Title
Changes in Hydrologic Extremes: Impacts of Nonstationarity on Water Resource Management

Author
Wang, Kimberly

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Changes in Hydrologic Extremes: Impacts of

Nonstationarity on Water Resource Management

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Civil Engineering

by

Kimberly Wang

2022

ABSTRACT OF THE DISSERTATION

Changes in Hydrologic Extremes: Impacts of

Nonstationarity on Water Resource Management

by

Kimberly Wang
Doctor of Philosophy in Civil Engineering
University of California, Los Angeles, 2022
Professor Dennis Lettenmaier, Chair

Climate change brings about several new hydrological changes including changes in intensity and frequency of extreme events such as drought and extreme precipitation. For a more accurate understanding of the impact of these extreme events on water resources, work is being done to move towards methods capable of accounting for nonstationarity. In this work, I examine how the presence of spatial nonstationarity and temporal nonstationarity can affect water resource management in two contexts. First, I consider drought in California linked to warming winter temperatures. The spatial nonstationarity present in California temperature records complicate attempts to create gridded datasets from station observations, and ability of the interpolation method to account for spatial nonstationarity greatly impacts the final products. I examine five gridded datasets to find substantial differences between winter temperature trend magnitudes and spatial patterns largely due to the interpolation method selected as well as to differences in station lists and bias correction methods. The differences in temperature also strongly impact modeled changes in Snow Water Equivalent (SWE) over the last century, which

are a vital component of water resources in California. Second, I examine extreme precipitation using Monte Carlo simulations representing events typical in the US to determine the bias and variance in error associated with using a temporally stationary versus temporally nonstationary model. In comparing the performance between a stationary and nonstationary version of the generalized extreme value distribution, I find the classic bias-variance tradeoff in modeling limits the usefulness of using a nonstationary distribution, which should be applied when record length is long, the coefficient of variation of the data is low, and the behavior of temporal nonstationarity is high (large increases in the mean and standard deviation of the data with time). This result complicates the reality of transferring from a stationary method of analysis to a nonstationary, revealing that for many stations in the US, record lengths are too short (fifty years and under) for the nonstationary application to be definitively useful.

The dissertation of Kimberly Wang is approved.

Mekonnen Gebremichael

Steven Adam Margulis

Karen McKinnon

Dennis P Lettenmaier, Committee Chair

University of California, Los Angeles

2022

dedicated to Ovik Banerjee

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am grateful for my committee members who have shown tremendous patience and kindness to me through the years. Thank you to Steve Margulis and Mekonnen Gebremichael for your time and your investment in my work as well as for all that you have taught me. Thank you to my advisors, Karen McKinnon and Dennis Lettenmaier, who were integral in shaping my experience as a graduate student.

I would like to thank the professors at UCLA who have met with me and taught me over the years: Jenny Jay, Don Kendall, and William Yeh. Thank you to Jesse and Mimi in CEE. Thank you to Kasi, Jenee, Vanessa, Rebecca, Brian, and Matt in the Geography department. I would like to thank my friends and family for their support.

VITA

EDUCATION

B.S.E.	Civil Engineering, Duke University	2013
M.S.E.	Civil Engineering, University of California, Los Angeles	2015
C. Phil	Civil Engineering, University of California, Los Angeles	2021

EXPERIENCE

Teaching Fellow	Geography, University of California, Los Angeles	2021 - 22

PUBLICATIONS

Published:

Wang, K. J., A. P. Williams, D. P. Lettenmaier, (2016), How much have California winters warmed over the last century?. Geophysical Research Letters, 44(17), 8893 – 8900.

Wanders N., A. Bachas, X.G. He, H. Huang, A. Koppa, Z.T. Mekonnen, B.R. Pagán, L.Q. Peng, N. Vergopolan, **K.J. Wang**, M. Xiao, S. Zhan, D.P Lettenmaier and E.F. Wood (2017), Forecasting the hydroclimatic signature of the 2015-16 El Niño event on the western U.S. Journal of Hydrometeorology, 18(1).

Submitted or in preparation:

Wang, K. J., K. McKinnon, D. P. Lettenmaier (2022), The bias-variance trade-off applied to nonstationary modeling of extreme precipitation evaluated using Monte Carlo simulations. Journal of Hydrology (in prep).

Schlef, K., K. E. Kunkel, C. Brown, Y. Demissie, D. P. Lettenmaier, A. Wagner, M. S. Wigmosta, T. R. Karl, D. R. Easterling, **K. J. Wang**, B. Francois, E. Yan (2022), Review and Best Practices for Incorporating Non-Stationarity from Climate Change into Intensity-Duration-Frequency (IDF) Curves. Journal of Water Resources Planning and Management (in review).

Chapter 1. Introduction

1.1 Background

Weather disasters are capable of disrupting and endangering human lives as well as causing significant economic loss and property damage. In 2021, the United States witnessed twenty natural disasters that caused over 1 billion USD in damages (NOAA, 2022) from droughts, flooding, severe storms, and other natural disasters. Due to global warming, some types of extreme weather events have been increasing in frequency and severity in the United States (Stott, 2016; National Academy of Sciences, 2016; Konisky et al., 2016). Increased periods of high temperatures lead to drought events while heavier extreme precipitation lead to increased flooding (Lall et. al, 2018). Events that used to be considered stationary are now being re-examined as this condition is being questioned (e.g. Jain and Lall, 2001; Milly et al., 2008; Villarini et al., 2009; Katz et al., 2010). The assumption of stationarity is deeply embedded in the statistical theoretical basis of traditional risk analysis and thus affects water management decision-making (Jacob, 2013). Being able to account for nonstationarity is critical to the future of water resource management. In this work, I examine how aspects of nonstationarity can impact water resource management decision making in the context of two hydrological extremes: meteorological drought and extreme precipitation events.

Meteorological drought, characterized by precipitation deficiency along with increases in evapotranspiration (Loon, 2015), is projected to increase in frequency and intensity under global warming with the United States Southwest and Rocky Mountain states predicted to experience the highest increases in drought frequency (Strzepek et al., 2010). The role of anthropogenically warming temperatures in drought is increasingly important (Williams et al., 2015), especially in

California where much of the state's water resources are naturally stored as snowpack in the mountains providing runoff in the spring (Kapnick and Hall, 2010). California saw a record-breaking low April 1st snow water equivalent (SWE) in 2015 after experiencing an also record-breaking warm winter (Belmecheri et al., 2015). Changes in temperature deeply affect snowpack through impacting how much precipitation falls as rain versus snow, thus affecting peak snow mass, and through altering when spring runoff begins, which affects streamflow (Regonda et al., 2005). It was found that during the historic drought in California that began in 2011/12 and ended in 2014/15, the long-term warming trend exacerbated the drought (Seager et al., 2015).

With the increasing influence of temperature on drought, accurate assessment of temperature is needed to model and plan management of water resources in the future. Direct measurements of temperature are often more scarce in mountainous regions (where most of the water in Western U.S. streams originates) due to the challenges of accessibility. The lack of consistent spatial coverage, particularly over complex terrain, makes hydrological models dependent on statistical interpolation wherever a continuous temperature dataset is needed. Due to the relationship between surface temperature, elevation, and landcover, the ability of the regression method selected to account for spatial nonstationarity can highly impact estimated surface temperatures (Su et al., 2012). Several gridded temperature datasets exist and cover California but discrepancies between interpolation methods and data selection and processing lead to differences amongst temperature products. Changes in instrumentation, observation time, and (in some cases) station location have a further effect on recorded temperature at-site, thus raising the question as to how much California winters have warmed. The ability to answer this question is critical for an accurate assessment of the mountain snowpack.

Extreme precipitation has been shown to exhibit nonstationary behavior with events increasing in frequency and intensity (e.g. Groisman et al., 2001; DeGaetano, 2009; Higgins and Kousky, 2013; Kunkel et al., 2013; Alfieri et al., 2015) and with changing spatiotemporal patterns (Jiang et al., 2016; Dhakal and Jain, 2019). Accordingly, there is much work being done now to incorporate nonstationarity into the frequency analyses so often used to assess extreme precipitation. A simple way to represent nonstationarity is to add a time-dependent parameter directly into the parameters of the probability distribution function (Katz, 2013). Cheng et al., 2014; Salas and Obeysekera, 2014; and Sarhadi and Soulis, 2017, introduce a linear trend in the location parameter or in the location and scale parameters of the probability density function chosen to represent extreme precipitation. There are also nonstationary probability distributions whose parameters vary with time, as in Leclerc and Ouarda, 2007; Begouria et al., 2011; Panagoulia et al.; 2013, and Um et al., 2017. Other developments include the use of a physical covariate rather than time (Agilan and Umamahesh, 2017) such as the El-Nino Southern Oscillation (e.g. Verdon, et al., 2005; Sun et al., 2017; Su et al., 2019). These nonstationary probability distributions have been used to update intensity-duration- frequency (IDF) curves which are extensively used in engineering practice. The implications of changing IDF curves include the need to update hydraulic infrastructure (Cheng and AghaKouchak, 2014; Yilmaz and Perera, 2014; Ganguli and Coulibaly, 2017; Vu and Mishra, 2019; Silva et al., 2021).

The move towards nonstationarity is not unanimous, however. Matalas, 2011, argues that nonstationarity cannot be accurately modeled as a trend. Montanari and Koutsoyiannis, 2014, argue that elements of hydrological frequency analysis are time invariant and that the use of nonstationary models without first identifying the deterministic relationship of changes in extreme precipitation is fundamentally problematic. Lins and Cohn, 2011, acknowledge the existence of nonstationarity in hydrologic time series but they, along with Serinaldi and Kilsby, 2015, believe there is too much uncertainty in the nonstationarity for it to be used practically. Much remains to be done before the use of nonstationary modeling, particularly based on at-site (Serinaldi and Kilsby, 2015) data, can be applied in practice. In particular, the increased variance associated with adding one or more parameter to a probability distribution function has not been thoroughly examined. The added variance is important given the inherent variability in estimated parameters even in the stationary case.

