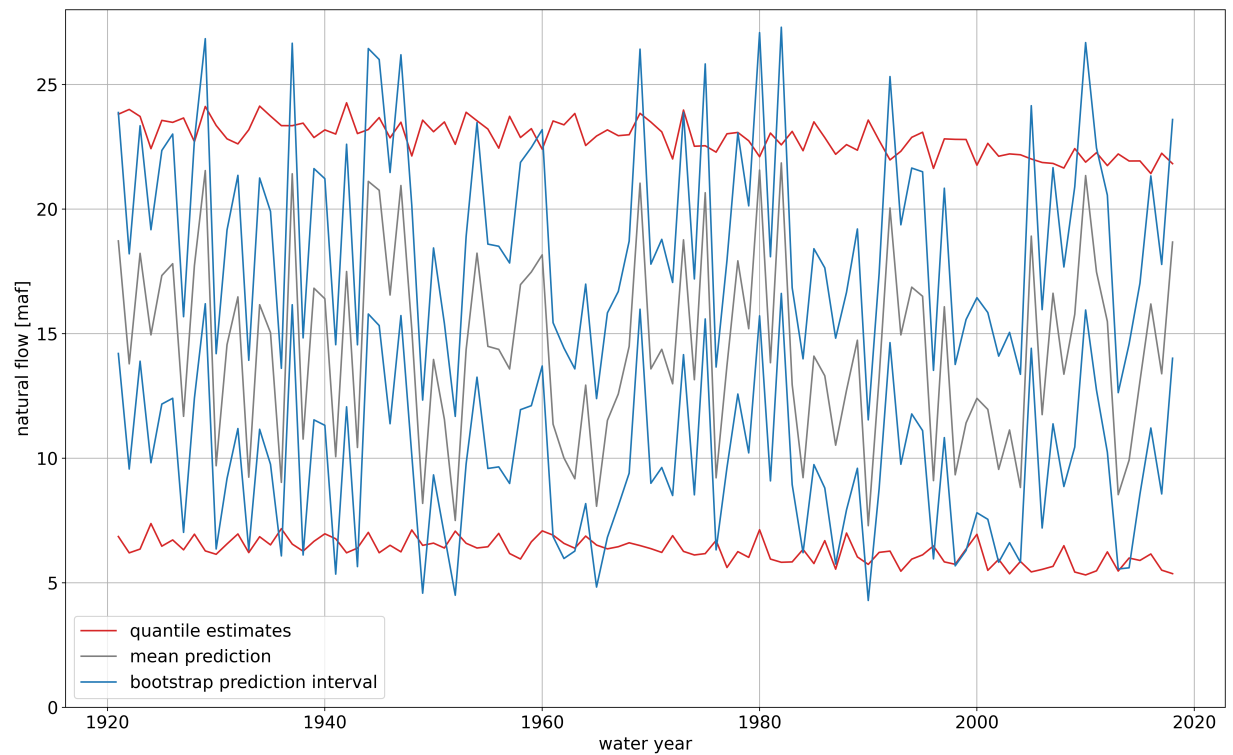


Figure C.4: The bootstrap 95% Prediction Interval for the synthetic flows conditioned on a single set of synthetic precipitation and time series records along with the mean prediction. The quantile estimates over the full distribution of synthetic precipitation and temperature is shown in red.

Bibliography

Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *Annals of statistics*, 43(5):2055–2085, 2015. doi:[10.1214/15-aos1337](https://doi.org/10.1214/15-aos1337).

Mathew Barlow, William J. Gutowski, John R. Gyakum, Richard W. Katz, Young-Kwon Lim, Russ S. Schumacher, Michael F. Wehner, Laurie Agel, Michael Bosilovich, Allison Collow, Alexander Gershunov, Richard Grotjahn, Ruby Leung, Shawn Milrad, and Seung-Ki Min. North American extreme precipitation events and related large-scale meteorological patterns: A review of statistical methods, dynamics, modeling, and trends. *Climate Dynamics*, 53(11):6835–6875, December 2019. doi:[10.1007/s00382-019-04958-z](https://doi.org/10.1007/s00382-019-04958-z).