Fundamentally, at-site measurements, valued for their accuracy, is critical in assessing water resources. Regular observations of surface weather began in the early 1800s, with a large network of volunteer observers being organized by the Smithsonian in the 1850s (Fiebrich, 2009). Despite having a long history of implementation, however, most stations had very limited record lengths and large gaps of missing data. These limitations have long existed, and there have been numerous methods developed to fill in missing data (e.g. Wallis et al., 2007; Sattari et al., 2017; Sun et al., 2017; Tang et al., 2021). However, these methods were developed before the context of nonstationarity became so important and are often themselves based in stationarity. Further, stations represent a very small area and interpolation to create continuous or gridded products of data are affected by station density (Hofstra et al., 2009). The inclusion of nonstationarity only further complicates the methods used to address temporal gaps and spatial gaps. The lack of understanding of how to account for nonstationarity in at-site data can lead to poor decision making, which can be costly and even deadly. When water resources are over-estimated, droughts become even more difficult to endure and alternative water supplies are costly. When water resources are under-estimated, incorrectly sized infrastructure can lead to flooding and destruction. Thus, it is important to have an understanding as to how limitations of at-site data can impact water resource decision making. My work explores how nonstationarity can affect our estimation and therefore management of water resources in the face of both spatial and temporal nonstationarity.

1.2. Research Questions

In my work, I seek to further our understanding of how station data can best be used in the context of extreme event estimation in a nonstationary environment. The two questions I aim to answer are:

1. What contributes to the differences in temperature datasets in California, and what is the significance of these differences in the context of the 2011 – 2014 period of drought?

2. When does the application of a nonstationary model to extreme precipitation outperform the application of a stationary model when the assumption of stationarity is incorrect?

I address these questions in two core chapters. In Chapter 2, I study the differences between five gridded temperature datasets that cover California to discern how much of a difference exists between minimum and maximum winter temperatures. I look into differences in construction of the datasets to explain why the resulting products vary. In Chapter 3, I use Monte Carlo simulations to assess the error associated with modeling extreme precipitation with a nonstationary model versus a stationary model. I consider how this error changes with station characteristics such as the coefficient of variation (CV), skewness, and magnitude of increasing trend in the mean. I also consider how the record length affects this error. Chapter 4 integrates the implications of my work and possible future applications.

References

Agilan, V., N. V. Umamahesh (2017), What are the best covariates for developing non-stationary rainfall intensity-duration-frequency relationship? Advances in Water Resources, 101, 11 – 12.

Alfieri, L., P. Burek, L. Feyen, G. Forzieri (2015): Global warming increases the frequency of river floods in Europe. Hydrology and Earth System Sciences, 19(5), 2247 – 2260.

Begueria, S., M. Angulo-Martinez, S. M. Vicente-Serrano, J. I. Lopez-Moreno, A. El-Kenawy (2010), Assessing trends in extreme precipitation events intensity and magnitude using non-stationary peaks-over-threshold analysis: a case study in northeast Spain from 1930 to 2006. International Journal of Climatology, 31(14), 2102 – 2114.

Belmecheri, S. F. Babset, E. R. Wahl, D. W. Stahle, V. Touret (2015): Multi-century evaluation of Sierra Nevada snowpack. Nature Climate Change, 6(2).

Bracken, C., K. D. Holman, B. Rajagopalan, H. Moradkhani (2018), A Bayesian hierarchical approach to multivariate nonstationary hydrologic frequency analysis. Water Resources Research, 54(1), 243 – 255.

Cheng, L., AghaKouchak, A., Gilleland, E. et al. (2014), Non-stationary extreme value analysis in a changing climate. Climatic Change 127, 353–369.

DeGaetano, A.T. (2009), Time-dependent changes in extreme-precipitation return-period amounts in the continental United States. Journal of Applied Meteorology and Climatology, 48(10), 2086 – 2099.

Dhakal, N., S. Jain (2019), Nonstationary influence of the North Atlantic tropical cyclones on the spatio-temporal variability of the eastern United States precipitation extremes. International Journal of Climatology, 40(7), 3486 – 3499.

Fiebrich, C. A. (2009), History of surface weather observations in the United States. Earth-Science Reviews, 93(3-4), 77 – 84.

Ganguli, P., P. Coulibaly (2017), Does nonstationarity in rainfall require nonstationary intensity–duration–frequency curves? Hydrology and Earth System Sciences, 21(2).

Groisman, P.Y., Knight, R.W. and Karl, T.R. (2001), Heavy precipitation and high streamflow in the contiguous United States: trends in the twentieth century. Bulletin of the American Meteorological Society, 82(2), 219–246.

Higgins, R. W., and V. E. Kousky (2013), Changes in observed daily precipitation over the United States between 1950–79 and 1980–2009, J. Hydrometeorol., 14, 105–121.

Hofstra, N., M. New, C. McSweeney (2009), The influence of interpolation and station network density on the distributions and trends of climate variables in gridded daily data. Climate Dynamics, 35, 841 – 858.

Jacob, D. (2013), Nonstationarity in extremes and engineering design. Extremes in a Changing Climate [AghaKouchak et al. (eds.)], 363–417.

Jain, S., U. Lall (2001), Floods in a changing climate: does the past represent the future? Water Resources Research, 37(12), 3193 – 3205.

Jiang P., Z. Yu, M. R. Gautam, F. Yuan, K. Acharya (2016), Changes of storm properties in the United States: observations and multimodel ensemble projections. Global and Planetary Change 142(2016): 41–52.

Kapnick, S., A. Hall (2010), Observed Climate–Snowpack Relationships in California and their Implications for the Future. Journal of Climate, 23(13), 3446 – 3456.

Katz, R. W. (2010), Statistics of extremes in climate change. Climatic Change, 100 (1), 71 – 76.

Katz, R.W. (2013). Statistical Methods for Nonstationary Extremes. In: AghaKouchak, A., Easterling, D., Hsu, K., Schubert, S., Sorooshian, S. (eds) Extremes in a Changing Climate. Water Science and Technology Library, vol 65. Springer, Dordrecht.

Konisky, D. M., L. Hughes, C. H. Kaylor (2016), Extreme weather events and climate change concern. Climatic Change, 134(2016), 533 – 547.

Kunkel, K. E., T. R. Karl, D. R. Easterling, K. Redmond, J. Young, X. Yin, P. Hennon (2013), Probable maximum precipitation and climate change. Geophysical Research Letters, 40(7), 1402 – 1408.

Lall, U., T. Johnson, P. Colohan, A. Aghakouchak, C. Brown, G. McCabe, R. Pulwarty, and A. Sankarasubramanian (2018), Water. In Impacts, Risks, and Adaptation in the United States: Fourth National Climate Assessment, Volume II [Reidmiller, D.R., C.W. Avery, D.R. Easterling, K.E. Kunkel, K.L.M. Lewis, T.K. Maycock, and B.C. Stewart (eds.)]. U.S. Global Change Research Program, Washington, DC, USA, 145–173.

Leclerc, M., T. B. M. J. Ouarda (2007): Non-stationary regional flood frequency analysis at ungauged sites. Journal of Hydrology, 343(3-4), 254 – 265.

Libertino, A., P. Allamano, F. Laio, P. Claps (2018), Regional-scale analysis of extreme precipitation from short and fragmented records. Advances in Water Resources, 112, 147 – 159.

Loon, A. F. V. (2015), Hydrological drought explained. WIREs Water, 2, 359 – 392.

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., & Stouffer, R. J., 2008: Stationarity is dead: Whither water management? Science, 319(5863), pp. 573–574.

National Academy of Sciences Engineering, and Medicine, 2016; Attribution of Extreme Weather Events in the Context of Climate Change. Washington, DC: The National Academies Press.

NOAA National Centers for Environmental Information (NCEI), 2022: U.S. Billion-Dollar Weather and Climate Disasters, https://www.ncei.noaa.gov/access/monitoring/billions/.

Panagoulia, D., P. Economou, C. Caroni (2013), Stationary and nonstationary generalized extreme value modelling of extreme precipitation over a mountainous area under climate change. Environmetrics, 25(1), 29 – 43.

Regonda, S. K., B. Rajagopalan, M. Clark, J. Pitlick (2005), Seasonal Cycle Shifts in Hydroclimatology over the Western United States. Journal of Climate, 18(2), 372 – 384.

Salas, J. D., J. Obeysekera (2015), Revisiting the Concepts of Return Period and Risk for Nonstationary Hydrologic Extreme Events. Journal of Hydrologic Engineering, 19(3).

Sarhadi, A., E.D. Soulis (2017), Time-varying extreme precipitation intensity-duration-frequency curves in a changing climate. Journal of Geophysical Research, 44 (2017), 2454.

Sattari, M., A. Rezazadeh-Joudi, A. Kusiak (2017). Assessment of different methods for estimation of missing data in precipitation studies. Hydrology Research, 48(4), 1032 – 1044.

Seager, R., M. Hoerling, S. Schubert, H. Wang, et al. (2015), Causes of the 2011–14 California Drought. Journal of Climate, 28(18), 6997 – 7024.

Silva, D. F., S. P. Simonovic, A. Schardong, J. A. Goldenfum (2021), Assessment of non-stationary IDF curves under a changing climate: Case study of different climatic zones in Canada. Journal of Hydrology: Regional Studies, 36(100870).

Slott, P. 2016: How climate change affects extreme weather events. Science, 352(6239), pp. 1517-1518.

Strzepek, K, G. Yohe, J. Neumann, B. Boehlert (2010), Characterizing changes in drought risk for the United States from climate change. Environmental Research Letters, 5(4).

Su, C., X. Chen (2019), Covariates for nonstationary modeling of extreme precipitation in the Pearl River Basin, China. Atmospheric Research, 229(15), 224 – 239.

Sun, Q., C. Miao, Q. Duan, H. Ashouri, S. Sorooshian, K. Hsu (2017), A review of global precipitation data sets: data sources, estimation, and intercomparisons. Reviews of Geophysics, 56(1), 79 – 107.

Sun, Q., C. Miao, Y. Qiao, Q. Duan (2017), The nonstationary impact of local temperature changes and Enso on extreme precipitation at the global scale. Climate Dynamics, 49(11-12), 4281 – 4292.