Yoav Benjamini and Ruth Heller. False Discovery Rates for Spatial Signals. *Journal of the American Statistical Association*, 102(480):1272–1281, 2007. doi:[10.1198/016214507000000941](https://doi.org/10.1198/016214507000000941).

Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B, Methodological*, 57(1):289–300, 1995. doi:[10.1111/j.2517-6161.1995.tb02031.x](https://doi.org/10.1111/j.2517-6161.1995.tb02031.x).

Yoav Benjamini and Daniel Yekutieli. The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of statistics*, 29(4):1165–1188, 2001. doi:[10.1214/aos/1013699998](https://doi.org/10.1214/aos/1013699998).

G. E. P. Box and D. R. Cox. An Analysis of Transformations. *Journal of the Royal*

Statistical Society Series B: Statistical Methodology, 26(2):211–243, 1964. doi:[10.1111/j.2517-6161.1964.tb00553.x](https://doi.org/10.1111/j.2517-6161.1964.tb00553.x).

Jie Chen, François P. Brissette, and Robert Leconte. A daily stochastic weather generator for preserving low-frequency of climate variability. *Journal of Hydrology*, 388(3):480–490, July 2010. doi:[10.1016/j.jhydrol.2010.05.032](https://doi.org/10.1016/j.jhydrol.2010.05.032).

Cohen, Michael J. Municipal Deliveries of Colorado River Water. Technical report, Pacific Institute, 2011.

Christopher Daly, GH Taylor, and WP Gibson. The PRISM approach to mapping precipitation and temperature. In *Proc., 10th AMS Conf. on Applied Climatology*, volume 675, 1997.

G. Danabasoglu, J.-F. Lamarque, J. Bacmeister, D. A. Bailey, A. K. DuVivier, J. Edwards, L. K. Emmons, J. Fasullo, R. Garcia, A. Gettelman, C. Hannay, M. M. Holland, W. G. Large, P. H. Lauritzen, D. M. Lawrence, J. T. M. Lenaerts, K. Lindsay, W. H. Lipscomb, M. J. Mills, R. Neale, K. W. Oleson, B. Otto-Bliesner, A. S. Phillips, W. Sacks, S. Tilmes, L. van Kampenhout, M. Vertenstein, A. Bertini, J. Dennis, C. Deser, C. Fischer, B. Fox-Kemper, J. E. Kay, D. Kinnison, P. J. Kushner, V. E. Larson, M. C. Long, S. Mickelson, J. K. Moore, E. Nienhouse, L. Polvani, P. J. Rasch, and W. G. Strand. The Community Earth System Model Version 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, 12(2):e2019MS001916, 2020. doi:[10.1029/2019MS001916](https://doi.org/10.1029/2019MS001916).

A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Application*. Cambridge University Press, 1997. ISBN 978-0-521-57471-6.

C. Deser, F. Lehner, K. B. Rodgers, T. Ault, T. L. Delworth, P. N. DiNezio, A. Fiore, C. Frankignoul, J. C. Fyfe, D. E. Horton, J. E. Kay, R. Knutti, N. S. Lovenduski, J. Marotzke, K. A. McKinnon, S. Minobe, J. Randerson, J. A. Screen, I. R. Simpson, and M. Ting. Insights from Earth system model initial-condition large ensembles and future prospects. *Nature Climate Change*, 10(4):277–286, April 2020. doi:[10.1038/s41558-020-0731-2](https://doi.org/10.1038/s41558-020-0731-2).