Tang, G., M. P. Clark, S. M. Papalexiou (2021), The use of serially complete station data to improve the temporal continuity of gridded precipitation and temperature estimates. American Meteorological Society, 22(6), 1553 – 1568.

Um, M. Y. Kim, M. Markus, D. J. Wuebbles (2017), Modeling nonstationary extreme value distributions with nonlinear functions: An application using multiple precipitation projections for U.S. cities. Journal of Hydrology, 552, 396 – 406.

Verdon D.C., A.M. Wyatt, A.S. Kiem, S.W. Franks, S. Franks, T. Wagener, et al. (2005),

Multi-decadal variability of rainfall and streamflow across eastern Australia

Water Resources Research, 40(10), 333 – 341.

Villarini, G., J.A. Smith, F. Serinaldi, J. Bales, P.D. Bates, W.F. Krajewski (2009), Flood frequency analysis for nonstationary annual peak records in an urban drainage basin. Advances in Water Resources, 32(8), pp. 1255 – 1266.

Vu, T. M., A. K. Mishra (2019), Nonstationary frequency analysis of the recent extreme precipitation events in the United States. Journal of Hydrology, 575(2019), 999 – 1010.

Wallis, J. R. , M. G. Schaefer, B. L. Barker, G. H. Taylor (2007), Regional precipitation-frequency analysis and spatial mapping for 24-hour and 2-hour durations for Washington state. Hydrology and Earth Science Systems, 11(1), 415 – 442.

Williams, A. P., R. et al. (2015), Contribution of anthropogenic warming to California drought during 2012–2014. Geophysical Research Letters, 42(16), 6819 – 6828.

Chapter 2. How much have California winters warmed over the last century?

This chapter has been published in its current form in *Geophysical Research Letters* © American Geophysical Union with modifications for formatting. The supplemental material is attached as Appendix A.

Wang, K. J., A. P. Williams, D. P. Lettenmaier (2017), How much have California winters warmed over the last century? Geophysical Research Letters, 44(17), 8893 – 8900. https://doi.org/10.1002/2017GL075002

2.1 Introduction

High temperatures in California accompanied by severe drought from 2012 to 2016 captured national attention, with the National Climate Data Center reporting that the winters (Nov – Mar) of 2014-15 and 2013-14 were the warmest and second warmest winters on record, with the winter of 2014-15 over 1°C warmer than 2013-2014 [National Centers for Environmental Information, 2016]. The drought has been analyzed in context of historic climate variations [e.g. Mao et al., 2015, Williams et al., 2015] based on particular datasets. However, the evolution of temperature over the instrumental period of nearly 100 years has not been well documented. Here, we compare five gridded temperature datasets to assess differences between the products and to understand the nature of the differences.

Surface air temperature is a key climate variable and is intricately linked to many hydrological processes. It directly influences soil water storage (as a key driver of evapotranspiration, both directly as sensible heat and downward and emitted longwave radiation, and indirectly in the vapor pressure deficit and, through the daily temperature range, as downward solar radiation). Surface air temperature is used as input for nearly all hydrologic models and influences all surface water balance variables including evapotranspiration, runoff, SWE, soil moisture, and groundwater storage. Further, surface air temperature plays a central role in determining the partitioning of precipitation into rainfall versus snow. This partitioning is important in high elevations such as mountainous regions as it determines whether snowpack accumulates or decreases. California depends heavily on SWE for its water supplies [e.g. Mote et al., 2005, Cayan et al., 1993] and a better understanding of surface air temperature changes improves ability to plan and manage water resources.

Most analyses of long-term temperature trends and variability rely on gridded datasets, which process station observations in different ways. Among these datasets are UCLA'S California and Nevada Drought Monitoring System [Xiao et al. 2016], used by Mao et al. [2015] and termed SWM hereafter; the National Climatic Data Center's gridded National Temperature Index dataset [Vose et al. 2014], termed VOSE; the Precipitation Regression on Independent Slopes Method (PRISM) dataset [Di Luzio et al. 2008]; the Berkeley Earth Surface Temperatures (BEST) dataset [Rohde et al. 2013]; and an update of the gridded dataset described by Hamlet et al. [2005], termed HL here. The longest common period of record among all datasets begins in 1920 and is nearly a century long.

The variations among long-term gridded temperature datasets at both a local and state-wide scale have not received much attention. To understand the differences among datasets and their implications, we perform three analyses. First, we investigate the variations among datasets by estimating and comparing winter temperature trends for both the entire state of California and for the gridded daily maximum and minimum temperatures at the grid cell level. Second, we investigate the methods used to create each gridded dataset as well as the station records used. Third, we use a hydrologic model to show how variations among the temperature datasets impact model–reconstructed snowpack over California.

2.2 Data and Methods

We use the following gridded surface air temperature datasets: Berkeley Earth Surface Temperatures (BEST; Rohde et al. 2014) (1 degree), Hamlet and Lettenmaier (HL; Hamlet et al. 2005) ($1/16^{th}$ degree), Precipitation Regression on Independent Slopes Method (PRISM; Di Luzio et al. 2008) ($1/24^{th}$ degree), Drought Monitoring System (SWM; Xiao et al. 2016) ($1/16^{th}$ degree), VOSE (Vose et al. 2014) ($1/24^{th}$ degree). We also investigate the Historical Climatology Network (HCN) stations [Menne et al. 2009], which do not have a gridded format. We summarize each of the datasets briefly here and provide more details in the Supplementary Material (section S1).

The native resolution of BEST is 1 degree, but we re-gridded it to $1/24^{th}$ degree to match VOSE and PRISM. Because the BEST data are provided as anomalies, we calibrate the BEST anomalies to the shorter TopoWx dataset, which incorporates a novel homogenization of SNOTEL station data [Oyler et al. 2015]. VOSE heavily smooths station observations with a thin-plate smoothing spline procedure before performing a topographic adjustment based on PRISM mean fields. The PRISM gridding product uses a linear climate-elevation function that varies locally based on nearby station observations including snow pillow sites. The BEST dataset uses a kriging algorithm to interpolate from station data.

The SWM data relate current conditions to historic conditions; this requires relatively long climatologies, so the system uses stations with long records to the greatest extent possible but also requires that the stations report in real-time. Some aspects of the SWM data production are common to the Livneh et al. [2013] historical gridded dataset for the continental U.S. For instance, both use a constant lapse rate to account for elevation-related temperature differences.

The HCN is a subset of Cooperative Observer (Co-Op) stations with long records, 54 of which are in California. The station data have been subjected to data quality checks as described in Menne et al. [2009] and adjustments have been made where necessary for station moves and changes in instrumentation. While there is no gridded product for HCN, the extended HL dataset [Hamlet et al. 2005] uses HCN station observations to adjust its monthly Tmin and Tmax values and is used as a surrogate for a gridded HCN product. HL and SWM both use the SYMAP algorithm of Shepard [1984].

Where we analyzed station records, we only considered stations in California, even though nearby stations beyond the California border can affect the gridded temperature products. Subject to data availability limitations, we utilized station data to investigate trends on a station-by-station basis and to compare spatial patterns in the station trends with those inferred from the gridded data.

In the mid-1980s, many stations within the Co-Op network switched instrumentation from a liquid-in-glass thermometer within a Cotton Regional Shelter (LiG/CRS) to a thermistor-based system called a Maximum-Minimum Temperature System (MMTS). Quayle et al [1991] found that the change in instrumentation lead to an average 0.4 °C decrease in mean daily maximum temperature and an increase of 0.3 °C in mean daily minimum temperature. Both HCN and VOSE adjust for the instrumentation change while the others do not. We examined the change in winter temperatures for stations used in the SWM network based on documentation from Historical Observing Metadata Repository (HOMR) and found that about 60% of the stations within the network had clearly documented changes to MMTS. To study the effects of instrumentation changes on the SWM product, we removed the bias as estimated by Quayle et al. [1991] from the applicable station records within the SWM network and re-gridded using the same SWM gridding methodology previously applied to the unadjusted data.

In gridded datasets, there are no missing values by construct and each gridding method has a different procedure for filling in missing values. We aggregated daily or monthly temperatures to winter averages, where we define winter as November 1 through March 31, at the grid cell level and then use an area-weighted average to find the statewide winter average. Boundary grid cells may cross state boundaries and include small areas of the surrounding states so that the entire area of California is covered.

We used the Theil-Sen estimator [Sen, 1968] of temperature trends at a 95% confidence level. The Theil-Sen estimator is our primary method (over least squares linear regression; see S5) because it is less sensitive to outliers than parametric methods like linear regression, and has analytical estimates of confidence intervals that do not require distributional assumptions [Fernandes and Leblanc, 2005, Lavagnini et al. 2011]. We computed the trend in the aggregate statewide averages over the winter of 1920-21 to the winter of 2014-15 (2013-14 for HL). The results represent the average temperature trends. We used Theil-Sen estimators to examine temperature trends at the grid cell level. We also evaluated pattern correlations between the winter temperature trend spatial fields using the Pearson product-moment correlation coefficient (see section S6).

For our station analysis, we computed trends for long-term stations with minimal missing data such that daily datasets must have 90% of winter days reported and monthly datasets must report all five winter months. Further, 90% of winters must be reported for the winter of 1920/21 through 2014/15 (or to 2013/14 for HL). Missing values were not substituted. There is substantial overlap among stations from the different datasets that meet our length-of-record requirements, as most stations are from NOAA's Co-Op network of long-term climate stations. More details on the specific thresholds are provided in section S4.

To demonstrate the influence of different temperature trends on hydrologic variables, we implemented the Variable Infiltration Capacity (VIC) macroscale hydrologic model, version 4.0, [Liang et al., 1994] to reconstruct SWE from 1920 through 2015 (2014 for HL) for the entire state of California. The meteorological inputs required to run VIC include daily precipitation, wind speed, Tmax, and Tmin. We used the same wind speed and precipitation data for all runs and changed only temperature inputs. Wind speed and precipitation were taken from the SWM dataset. For BEST, PRISM, and VOSE, we disaggregated the monthly temperatures to daily by adjusting the SWM daily data such that the monthly averages matched. We resampled using a nearest neighbor technique so that the datasets at 1/24th degree matched the 1/16th degree resolution of SWM. In running VIC, we used ten years for spin up and then ran the model over the state for the period 1920 to 2015 (2014 for HL) enabling the full energy balance as described in Liang et al., [1999].

2.3 Results and Discussion

a. Statewide Domain

For all datasets, we constructed time series of statewide winter average Tmax and Tmin. Figure 1 shows the statewide time series for the five gridded datasets starting from the winter of 1920 to 2015 (2014 for HL). Tmin trend magnitudes vary across datasets but are all positive. The

SWM data has an offset (which appears to be a result of their using a different temperature lapse rate than the other datasets) but the overall trends are similar to the other datasets. All Tmin datasets have positive trends over the period of record, which range from 1.2 to 1.9 °C per century, with SWM reporting the smallest trend and PRISM reporting the largest.

There is more disagreement regarding trends in Tmax, which range from -0.3 to 1.2 °C per century. SWM is the only dataset with a negative trend, but in all cases, trends in Tmax are less positive than the corresponding trends in Tmin. We show all the statewide long-term trends defined as 1920/21 to 2014/15 (or 2013/14 for HL) and the trends over the shorter and typically warmer period 1970/71 to 2014/15 (or 2013/14) in Table 1. Trend magnitudes over the more recent period (beginning 1970/71) are always larger than the long-term trend for both maximum and minimum. Unlike the long-term trends, the minimum temperature trends for the most recent ~45 years are larger than the maximum trend for only PRISM and SWM.

Differences in trend magnitudes are likely due to both the differences in the stations used by each dataset as well as differences in the gridding methods used. Each dataset has different station requirements for choosing the stations and station data to incorporate into their gridded product; further, each dataset has its own procedures for adjusting station data and estimating missing values. This results in different station lists and densities as well as different station data being used as input to create each gridded dataset. The gridding resolution and interpolation method used also creates differences among datasets that can be reflected in the statewide temperature trend. Many papers have compared different gridding methods [e.g. Lam et al., 1983]. A key difference in the production of gridded datasets for regions with complex terrain is the representation of lapse rates. Constant lapse rates, as used in SWM and HL, fail to capture daily variations, while the locally-varying lapse rates employed by BEST, PRISM, and VOSE (see S1 for more detail) can result in anomalous behavior in elevational differences as a result of spatial variations in station density at high elevations [Stahl et al., 2006]. Higher lapse rates lead to colder temperatures at higher elevations and is partially why SWM has the lowest Tmin of all the datasets. HL also has a constant lapse rate, but other adjustments used in the gridding process lead to differences between SWM and HL.

b. Station-based Analysis

In addition to our analysis of gridded data, we analyzed trends on a station-by-station basis for stations with long-term records, defined in the methods. Based on these screens for missing values, BEST had 68 stations that met the record length requirements for Tmax and 62 for Tmin. VOSE and HL had the same 44 stations for both Tmax and Tmin as they both draw on the daily temperature products from the Global Historical Climatology Network (GHCN-D). PRISM had 46 stations for Tmax and 45 for Tmin. SWM had 16 stations for Tmax and 15 for Tmin. There is considerable overlap among the stations from the different datasets (see Figure S3).

Results of the Theil-Sen trend analyses for each of the stations identified using the above criteria are shown in Figure 2. The station trends generally are consistent with state-average temperatures and gridded product trends. Specifically, station Tmin series show more warming than do Tmax trends, which is also reflected in the larger magnitudes of Tmin as compared with Tmax trends at the statewide level. Further, the number of stations reporting warming trends are considerably greater than those reporting cooling trends for most of the datasets, and more stations have statistically significant trends for Tmin than for Tmax. For Tmax, the greatest number of statistically significant warming trends is in southern and central California, while many of the stations reporting cooling trends are in coastal regions and near mountainous regions. For Tmin, there are few stations reporting cooling trends, and for those stations that do have cooling trends, the magnitudes are smaller in absolute value than the warming trends.

Station density is lowest in all cases in northern California, in the Sierra Nevada range, and in the desert region of southeastern California. Station density affects the gridded temperature estimates and can introduce biases in local estimates from the gridded products since all methods of constructing a gridded product rely on some distance-temperature relationship.

c. Spatial Variations of Gridded Trend Magnitudes

We performed trend analyses on the winter temperature time series at each grid cell for each dataset, shown in Figure 3. The results vary widely among datasets, and in some instances the relationship between gridded products and long-term reporting stations is unclear.

The maps for BEST indicate high-resolution artifacts due to the calibration to the TopoWx dataset but nonetheless indicate relatively spatially homogeneous warming trends. BEST is the only dataset that shows only warming in both Tmin and Tmax across the entire state. The spatial patterns that might be expected based on station locations and their trends, some of which are negative, are not visible in the gridded trend map for BEST due to the low native resolution of this dataset. Of all the datasets, BEST has the smallest spatial standard deviation for both Tmax and Tmin trends at the grid cell level. Though the variation is slight, BEST Tmin trends are larger than Tmax trends.

The list of long-term stations used by PRISM is similar to that for VOSE, but the gridded trend maps for these two datasets are very different. PRISM shows more local variation in trends and hence has a wider range of trend magnitude. Like the VOSE and SWM gridded datasets, PRISM shows some cooling trends in maximum temperature along the coast. Other areas in the PRISM gridded data with cooling trends, such as southern central California, are not reflected in the long-term station trends; the PRISM datasets allow for the introduction of stations with shorter-term records compared to SWM and HL, increasing local-scale accuracy for portions of the record but apparently in a few cases resulting in artificial local trends. Nonetheless, general patterns of Tmin and Tmax trends for PRISM are consistent between station trends and grid cell trends in that Tmin trends are almost entirely positive and have larger magnitudes in absolute value than Tmax trends.

The VOSE Tmax trends show cooling in areas close to the stations with cooling trends (see Figure 2). In contrast, VOSE Tmin trends do not show the signatures of station trends so clearly. At the statewide scale, the station trends and VOSE gridded trends are similar, with larger warming trends in Tmin than Tmax and with Tmin having positive trends almost everywhere. Tmin trends have higher average trends than Tmax.

SWM has the largest range of trends for both Tmin and Tmax. SWM's maximum temperatures are slightly dominated by negative trends, particularly in Southern California, despite the fact that all but one of the statistically significant station trends were upward. The large areas for which cooling trends were inferred in Tmax is consistent with the aggregated SWM statewide trend being slightly negative. For Tmin, a few areas along the coast and the southwest border of California experienced the largest warming trends, while portions in the north and central parts of the state have cooling trends. The SWM dataset shows the largest cooling areas in both Tmax and Tmin of all datasets. The SWM Tmin trend decreased from 1.2 °C/century to 0.99 °C/century after removing the bias in stations that underwent instrumentation change; the SWM Tmax trend increased from -0.30 °C/century to -0.05 °C/century. We include details about the instrument change and bias in the Supplementary Material (section S4).

The HL dataset shows the least spatial variation after BEST (likely because of the relatively sparse HCN station network). Although SWM and HL both use the SYMAP gridding algorithm, there are differences between the station lists used and other processes in creating the dataset, such as HL's use of the HCN monthly product for corrections at the monthly time-step and use of the PRISM temperature data as bias correction for climatology values, leading to different trend maps. Also, HL and VOSE have the same long-term reporting stations from the GHCN-D network, yet the final gridded product of the two have different spatial patterns, particularly for Tmin. HL Tmax trends show slight to no warming in the San Francisco Bay area as well as some cooling in parts of the northern-central Sierra Nevada range, similar to VOSE. Unlike SWM, there are no dominant cooling trends in Southern California nor are there areas of intense warming trends. HL Tmin does show large positive trends in Southern California, but there are still small areas of cooling trends, the pattern of which is unique to the HL trend map. The cooling areas seen in HL are not shown in VOSE despite having the same long-term stations.

Pattern correlation coefficients reinforce the visual similarities and differences in the spatial patterns shown in Figure 3. The maps of similar texture (e.g., VOSE and HL) are the most highly correlated (while SWM has similar texture, the large cooling trends seen in Tmax result in slightly negative correlations with most other results). BEST and PRISM contrast the most in terms of smoothness and have near zero correlation for Tmax and a slightly negative correlation for Tmin (see Table S4). It is worth noting that while the trend maps have relatively low pattern correlations, the mean (Tmax and Tmin) fields generally are highly correlated (all pairs of monthly and annual mean fields have pattern correlations that exceed 0.9). Differences in gridding algorithms and station selection make far more difference to the spatial fields of trends than to the means.

For our applied example, we extracted April 1 SWE from the VIC model runs to find the statewide SWE volume for each dataset. SWM has the highest average SWE volume (23.6 km^3) and the lowest Tmin across the state. The other four datasets, BEST, HL, PRISM, and VOSE, have lower mean SWE values of 12.6, 13.9, 15.0, and 12.5 km^3 respectively. These snowpack estimates are lower due to overall higher Tmin values and particularly due to warmer temperatures at higher elevations. When we evaluate the trends in SWE across the state, we find that HL has the smallest decreasing trend (-5.04 km^3 per century) while PRISM has the largest (-7.64 km^3 per century). BEST, SWM, and VOSE have trends of -6.32, -7.07, and -5.36 km^3/year respectively. It is important to note that the SWM average SWE is roughly similar to other methods of SWE estimates across California (Margulis et al., 2016, estimate an average snowpack of around 20 km^3 for the Sierra Nevada range for 1985 to 2015). The resulting SWE simulations for BEST, HL, PRISM, and VOSE are much lower in comparison. This emphasizes how differences between temperature datasets affect secondary variables in analyses and that some datasets may be more appropriate than others for specific types of hydrologic analyses.