- Clara Deser, Adam S. Phillips, and James W. Hurrell. Pacific Interdecadal Climate Variability: Linkages between the Tropics and the North Pacific during Boreal Winter since 1900. *Journal of Climate*, 17(16):3109–3124, August 2004. doi:[10.1175/1520-0442\(2004\)017<3109:PICVLB>2.0.CO;2](https://doi.org/10.1175/1520-0442(2004)017<3109:PICVLB>2.0.CO;2).
- Clara Deser, Isla R. Simpson, Adam S. Phillips, and Karen A. McKinnon. How Well Do We Know ENSO’s Climate Impacts over North America, and How Do We Evaluate Models Accordingly? *Journal of Climate*, 31(13):4991–5014, July 2018. doi:[10.1175/JCLI-D-17-0783.1](https://doi.org/10.1175/JCLI-D-17-0783.1).
- Peter Diggle. *Model-Based Geostatistics / Peter J. Diggle, Paulo J. Ribeiro, Jr.* Springer Series in Statistics. Springer, New York, 2007. ISBN 978-0-387-32907-9.
- Lu Dong, L. Ruby Leung, Fengfei Song, and Jian Lu. Roles of SST versus Internal Atmospheric Variability in Winter Extreme Precipitation Variability along the U.S. West Coast. *Journal of Climate*, 31(19):8039–8058, October 2018. doi:[10.1175/JCLI-D-18-0062.1](https://doi.org/10.1175/JCLI-D-18-0062.1).
- Alessio Farcomeni. More Powerful Control of the False Discovery Rate Under Dependence. *Statistical methods & applications*, 15(1):43–73, 2006. doi:[10.1007/s10260-006-0002-z](https://doi.org/10.1007/s10260-006-0002-z).
- Leela M. Frankcombe, Anna von der Heydt, and Henk A. Dijkstra. North Atlantic Multi-decadal Climate Variability: An Investigation of Dominant Time Scales and Processes. *Journal of Climate*, 23(13):3626–3638, July 2010. doi:[10.1175/2010JCLI3471.1](https://doi.org/10.1175/2010JCLI3471.1).
- Reinhard Furrer, Marc G Genton, and Douglas Nychka. Covariance Tapering for Interpolation of Large Spatial Datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523, September 2006. doi:[10.1198/106186006X132178](https://doi.org/10.1198/106186006X132178).
- R. García-Herrera, C. Wilkinson, F. B. Koek, M. R. Prieto, N. Calvo, and E. Hernández. Description and General Background to Ships’ Logbooks as a Source of Climatic Data. *Climatic Change*, 73(1):13–36, November 2005. doi:[10.1007/s10584-005-6954-4](https://doi.org/10.1007/s10584-005-6954-4).

- Alan E. Gelfand, Peter Diggle, Montse Fuentes, and Peter Guttorp. *Handbook of Spatial Statistics / Edited by Alan E. Gelfand, Peter J. Diggle, Montserrat Fuentes, Peter Guttorp*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, 2010. ISBN 978-1-4200-7287-7.
- Jose George and Athira P. A Multi-stage Stochastic Approach for Statistical Down-scaling of Rainfall. *Water Resources Management*, 37(14):5477–5492, November 2023. doi:[10.1007/s11269-023-03615-9](https://doi.org/10.1007/s11269-023-03615-9).
- J. J. Goeman, J. Hemerik, and A. Solari. Only closed testing procedures are admissible for controlling false discovery proportions. *Annals of Statistics*, 49(2):1218–1238, 2021. doi:[10.1214/20-AOS1999](https://doi.org/10.1214/20-AOS1999).
- Jelle J. Goeman and Aldo Solari. Multiple Testing for Exploratory Research. *Statistical science*, 26(4):584–597, 2011. doi:[10.1214/11-STS356](https://doi.org/10.1214/11-STS356).
- David F. Gold, Rohini S. Gupta, and Patrick M. Reed. Exploring the Spatially Compounding Multi-Sectoral Drought Vulnerabilities in Colorado’s West Slope River Basins. *Earth’s Future*, 12(11):e2024EF004841, 2024. doi:[10.1029/2024EF004841](https://doi.org/10.1029/2024EF004841).
- Timothy C. Haas. Kriging and automated variogram modeling within a moving window. *Atmospheric Environment. Part A. General Topics*, 24(7):1759–1769, January 1990. doi:[10.1016/0960-1686\(90\)90508-K](https://doi.org/10.1016/0960-1686(90)90508-K).
- H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, and J-N. Thépaut. ERA5 hourly data on single levels from 1940 to present, 2023.
- Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata

- Biavati, Jean Bidlot, Massimo Bonavita, Giovanna Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730): 1999–2049, 2020. doi:[10.1002/qj.3803](https://doi.org/10.1002/qj.3803).
- Kyle B. Heyblom, Hansi A. Singh, Philip J. Rasch, and Patricia DeRepentigny. Increased Variability of Biomass Burning Emissions in CMIP6 Amplifies Hydrologic Cycle in the CESM2 Large Ensemble. *Geophysical Research Letters*, 49(5):e2021GL096868, 2022. doi:[10.1029/2021GL096868](https://doi.org/10.1029/2021GL096868).
- David Higdon. A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environmental and Ecological Statistics*, 5(2):173–190, June 1998. doi:[10.1023/A:1009666805688](https://doi.org/10.1023/A:1009666805688).
- M. Hoerling, J. Barsugli, B. Livneh, J. Eischeid, X. Quan, and A. Badger. Causes for the Century-Long Decline in Colorado River Flow. *Journal of Climate*, 32(23):8181–8203, December 2019. doi:[10.1175/JCLI-D-19-0207.1](https://doi.org/10.1175/JCLI-D-19-0207.1).
- Martin P. Hoerling, Jon K. Eischeid, Henry F. Diaz, Balaji Rajagopalan, and Eric Kuhn. Critical Effects of Precipitation on Future Colorado River Flow. *Journal of Climate*, 37(16):4079–4093, July 2024. doi:[10.1175/JCLI-D-23-0617.1](https://doi.org/10.1175/JCLI-D-23-0617.1).
- Boyin Huang, Peter W. Thorne, Viva F. Banzon, Tim Boyer, Gennady Chepurin, Jay H. Lawrimore, Matthew J. Menne, Thomas M. Smith, Russell S. Vose, and Huai-Min Zhang. NOAA Extended Reconstructed Sea Surface Temperature (ERSST), Version 5 (years 1854-2023). NOAA National Centers for Environmental Information., 2017.
- Liwei Jia, Gabriel A. Vecchi, Xiaosong Yang, Richard G. Gudgel, Thomas L. Delworth,

- William F. Stern, Karen Paffendorf, Seth D. Underwood, and Fanrong Zeng. The Roles of Radiative Forcing, Sea Surface Temperatures, and Atmospheric and Land Initial Conditions in U.S. Summer Warming Episodes. *Journal of Climate*, 29(11):4121–4135, June 2016. doi:[10.1175/JCLI-D-15-0471.1](https://doi.org/10.1175/JCLI-D-15-0471.1).
- C. J. Keylock. A resampling method for generating synthetic hydrological time series with preservation of cross-correlative structure and higher-order properties. *Water Resources Research*, 48(12), 2012. doi:[10.1029/2012WR011923](https://doi.org/10.1029/2012WR011923).
- Marlene Kretschmer, Jakob Runge, and Dim Coumou. Early prediction of extreme stratospheric polar vortex states based on causal precursors. *Geophysical Research Letters*, 44(16):8592–8600, 2017. doi:[10.1002/2017GL074696](https://doi.org/10.1002/2017GL074696).
- Yan-Ning Kuo, Flavio Lehner, Isla R. Simpson, Clara Deser, Adam S. Phillips, Matthew Newman, Sang-Ik Shin, Spencer Wong, and Julie M. Arblaster. Recent southwestern US drought exacerbated by anthropogenic aerosols and tropical ocean warming. *Nature Geoscience*, 18(7):578–585, July 2025. doi:[10.1038/s41561-025-01728-x](https://doi.org/10.1038/s41561-025-01728-x).
- Mikael Kuusela and Michael L. Stein. Locally stationary spatio-temporal interpolation of Argo profiling float data. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2220):20180400, December 2018. doi:[10.1098/rspa.2018.0400](https://doi.org/10.1098/rspa.2018.0400).
- Daniel J. Leathers, Brent Yarnal, and Michael A. Palecki. The Pacific/North American Teleconnection Pattern and United States Climate. Part I: Regional Temperature and Precipitation Associations. *Journal of Climate*, 4(5):517–528, May 1991. doi:[10.1175/1520-0442\(1991\)004<0517:TPATPA>2.0.CO;2](https://doi.org/10.1175/1520-0442(1991)004<0517:TPATPA>2.0.CO;2).
- Flavio Lehner, Clara Deser, Isla R. Simpson, and Laurent Terray. Attributing the U.S. Southwest’s Recent Shift Into Drier Conditions. *Geophysical Research Letters*, 45(12): 6251–6261, 2018. doi:[10.1029/2018GL078312](https://doi.org/10.1029/2018GL078312).

- Flavio Lehner, Andrew W. Wood, Julie A. Vano, David M. Lawrence, Martyn P. Clark, and Justin S. Mankin. The potential to reduce uncertainty in regional runoff projections from climate models. *Nature Climate Change*, 9(12):926–933, December 2019. doi:[10.1038/s41558-019-0639-x](https://doi.org/10.1038/s41558-019-0639-x).
- Robert E. Livezey and W. Y. Chen. Statistical Field Significance and its Determination by Monte Carlo Techniques. *Monthly Weather Review*, 111(1):46–59, 1983. doi:[10.1175/1520-0493\(1983\)111<0046:sfsaid>2.0.co;2](https://doi.org/10.1175/1520-0493(1983)111<0046:sfsaid>2.0.co;2).
- Ben Livneh, Eric A. Rosenberg, Chiyu Lin, Bart Nijssen, Vimal Mishra, Kostas M. Andreadis, Edwin P. Maurer, and Dennis P. Lettenmaier. A Long-Term Hydrologically Based Dataset of Land Surface Fluxes and States for the Conterminous United States: Update and Extensions. *Journal of Climate*, 26(23):9384–9392, December 2013. doi:[10.1175/JCLI-D-12-00508.1](https://doi.org/10.1175/JCLI-D-12-00508.1).
- N. Maher, D. Matei, S. Milinski, and J. Marotzke. ENSO Change in Climate Projections: Forced Response or Internal Variability? *Geophysical Research Letters*, 45(20): 11,390–11,398, 2018. doi:[10.1029/2018GL079764](https://doi.org/10.1029/2018GL079764).
- Nicola Maher, Adam S. Phillips, Clara Deser, Robert C. Inglin Wills, Flavio Lehner, John Fasullo, Julie M. Caron, Lukas Brunner, Urs Beyerle, and Jemma Jeffree. The updated Multi-Model Large Ensemble Archive and the Climate Variability Diagnostics Package: New tools for the study of climate variability and change. *Geoscientific Model Development*, 18(18):6341–6365, September 2025. doi:[10.5194/gmd-18-6341-2025](https://doi.org/10.5194/gmd-18-6341-2025).
- Kyle R. McEvoy and Karen A. McKinnon. Increasing Power When Controlling Multiple Hypothesis Testing with Climate Data via Covariate Smoothing. *Journal of Agricultural, Biological and Environmental Statistics*, May 2026. doi:[10.1007/s13253-026-00738-5](https://doi.org/10.1007/s13253-026-00738-5).
- K. A. McKinnon, A. Rhines, M. P. Tingley, and P. Huybers. Long-lead predictions of eastern

- United States hot days from Pacific sea surface temperatures. *Nature Geoscience*, 9(5): 389–394, May 2016. doi:[10.1038/ngeo2687](https://doi.org/10.1038/ngeo2687).
- Karen A. McKinnon and Clara Deser. Internal Variability and Regional Climate Trends in an Observational Large Ensemble. *Journal of Climate*, 31(17):6783–6802, September 2018. doi:[10.1175/JCLI-D-17-0901.1](https://doi.org/10.1175/JCLI-D-17-0901.1).
- Karen A. McKinnon and Clara Deser. 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(24):9605–9622, December 2021. doi:[10.1175/JCLI-D-21-0251.1](https://doi.org/10.1175/JCLI-D-21-0251.1).
- Karen A. McKinnon, Andrew Poppick, Etienne Dunn-Sigouin, and Clara Deser. An “Observational Large Ensemble” to Compare Observed and Modeled Temperature Trend Uncertainty due to Internal Variability. *Journal of Climate*, 30(19):7585–7598, October 2017. doi:[10.1175/JCLI-D-16-0905.1](https://doi.org/10.1175/JCLI-D-16-0905.1).
- Nasser Najibi, Alejandro J. Perez, Wyatt Arnold, Andrew Schwarz, Romain Maendly, and Scott Steinschneider. A statewide, weather-regime based stochastic weather generator for process-based bottom-up climate risk assessments in California – Part I: Model evaluation. *Climate Services*, 34:100489, April 2024. doi:[10.1016/j.cliser.2024.100489](https://doi.org/10.1016/j.cliser.2024.100489).
- Jerome Namias. Spring and Summer 1988 Drought over the Contiguous United States—Causes and Prediction. *Journal of Climate*, 4(1):54–65, January 1991. doi:[10.1175/1520-0442\(1991\)004<0054:SASDOT>2.0.CO;2](https://doi.org/10.1175/1520-0442(1991)004<0054:SASDOT>2.0.CO;2).
- Albert Ossó, Philip Craig, and Richard P. Allan. An assessment of CMIP6 climate signals and biases in temperature, precipitation and soil moisture over Europe. *International Journal of Climatology*, 43(12):5698–5719, 2023. doi:[10.1002/joc.8169](https://doi.org/10.1002/joc.8169).
- Angeline G. Pendergrass, Reto Knutti, Flavio Lehner, Clara Deser, and Benjamin M. Sander-