2.4 Conclusions

We evaluated century-scale winter temperature trends over California using five gridded datasets. Although the datasets use different stations, many stations are common to multiple networks. We find that:

- All datasets agree that long-term (century) trends in Tmin have been greater than in Tmax (as found in many previous studies). In all but one case, both Tmin and Tmax trends are increasing.
- Inferred trend magnitudes are greater in recent decades (1970s – present) than in the longer 100-year record.
- Generally, statewide average trends, station trends, and gridded trends in both Tmax and Tmin agree as to sign and approximate magnitudes; e.g., trends in Tmin are larger than in Tmax. Furthermore, spatial patterns in trends at long-term stations are similar across datasets.
- There is a lack of agreement in the spatial pattern of temperature trends in the gridded products with some datasets showing much greater spatial variability in trends across the state than others. This is related mostly to the degree of smoothness imposed by the different gridding methods, since trends in long-term stations used in the different datasets are similar. Differences in temperatures among datasets imply modeled snow pack trends that vary by a factor of one and a half (-5 to -7.6 km^3 per century).
- In correcting the SWM dataset for a network-wide instrumentation change, we found that the statewide Tmin trend decreased by 0.21 °C per century while the statewide Tmax trend increased by 0.25 °C per century. There may be a similar change in trend magnitudes for the other uncorrected datasets, BEST, HL, and PRISM.

Acknowledgments

The VOSE dataset is publicly available at ftp://ftp.ncdc.noaa.gov/pub/data/climgrid/. The PRISM data set is publicly available at http://prism.nacse.org/. The BEST data set is available at http://www.berkeleyearth.org/, and the Oyler et al. data set is available at https://github.com/jaredwo/topowx. The regridded BEST data set we use is available at http://www.hydro.ucla.edu/SurfaceWaterGroup/data.php. The SWM data set is available at http://uw-hydro.github.io/data/.

References

Casola, J.H., L. Cuo, B. Livneh, D.P. Lettenmaier, M. Stoelinga, P.W. Mote and J. M. Wallace (2009), Assessing the Impacts of Global Warming on Snowpack in the Washington Cascades, *Journal of Climate*, 22(10), 2758-2772.

Cayan, Daniel R., L. G. Riddle, E. Aguado (1993), The influence of precipitation and temperature on seasonal streamflow in California, *Water Resources Research*, 29(4), 1127 – 1140.

Cheng, Linyin, et al. (2016), How has human-induced climate change affected California drought risk?, *American Meteorological Society*, 29(1), 111-120.

Cooper, Matthew G., Anne W. Nolin, and Mohammed Safeeq (2016), Testing the recent snow drought as an analog for climate warming sensitivity of Cascades snowpacks, *Environmental Research Letters*, 11(8), doi: 10.1088/1748-9326/11/8/084009.

Di Luzio, M., G.L. Johnson, C. Daly, J.K. Eischeid, and J.G. Arnold (2008), Constructing restrospective gridded daily precipitation and temperature datasets for the conterminous United States, *J. Applied Meteorology and Climatology*, 47(2), 475-497.

Diffenbaugh, N. S., D. L. Swain, D. Touma (2015), Anthropogenic warming has increased drought risk in California, Proceedings of the National Academy of Sciences USA, 112(13), 3931-3936, doi:10.1073/pnas.1422385112.

Fernandes, R., S. G. Leblanc (2005), Parametric (modified least squares) and non-parametric (Theil-Sen) linear regressions for predicting biophysical parameters in the presence of measurement errors, *Remote Sensing of Environment*, 95(3), 303-316.

Hamlet A.F., Lettenmaier D.P. (2005), Production of temporally consistent gridded precipitation and temperature fields for the continental U.S., *Journal of Hydrometeorology*, 6(3), 330-336.

Hubbard, K. G., and X. Lin (2006), Reexamination of instrument change effects in the U. S. Historical Climatology Network, *Geophysical Research Letters*, 33(15).

Lam, N. (1983), Spatial Interpolation Methods: A Review, *The American Cartographer*, 10(2), 129-150.

Lavagnini, I., D. Badocco, P. Pastore (2011), Theil-Sen nonparametric regression technique on univariate calibration, inverse regression and detection limits, *Talanta*, 87, 180-188.

Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J. Burges (1994), A Simple hydrologically Based Model of Land Surface Water and Energy Fluxes for GSMs, *Journal of Geophysical Research*, 99(D7), 14,415-14,428.Liang, X., E. F. Wood, and D. P. Lettenmaier (1999), Modeling ground heat flux in land surface parameterization schemes, *Journal of Geophysical Research*, 104(D8), 9581-9600.

Livneh B, E.A. Rosenberg, C. Lin, B. Nijssen, V. Mishra, K. Andreadis, E.P. Maurer, and D.P. Lettenmaier (2013), A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States: Updates and extensions, *Journal of Climate*, 26(23), 9384-9392.

Mao, Y., B. Nijssen, and D.P. Lettenmaier (2015), Is climate change implicated in the 2013-2014 California dought? A hydrologic perspective, *Geophysical Research Letters*, 42(8), 2805–2813.

Margulis, S. and G. Cortes, M. Girotto, and M. Durand (2016), A Landsat-Era Sierra Nevada Snow Reanalysis (1985-2015), *American Meteorological Society,* 17, 1203-1221.

Menne, M.J, C.N. Williams, and R.S. Vose (2009), The US historical climatology network monthly temperature data, version 2, *Bulletin of the American Meteorological Society* 90(7), 993-1007.

Mote P.W., Hamlet A.F., Clark M.P., Lettenmaier D.P. (2005), Declining mountain snowpack in western North America, *Bulletin of the American Meteorological Society*, 86(1), 39-49.

Mote P.W., D. E. Rupp, S. Li, D. J. Sharp, F. Otto, P. F. Uhe, M. Xiao, D.P. Lettenmaier, H. Cullen, and M. R. Allen (2016), Perspectives on the causes of exceptionally low 2015 snowpack in the western United States, *Geophysical Research Letters*, 43(20), 10980-10988, doi: 10.1002/2016GL069965.

National Centers for Environmental Information (2016), Climate at a Glance, *ncdc.noaa.gov*. [Available at http://www.ncdc.noaa.gov/cag/time-series/us/4/0/tavg/5/3/1895-2016].

Oyler, J. W., A. Ballantyne, K. Jencso, M. Sweet, S. W. Running (2015), Creating a topoclimatic daily air temperature dataset for the conterminous United States using homogenized station data and remotely sensed land skin temperature, *International Journal of Climatology*, 35(9), 2258-2279, doi:10.1002/joc.4127.

Quayle, R.G., D.R. Easterling, T.R. Karl, and P.Y. Hughes (1991), Effects of Recent Thermometer Changes in the Cooperative Station Network, *American Meteorological Society*, 72(11), 1718-1723.

Rohde, R., R. Muller, R. Jacobsen, S. Perlmutter, A. Rosenfeld, J. Wurtele, J. Curry, C. Wickham,, and S. Mosher (2013), Berkeley Earth Temperature Averaging Process, *Geoinformatics and Geostatics,* 1(2).

Seager, R., N. Henderson (2016), Tropical ocean forcing of the persistent North American west coast ridge of winter 2013/14, *Journal of Climate*, 10.1175/JCLI-D-16-0145.1, In press, doi:10.1175/JCLI-D-16-0145.1.

Sen, P.K (1968), Estimates of the Regression Coefficient Based on Kendall's Tau, *Journal of the American Statistical Association*, 63(324), 1379-1389.

Shepard, D.S. (1984), Computer Mapping: The SYMAP Interpolation Algorithm, *Spatial Statistics and Models,* 40, 133 – 145.

Stahl, K., R. D. Moore, J. A. Floyer, M. G. Asplin, I. G. McKendry (2006), Comparison of approaches for spatial interpolation of daily air temperature in large region with complex topography and highly variable station density, *Agricultural and Forest Meteorology,* 139(3-4), 224-236.

Swain, D.L., M. Tsiang, M. Haugen, D. Singh, A. Charland, B. Rajaratnam, and N.S. Diffenbaugh (2014), The extraordinary California drought of 2013/14: character, context, and the role of climate change, *American Meteorological Society*, S3 – S7.

Vose, R.S., S. Applequist, M. Squires, I. Durre, M.J. Menne, C.N. Williams, C. Fenimore, K. Gleason, and D. Arndt (2014), Improved historical temperature and precipitation time series for the US climate divisions, *Journal of Applied Meteorology and Climatology*, 53(5), 1232-1251.

Williams, A.P., R. Seager, J.T. Abatzoglou, B.I. Cook, J.E. Smerdon, and E.R. Cook (2015), Contribution of anthropogenic warming to California drought during 2012-2014, *Geophysical Research Letters*, 42(16), 6819-6828, doi:10.1002/2015GL064924.

Worland, J. (2015), How the California Drought is increasing the Potential for Devastating Wildfires, *Time*, last modified May 8, 2015, http://time.com/3849320/california-drought-wildfires/.

Xiao, M., Nijssen, B. and Lettenmaier, D.P. (2016), Drought in the Pacific Northwest, 1920–2013, *Journal of Hydrometeorology*, 17(9), 2391-2404.

Table 2.1: Inferred statewide trends (in degrees Celsius per century) for minimum and maximum winter temperatures.