- son. Precipitation variability increases in a warmer climate. *Scientific Reports*, 7(1):17966, December 2017. doi:[10.1038/s41598-017-17966-y](https://doi.org/10.1038/s41598-017-17966-y).
- M. Perone Pacifico, C. Genovese, I. Verdinelli, and L. Wasserman. False Discovery Control for Random Fields. *Journal of the American Statistical Association*, 99(468):1002–1014, 2004. doi:[10.1198/0162145000001655](https://doi.org/10.1198/0162145000001655).
- Geeta G. Persad, Daniel L. Swain, Claire Kouba, and J. Pablo Ortiz-Partida. Inter-model agreement on projected shifts in California hydroclimate characteristics critical to water management. *Climatic Change*, 162(3):1493–1513, October 2020. doi:[10.1007/s10584-020-02882-4](https://doi.org/10.1007/s10584-020-02882-4).
- David W. Pierce, Lu Su, Daniel R. Cayan, Mark D. Risser, Ben Livneh, and Dennis P. Lettenmaier. An Extreme-Preserving Long-Term Gridded Daily Precipitation Dataset for the Conterminous United States. *Journal of Hydrometeorology*, 22(7):1883–1895, July 2021. doi:[10.1175/JHM-D-20-0212.1](https://doi.org/10.1175/JHM-D-20-0212.1).
- David W. Pierce, Daniel R. Cayan, Jordan Goodrich, Tapash Das, and Armin Munévar. Evaluating Global Climate Models for Hydrological Studies of the Upper Colorado River Basin. *JAWRA Journal of the American Water Resources Association*, 58(5):709–734, 2022. doi:[10.1111/1752-1688.12974](https://doi.org/10.1111/1752-1688.12974).
- Paul Poli, Hans Hersbach, Dick P. Dee, Paul Berrisford, Adrian J. Simmons, Frédéric Vitart, Patrick Laloyaux, David G. H. Tan, Carole Peubey, Jean-Noël Thépaut, Yannick Trémolet, Elías V. Hólm, Massimo Bonavita, Lars Isaksen, and Michael Fisher. ERA-20C: An Atmospheric Reanalysis of the Twentieth Century. *Journal of Climate*, 29(11):4083–4097, June 2016. doi:[10.1175/JCLI-D-15-0556.1](https://doi.org/10.1175/JCLI-D-15-0556.1).
- Brian D. Richter, Dominique Bartak, Peter Caldwell, Kyle Frankel Davis, Peter Debaere, Arjen Y. Hoekstra, Tianshu Li, Landon Marston, Ryan McManamay, Mesfin M. Mekonnen, Benjamin L. Ruddell, Richard R. Rushforth, and Tara J. Troy. Water scarcity and fish

imperilment driven by beef production. *Nature Sustainability*, 3(4):319–328, April 2020. doi:[10.1038/s41893-020-0483-z](https://doi.org/10.1038/s41893-020-0483-z).

Mark D. Risser, Michael F. Wehner, John P. O’Brien, Christina M. Patricola, Travis A. O’Brien, William D. Collins, Christopher J. Paciorek, and Huanping Huang. Quantifying the influence of natural climate variability on in situ measurements of seasonal total and extreme daily precipitation. *Climate Dynamics*, 56(9):3205–3230, May 2021. doi:[10.1007/s00382-021-05638-7](https://doi.org/10.1007/s00382-021-05638-7).