Model	Tmin (1920/21-2014/15)	Tmin (1970/71-2014/15)	Tmax (1920/21-2014/15)	Tmax (1970/71-2014/15)
BEST	1.25	2.90	0.87	3.43
H&L*	1.44	2.01	0.90	3.34
PRISM	1.88	3.94	0.62	2.67
SWM	1.20	2.26	-0.30	1.98
VOSE	1.33	2.05	1.19	3.75

* indicates record ends in 2013/14

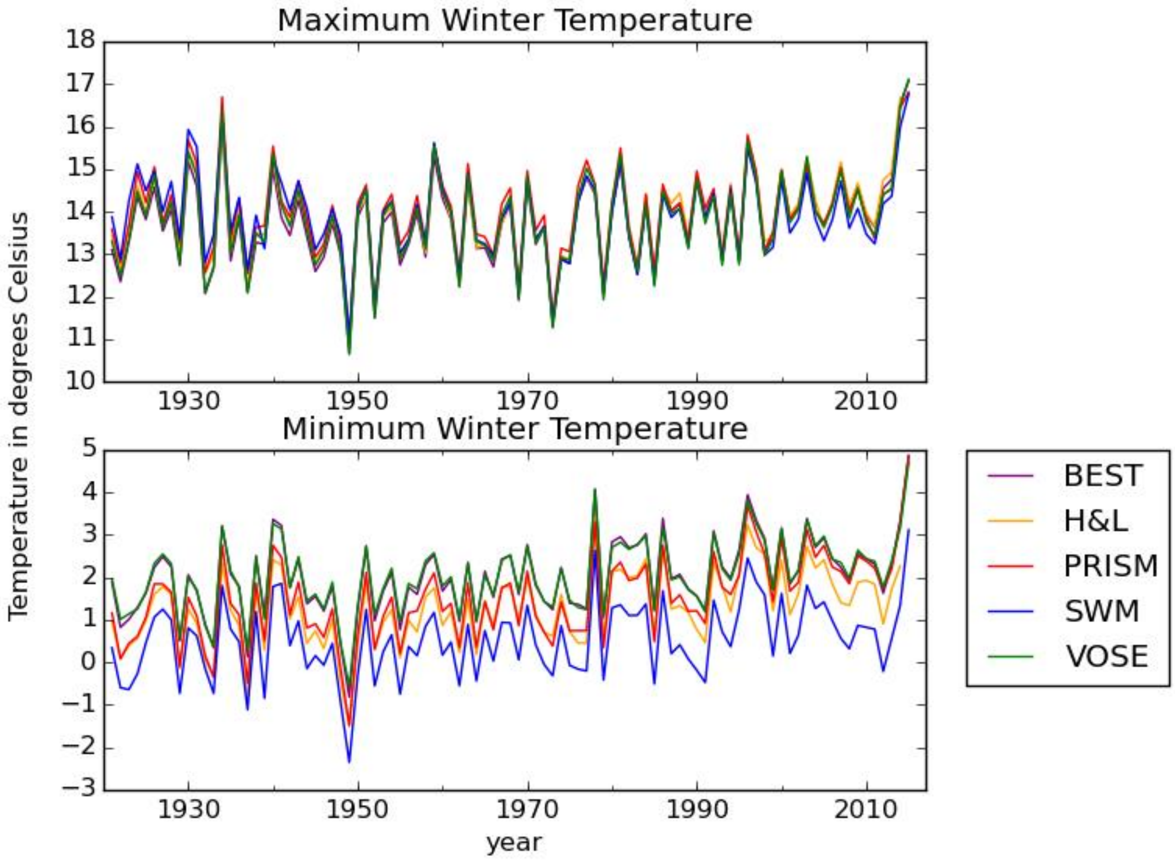


Figure 2.1: Statewide averages of minimum and maximum winter temperatures for gridded data sets BEST, H&L, PRISM, SWM, and VOSE.

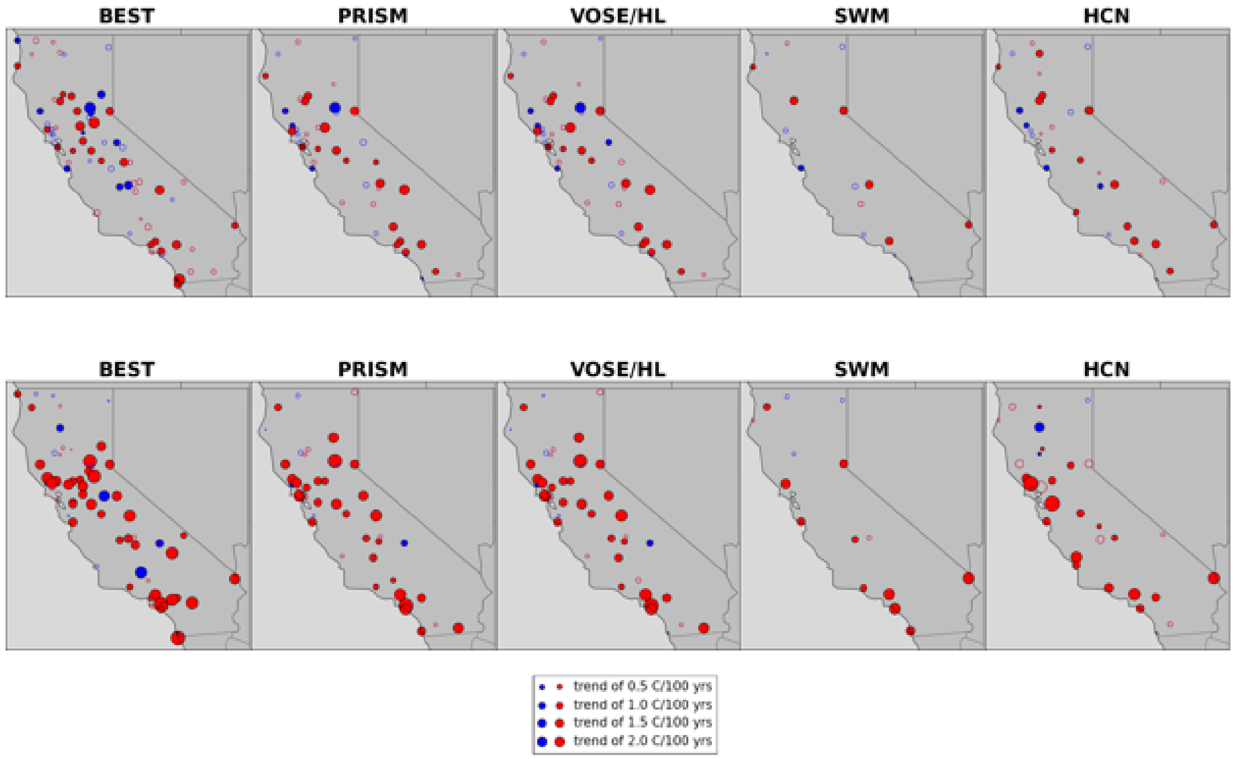


Figure 2.2: Station trends in degrees Celsius per century for (a—top row) Tmax and (b—bottom row) Tmin aggregated to winter season. Red circles signify warming trends; blue circles signify cooling trends. The sizes of the circles correspond to the magnitude of the trend. Open circles represent statistically insignificant trends at the 95% level, while closed or filled in circles are statistically significant at the same level.

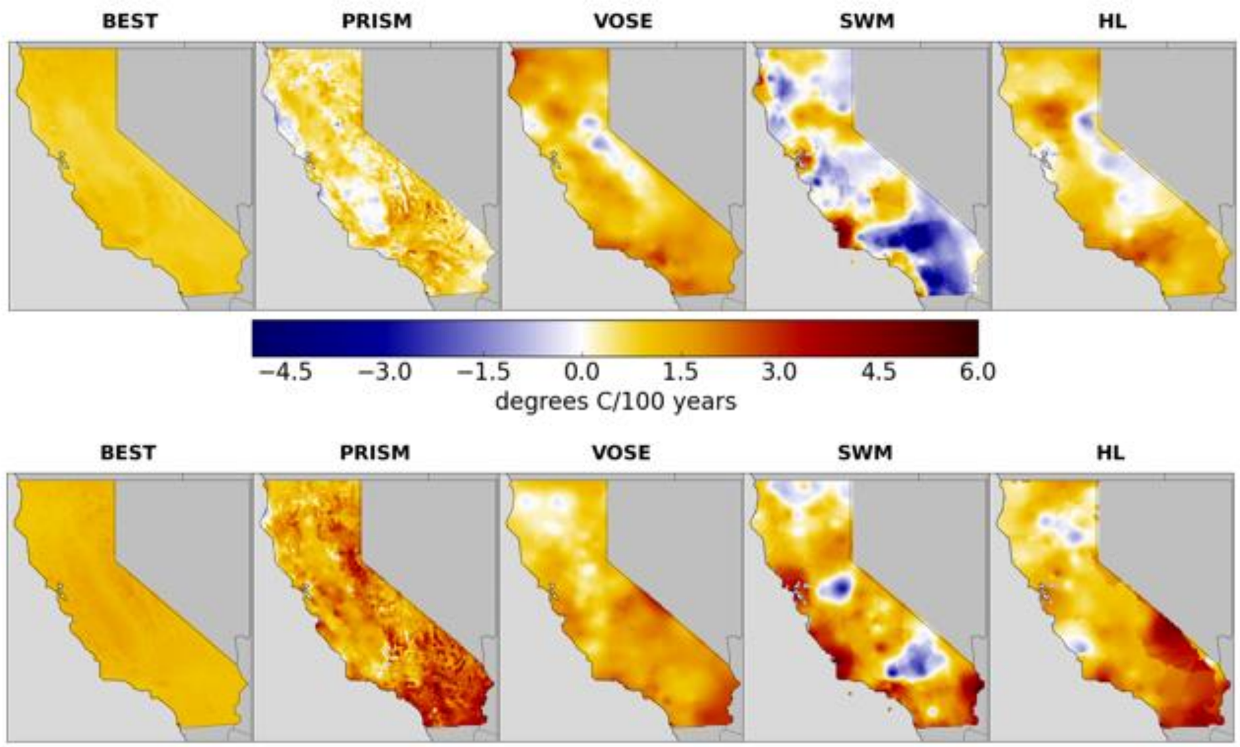


Figure 2.3: Grid cell temperature trends in degrees C per century for (a—top row) maximum temperatures and (b—bottom row) minimum temperatures aggregated to winter season. Orange to red areas indicate increasing or warming trends, while blues indicate cooling trends. Each data set uses the same scale for coloring. VOSE and BEST have the greatest spatial smoothing (related both to spatial resolution and the gridding processes); PRISM and SWM show more texture.

Chapter 3. The bias-variance tradeoff applied to nonstationary modeling of extreme precipitation evaluated using Monte Carlo simulations

This chapter will be submitted to *Journal of Hydrology* as the following. The supplemental material is attached as Appendix B.

Wang, K., K. McKinnon, D. P. Lettenmaier (2022), The bias-variance trade-off applied to nonstationary modeling of extreme precipitation evaluated using Monte Carlo simulations. *Journal of Hydrology* (in prep).

3.1 Introduction

Extreme precipitation events can be costly weather disasters (Kunkel, 2003; Smith and Matthews, 2015) and can further trigger natural disasters such as landslides and floods (e.g. Cavalcanti, 2012). Many recent studies have found increases in observed extreme precipitation, either in intensity, frequency, or both, and in station observations and gridded data (e.g. Karl and Knight, 1998; Easterling et al., 2000; Frich et al., 2002; Goisman et al., 2005; Min et al., 2011; Kunkel et al., 2012; Easterling, et al., 2017). The ability to estimate magnitude and frequency of these events is critical for mitigation of flood damage (Lopez-Cantu and Samaras, 2018), especially in urban areas where storm-induced flooding is closely linked with extreme precipitation [Zhou et al., 2017]. Designing new infrastructure and updating existing infrastructure for theses changes has proven to be challenging (Stedinger and Griffis, 2008), raising many questions about how best to move forward. A common concept used in engineering standards today is the "100-year event" which, in a stationary environment, has a specified magnitude with a 1% chance of occurrence in any one year. Moving into a nonstationary environment, the 100-year event will change depending on the year, or any specified magnitude will have changing chances of occurrence, again depending on the year.

Statistical extreme value theory (EVT) is commonly used to assess extreme events where the events are assumed to be independent and stationary. In EVT, the Fisher-Tippett-Gnedenko theorem states that the underlying distribution will converge to one of three forms characterized by different tail behaviors. All three distributions are special cases of the Generalized Extreme Value (GEV) distribution, which is a family of distributions of which the three forms (Gumbel, Fréchet, and Weibull) are special cases. The GEV has the advantages of being widely used, versatile, and well-understood (Wilks, 1993; El Adlouni et al., 2007; Svensson and Jones, 2010) as well as having more accurate estimates of tail behavior than the Gumbel distribution (Overeem et al., 2008). The probability density function and the cumulative distribution function of the stationary GEV (S-GEV) are shown in Eq. 1 and Eq. 2 respectively where $\mu, \sigma,$ and ξ represent the location, scale, and shape parameters, respectively.

$$f(x;\mu,\sigma,\xi) = \frac{1}{\sigma}\left\{1 + \xi\,\frac{x-\mu}{\sigma}\right\}^{(-1/\xi)-1} exp\left[-\left\{1 + \xi\,\frac{x-\mu}{\sigma}\right\}^{-1/\xi}\right] \qquad \text{Eq. 1}$$

$$F(x;\mu,\sigma,\xi) = exp\left[-\left\{1 + \xi\,\frac{x-\mu}{\sigma}\right\}^{-1/\xi}\right] \qquad \text{Eq. 2}$$

The evidence for increasing extreme precipitation events has led to a growing interest in risk estimation methods that are able to account for nonstationarity; Westra et al. (2013) and Cheng et al. (2014) incorporate nonstationarity into the GEV distribution by directly adding a time-dependent parameter into the model; this linear variation (with time) in the location parameter is shown in Eq. 3. We label this distribution as the nonstationary GEV (NS-GEV). There is discussion around using time versus a physical covariate (Villarini et al., 2009; Vasiliades et al., 2015; Lu et al., 2019), about which parameters may be nonstationary (Cunderlik and Burn, 2003; El Adlouni, 2007), and about use of linear versus nonlinear functions in the distribution (Leclerc and Ouarda, 2007; Gilroy and McCuen, 2012; Um et al., 2017), but essentially the conversion of the S-GEV into the NS-GEV necessitates increasing from a three-parameter model to a four-parameter (or higher) model.

$$\mu(t) = \mu_1 t + \mu_0 \qquad \text{Eq. 3}$$

The interest in developing nonstationary models is perhaps best captured in an article announcing “Stationarity is Dead” (Milly et al., 2008). This article sparked much conversation with responses including phrasing such as “Stationarity is Undead” (Serinaldi and Kilsby, 2015) and “Stationarity is Immortal” (Montanari and Koutsoyiannis, 2014). Part of the difficulty in fitting extreme precipitation data is that despite there being many ways to test for the best fit, in all cases the parent distribution - its parameters and whether it is stationary or not - are fundamentally unknown. Hence, the use of either the stationary and nonstationary version of the GEV distribution will have advantages and disadvantages. If the true distribution is nonstationary, use of the S-GEV will result in biased estimates, and the bias likely will increase with time, especially for estimates made beyond the last observation year. On the other hand, inclusion of the fourth parameter to represent nonstationarity has the effect of increasing the variance of quantile estimates at any fixed time because of the loss of a degree of freedom associated with estimation of the additional parameter; the use of the NS-GEV will lead to higher variability in quantile estimates compared to the S-GEV.

There is currently little literature discussing the costs associated with increasing the number of parameters in the distribution in specific application to the modeling of extreme precipitation. Further, there is a lack of understanding as to what the minimum record length should be to ensure accurate nonstationary estimations. Our objective is simple but necessary as it is currently missing from the conversation of modeling nonstationary extreme precipitation. We evaluate the tradeoff in bias and variance between using the S- and NS-GEV distribution in the environment where the parent distribution is unknown and with limitations of available historical observations.

Following a long line of evaluation of hydrologic risk estimation using synthetic data (e.g., Landwehr, 1979; Adamowski, 1989; Rahman et al., 2002; Nathan et al., 2003; Ngongondo et al., 2011), we use Monte Carlo (MC) simulation techniques to create both S-GEV and NS-GEV parent distributions. While our analysis is based entirely on synthetic data, we generate parent distributions with statistical characteristics that are comparable to historical extreme rainfall across the contiguous US. We then fit either an S-GEV or NS-GEV distribution to each (of many) synthetic data sequences and evaluate the error in the 100-year return period events.

3.2 Data and Methods

We used the quality-checked hourly precipitation data set for the conterminous U.S. (CONUS) from Mishra et al. (2012) as our basis for identifying population parameters for the synthetic data. This data set consists of 1029 stations distributed over the CONUS and spans the 60-year period 1950 through 2009 (although no station has complete coverage over the entire 60 years). We extended the data set through 2013, using the data available from the National Oceanic and Atmospheric Administration (NOAA) via File Transfer Protocol. As we are using these station data to establish realistic boundaries for extreme precipitation characteristics such as CV and skewness, γ, rather than for a detailed analysis of temporal trends, further extensions of the dataset are not deemed critical. Station density in the Mishra data set is fairly uniform over the CONUS, although somewhat sparser in the mountainous regions of the West (see Figure 1). The stations in this set were chosen partially for completeness of record, and any missing data is not replaced.

At each station, we first calculate the annual maxima series (AMS) as the largest hourly precipitation rate each year, and then normalize each AMS by its mean. From the normalized AMS, we find the CV and γ of the observations to establish the parameter space used to create synthetic datasets (the mean is always one due to the normalization). The resulting ranges of CV (top) and γ (bottom) values across CONUS are shown in Figure 1. We use these parameters to establish realistic ranges of CV and γ for extreme precipitation for our synthetic data.

We provide details for one station in the west, the southeast, and the northwest to provide more context as to how our work can be applied to at-site data. The coordinates of each station are given in Table 2. Each station has sixty years of record, ending in 2009. The hydrological regions are based on those outlined in the Fourth National Climate Assessment (USGCRP, 2018) where each region is associated with its own range of mean, γ, and other distinctive properties. We show the 1-hour AMS time series of these three stations in Figure 2 along with a plot of the at-site γ and CV of all stations within the three regions.

At each station, we normalize the AMS by the station's average extreme event so the mean is by definition 1.0. We use these normalized time series to establish a parameter space of CV and γ values based on station data and define the intervals of our simulations as shown in Table 1. This gives 12 sets of GEV location, scale, and shape parameters (3 CV values * 4 γ values). We cover a broad range of possible increases in the mean event, from a 10% increase (NSI = 0.1) per century up to a 50% increase (NSI = 0.5) per century. The compounding effect of nonstationarity increases the possible sets of parameters in the nonstationary environment to 60 combinations (12 parameter sets * 5 NSI values).

Below we describe in detail how we set up the two sets of experiments noted above – the first with stationary parents, the second with nonstationary parents. In both experiments, we first established the true set of statistics (mean, CV, and γ) at t = 0 which represents the present time. For each parent distribution the set of summary statistics (mean, CV, γ) corresponds to a unique set of parameters (location, scale, shape) for the distribution at t = 0.

Our first set of experiments examines the implications of using the NS-GEV when the population is in fact stationary (S-GEV parent; NS-GEV fit). This is the simplest case as it involves the estimation of the additional parameter, μ_1, but does not have compounding effects of nonstationarity in the parent distribution.

For each parent distribution, we generated 10,000 synthetic observation sequences where each sequence (a synthetic AMS) consisted of either 30, 50, or 100 observations (equivalent to

30, 50, or 100 years of record). For each synthetic AMS, t = 0 indicates the first year after end of the record, views as the present time, and *n* indicates the record length, which can cover a time span of 30, 50, or 100 years before the present. We consider a hypothetical project lifespan maximum of N = 50 years, representing the future period (where t > 0). We fit distributions to the synthetic observations as appropriate to each scenario using maximum likelihood estimators for the S-GEV and using the NEVA Matlab package for the time-dependent parameter of the NS-GEV. From the fitted distribution, we found the quantile representing the 100-year storm intensity denoted as $Q(n,t)$.