Juan A. Rivera and Gabriel Arnould. Evaluation of the ability of CMIP6 models to simulate precipitation over Southwestern South America: Climatic features and long-term trends (1901–2014). *Atmospheric Research*, 241:104953, September 2020. doi:[10.1016/j.atmosres.2020.104953](https://doi.org/10.1016/j.atmosres.2020.104953).

Keith B. Rodgers, Sun-Seon Lee, Nan Rosenbloom, Axel Timmermann, Gokhan Danabasoglu, Clara Deser, Jim Edwards, Ji-Eun Kim, Isla R. Simpson, Karl Stein, Malte F. Stuecker, Ryohei Yamaguchi, Tamás Bódai, Eui-Seok Chung, Lei Huang, Who M. Kim, Jean-François Lamarque, Danica L. Lombardozzi, William R. Wieder, and Stephen G. Yeager. Ubiquity of human-induced changes in climate variability. *Earth System Dynamics*, 12(4):1393–1411, December 2021. doi:[10.5194/esd-12-1393-2021](https://doi.org/10.5194/esd-12-1393-2021).

Udo Schneider, Peter Finger, Anja Meyer-Christoffer, Elke Rustemeier, Markus Ziese, and Andreas Becker. Evaluating the Hydrological Cycle over Land Using the Newly-Corrected Precipitation Climatology from the Global Precipitation Climatology Centre (GPCC). *Atmosphere*, 8(3):52, March 2017. doi:[10.3390/atmos8030052](https://doi.org/10.3390/atmos8030052).

Thomas Schreiber and Andreas Schmitz. Improved Surrogate Data for Nonlinearity Tests. *Physical Review Letters*, 77(4):635–638, July 1996. doi:[10.1103/PhysRevLett.77.635](https://doi.org/10.1103/PhysRevLett.77.635).

Thomas Schreiber and Andreas Schmitz. Surrogate time series. *Physica D: Nonlinear Phenomena*, 142(3):346–382, August 2000. doi:[10.1016/S0167-2789\(00\)00043-9](https://doi.org/10.1016/S0167-2789(00)00043-9).

- Richard Seager, Mingfang Ting, Patrick Alexander, Haibo Liu, Jennifer Nakamura, Cuihua Li, and Matthew Newman. Ocean-forcing of cool season precipitation drives ongoing and future decadal drought in southwestern North America. *npj Climate and Atmospheric Science*, 6(1):141, September 2023. doi:[10.1038/s41612-023-00461-9](https://doi.org/10.1038/s41612-023-00461-9).
- Ridwana Binte Sharif, Viviana Maggioni, and Ishrat Jahan Dollan. Changes in historical and future precipitation patterns across the contiguous United States. *Frontiers in Earth Science*, 13, April 2025. doi:[10.3389/feart.2025.1542536](https://doi.org/10.3389/feart.2025.1542536).
- Sebastian Sippel, Elizabeth C. Kent, Nicolai Meinshausen, Duo Chan, Christopher Kadow, Raphael Neukom, Erich M. Fischer, Vincent Humphrey, Robert Rohde, Iris de Vries, and Reto Knutti. Early-twentieth-century cold bias in ocean surface temperature observations. *Nature*, 635(8039):618–624, November 2024. doi:[10.1038/s41586-024-08230-1](https://doi.org/10.1038/s41586-024-08230-1).
- Luke Stone, Courtenay Strong, Husile Bai, Thomas Reichler, Greg McCabe, and Paul D. Brooks. Atlantic-Pacific influence on western U.S. hydroclimate and water resources. *npj Climate and Atmospheric Science*, 6(1):1–8, September 2023. doi:[10.1038/s41612-023-00471-7](https://doi.org/10.1038/s41612-023-00471-7).
- Wenguang Sun, Brian J. Reich, T. Tony Cai, Michele Guindani, and Armin Schwartzman. False Discovery Control in Large-Scale Spatial Multiple Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(1):59–83, January 2015. doi:[10.1111/rssb.12064](https://doi.org/10.1111/rssb.12064).
- Wesley Tansey, Oluwasanmi Koyejo, Russell A. Poldrack, and James G. Scott. False Discovery Rate Smoothing. *Journal of the American Statistical Association*, 113(523):1156–1171, July 2018. doi:[10.1080/01621459.2017.1319838](https://doi.org/10.1080/01621459.2017.1319838).
- Mingfang Ting and Hui Wang. Summertime U.S. Precipitation Variability and Its Relation to Pacific Sea Surface Temperature. *Journal of Climate*, 10(8):1853–1873, August 1997. doi:[10.1175/1520-0442\(1997\)010<1853:SUSPVA>2.0.CO;2](https://doi.org/10.1175/1520-0442(1997)010<1853:SUSPVA>2.0.CO;2).

- Bradley Udall and Jonathan Overpeck. The twenty-first century Colorado River hot drought and implications for the future. *Water Resources Research*, 53(3):2404–2418, 2017. doi:[10.1002/2016WR019638](https://doi.org/10.1002/2016WR019638).
- United States Bureau of Reclamation. Drought Response Operations Agreement. <https://www.usbr.gov/ColoradoRiverBasin/dcp/droa.html>, April 2026a.
- United States Bureau of Reclamation. Post-2026 Operational Guidelines and Strategies for Lake Powell and Lake Mead – Draft Environmental Impact Statement (EIS). Technical report, United States Department of the Interior, September 2026b.
- Julie A. Vano, Tapash Das, and Dennis P. Lettenmaier. Hydrologic Sensitivities of Colorado River Runoff to Changes in Precipitation and Temperature. *Journal of Hydrometeorology*, 13(3):932–949, June 2012. doi:[10.1175/JHM-D-11-069.1](https://doi.org/10.1175/JHM-D-11-069.1).
- Valérie Ventura, Christopher J. Paciorek, and James S. Risbey. Controlling the Proportion of Falsely Rejected Hypotheses when Conducting Multiple Tests with Climatological Data. *Journal of Climate*, 17(22):4343–4356, November 2004. doi:[10.1175/3199.1](https://doi.org/10.1175/3199.1).
- Andrew Verdin, Balaji Rajagopalan, William Kleiber, and Richard W. Katz. Coupled stochastic weather generation using spatial and generalized linear models. *Stochastic Environmental Research and Risk Assessment*, 29(2):347–356, February 2015. doi:[10.1007/s00477-014-0911-6](https://doi.org/10.1007/s00477-014-0911-6).
- D. S. Wilks. Resampling Hypothesis Tests for Autocorrelated Fields. *Journal of Climate*, 10(1):65–82, January 1997. doi:[10.1175/1520-0442\(1997\)010<0065:RHTFAF>2.0.CO;2](https://doi.org/10.1175/1520-0442(1997)010<0065:RHTFAF>2.0.CO;2).
- D. S. Wilks. “The Stippling Shows Statistically Significant Grid Points”: How Research Results are Routinely Overstated and Overinterpreted, and What to Do about It. *Bulletin of the American Meteorological Society*, 97(12):2263–2273, December 2016. doi:[10.1175/BAMS-D-15-00267.1](https://doi.org/10.1175/BAMS-D-15-00267.1).

- D. S. Wilks and R. L. Wilby. The weather generation game: A review of stochastic weather models. *Progress in Physical Geography: Earth and Environment*, 23(3):329–357, September 1999. doi:[10.1177/030913339902300302](https://doi.org/10.1177/030913339902300302).
- Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data, Second Edition*. MIT Press, Cambridge, 2 edition, 2010. ISBN 978-0-262-23258-6.
- Daniel Yekutieli and Yoav Benjamini. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82(1):171–196, 1999. doi:[10.1016/S0378-3758\(99\)00041-5](https://doi.org/10.1016/S0378-3758(99)00041-5).
- Chunming Zhang, Jianqing Fan, and Tao Yu. Multiple testing via FDRL for large-scale imaging data. *Annals of statistics*, 39(1):613–642, 2011. doi:[10.1214/10-AOS848](https://doi.org/10.1214/10-AOS848).
- Xuebin Zhang, Jiafeng Wang, Francis W. Zwiers, and Pavel Ya Groisman. The Influence of Large-Scale Climate Variability on Winter Maximum Daily Precipitation over North America. *Journal of Climate*, 23(11):2902–2915, June 2010. doi:[10.1175/2010JCLI3249.1](https://doi.org/10.1175/2010JCLI3249.1).