The above structure of the experiments resulted in 10,000 percentile estimates for the 100-year return period for the future time period ($0 \leq t \leq 50$). We summarize our results in terms of normalized root mean squared error (nRMSE) defined below where $\hat{\theta}$ indicates the fitted $Q(n,N)$ and θ represents the true storm to simplify notation.

$$nRMSE(\hat{\theta}) = \frac{\sqrt{Var(\hat{\theta}) + Bias^2(\hat{\theta},\theta)}}{\theta} \qquad \text{Eq. 5}$$

$$Var(\hat{\theta}) = \frac{\sum \ (\hat{\theta} - E[\hat{\theta}])^2}{n-1} \qquad \text{Eq. 6}$$

$$Bias(\hat{\theta},\theta) = \frac{\sum \ (\hat{\theta}-\theta)}{n} \qquad \text{Eq. 7}$$

The second set of experiments applies the S-GEV when the population is nonstationary (NS-GEV; S-GEV fit). This is a more complicated case as, while there is only one stationary population, there can be an infinity of nonstationary ones; we explore in particular three cases each representing a different relationship between the mean and the coefficient of variation (CV; defined as the standard deviation normalized by the mean) with time. We examine scenarios where the mean extreme storm per century increases from a minimum of 10% to a maximum of 50% over a century. The change per century is termed the nonstationary index (NSI). To fit the NS-GEV, we use the NEVA Matlab package created by Cheng et al. (2014). The time-dependent location parameter is related to the NSI as shown in Eq. 4.

$$NSI = \mu_1 * 100\ years \qquad \text{Eq. 4}$$

For the nonstationary experiments, we allow for three different scenarios differentiated by the behavior of the CV in the scenario: decreasing, constant, or increasing. For all three scenarios, the mean increases at a fixed, linear rate. We considered three cases as examples of an AMS typical of each case shown in Figure 3 where the parameter sets are the same at present time, t = 0.

I. **Decreasing CV:** For a range of NSI = 0.1 to 0.5, we keep the standard deviation constant, hence the CV decreases with time.

II. **Constant CV:** We use the same rates of increase in the mean as in Scenario I and we also increase the standard deviation at the same rate so that the CV remains constant with time.

III. **Increasing CV:** The mean increases at the same range of fixed increases, and the standard deviation both to vary in such a way that the CV increases at a fixed (linear) rate of 0.01/year.

3.3 Results

a. NS-GEV fit to S-GEV data

We first evaluate the results of the simplest case of misapplication: fitting the NS-GEV model to S-GEV data. In Figure 4 we show the histograms of the estimated $Q(n = 50, t = 0)$ separated by CV and γ where estimated $Q(50, 0)$ values are sorted into 20 bins. The difference between present (t = 0) to end of project time (t = 50) is slight, but results from the fact that the fitted non-stationary model will often have a non-zero trend parameter; for the histograms for t = 50 as well as for other record lengths see S1. We found the mean storm matches the true storm best at a γ = 1.25. Below this, the mean storm slight underestimates the true storm; for $\gamma \geq 2.50$, the mean storm slightly overestimates the true storm. If either CV or γ increases, the range of the normalized $Q(50, 0)$ values also increases. With increasing CV and the same γ, the true storm noticeably increases as well. At the lowest γ, the change in CV leads to small shifts of estimated quantiles. At $\gamma \geq 1.25$, changes in CV have a greater impact on the quantile estimates. Thus, the distribution of quantile estimates appears as sharp peaks at the $\gamma = 0.10$, where estimates are within a small range, and flatten as CV and γ increases when estimates cover a larger range.

The top panel of Figure 5 shows nRMSE as a function of CV, γ, and record length. Within the range of reasonable CV and γ based on observations, we find that γ and record length have more effect on error than does CV. Overall, increasing γ, increasing CV or decreasing record length all lead to poorer GEV fits and thus increased nRMSE. Increasing γ and CV inherently leads to wider ranges of possible storm intensity, making these environments harder to accurately model with limited data.

We proceed to split the nRMSE into bias fraction and variance fraction (Equation 5) such that the two components add to 1.0 (100% error), shown in the bottom panel of Figure 5. In the instance of applying a nonstationary model, both bias and variance change with time in relation to the magnitude of the time-dependent slope parameter estimation. We found the estimated slope is often small, since the parent data is stationary, so changes of Q(n, t) with time are slight. However, once the estimated storm crosses the true storm, bias can only increase from that point on. From looking at the fraction variance of nRMSE across the parameter space, we find the dominating factor of the nRMSE is the high variance associated with the additional fourth parameter. Larger γ in the data causes earlier overestimation of the true storm intensity within the historical time period so by present time, variance fraction decreases as γ increases.

b. S-GEV fit to NS-GEV data

The case of applying the S-GEV distribution to NS-GEV data is more likely to represent current practice, but it is also more difficult to simulate as there are infinite ways in which the nonstationary parent distribution could behave. To simplify the possibilities, we considered three specific scenarios: the "decreasing CV" scenario (Scen I) has a time-varying mean but fixed standard deviation; the "constant CV" scenario (Scen II) has both increasing means and standard deviations; and the "increasing CV" scenario (Scen III) is similar to the "constant CV" scenario but with the standard deviation increasing at a faster rate than the mean. As the three scenarios are differentiated by the behavior of the standard deviation with time (the mean increases with time uniformly across all scenarios), nonstationarity behavior is smallest in the “decreasing CV” scenario and largest in “increasing CV” scenario.

In the application of a stationary model, the estimated storm, based on historical record, remains constant with time while the true storm increases, shown in a schematic in Figure 6. Therefore, variance of the estimated storms is unchanging with time; bias must be driving any change in error with time. Typically, the estimated storm intensity over-estimates the true storm intensity in the beginning of its record, matches at some point (this point occurs later with higher CV and γ), and then proceeds to under-estimate the true storm intensity for the remainder of time. At the point where the two storm intensities are equal, the variance is accountable for 100% of the error as bias is 0. As the true storm intensity increases beyond the estimated storm intensity, bias increases, leading to a diminishing variance fraction with time. Whether or not the estimated storm intensity has under- or over-estimated the true storm intensity by the time the present time is reached (t = 0) depends on the CV, γ, and record length of the data.

When NSI is low such as when NSI = 0.1, the nRMSE values are similar to the resulting nRMSE from fitting the NS-GEV to S-GEV data, described in the previous section, meaning the trade-off of lower variance but higher bias leads to similar error magnitudes overall. The difference between the true storm and the estimated storm is small for Scenario I, NSI = 0.1, such that the variance fraction still controls the majority of the error when $\gamma \geq 1.25$. At $\gamma = 0.10$, variance fraction decreases rapidly with project time, beginning the project at 0.8 to 0.9 and decreasing to 0.5 to 0.8 by the end of the project lifespan.

For an example of the partitioning of error into bias and variance, we show in Figure 7 the variance fraction with project time for a chosen pairing of CV and γ under decreasing CV (Scen I). Variance fraction is lower when NSI is higher as a larger NSI indicates the mean storm is increasing at a faster rate throughout the record and the future period, leading to larger biases and accordingly, a larger bias fraction. The point at which the project turns from being variance dominated (variance fraction > 0.5) to being bias dominated (variance fraction < 0.5) depends on the NSI and record length. When NSI is low and record length is short, that point never occurs and instead variance drives the error throughout the project's intended lifespan. When record length is 100 years, bias drives error for nearly the entire project lifespan over all NSIs. Compared to Scen I results, the Scen II and III results show decreases in variance fraction and are discussed further in S2.

c. Station Application

In this section we discuss the errors at three example stations to demonstrate how the results of this paper can be interpreted. Table 2 includes the historical mean, CV, and γ for three precipitation stations chosen to represent different regions based within the US. We use nRMSE points, such as those shown in Figure 5, to interpolate between CV, γ, and record length, to estimate the nRMSE at each station. Significantly, we are using the CV and γ measured at-site as an accurate representation of true CV and γ, even though that may not be the case in practice due to limited data. While there is no significant trend at any of the stations using a Mann-Kendall test at the 5% significance level, the true condition of nonstationarity is unknown, and the possibly underlying NSI is unknown as well.

The station representing the northwest region has a high CV and mid-range γ within our parameter boundaries. Using NSI = 0.1 where applicable, the expected RMSE at the end of the project is in the range of 1.8 to 2.0 mm/hr, depending on which scenario is being followed. When the CV is decreasing or constant with time (Scen I or II), the S-GEV is preferred. If the CV is constant or increasing with time (Scen II or III) and the NSI is above 0.1, the NS-GEV is preferred. It is notable however that the preference towards either model is very slight and results in a close range of errors either way.

The western station has high CV and high γ. Under NSI = 0.1 conditions, the RMSE ranges between 3.2 to 3.5 with the misapplication of the S-GEV resulting in lower error than the misapplication of the NS-GEV for all three scenarios, indicating the price of higher variance in the NS-GEV model outweighs its possible benefits. If CV is decreasing with time, the S-GEV is preferred regardless of NSI. If the CV is constant or increasing with time, the S-GEV is preferred when NSI is at or below 0.3.

The southeastern station has the lowest CV and γ out of the three example stations, making it a more favorable location for using the NS-GEV. Indeed, there is a preference towards using the NS-GEV over all possible NSI and scenarios. This preference is smallest when NSI = 0.1 and the CV is decreasing with time (Scen I). As this site has the largest mean storm of the three sites, this corresponds with the largest range of possible RMSE. Under NSI = 0.1, RMSE ranges between 4.70 to 7.6 mm/hr, depending on the scenario used.

d. Visualization of model preference

The misapplication of the S-GEV model to nonstationary data results in an nRMSE that we will label as $nRMSE_S$. The other misapplication of the NS-GEV model to a stationary environment can similarly be labeled as $nRMSE_{NS}$. Figure 8 compares the error of either misapplication within the parameter space and using a color-code to indicate which nRMSE is larger. This allows us to make a statement about which model leads to larger error when being mis-applied. The metric used is the difference between the two nRMSE values divided by the $nRMSE_S$, which likely represents the current practice of using stationary model in a changing climate.

$$percent\ change = \frac{nRMSE_{NS} - nRMSE_S}{nRMSE_s} \qquad \text{Eq. 8}$$

When the NSI is lower or when the CV is decreasing with time (Scen I), the S-GEV is generally preferred, meaning that the higher variance associated with using the four parameter NS-GEV outweighs the possible likely benefits of having a more accurate (lower bias) estimated storm. The preference is slight, however, with the light coloring indicating that the percent change in error is near 0%. As NSI increases and standard deviation increases with time (leading to Scen II and III conditions), the NS-GEV proves to be possibly useful in application, meaning that despite the application of the S-GEV having lower variance due to being a three-parameter model, the resulting bias is exceedingly large, making the NS-GEV the better option. By the end of the project, the NS-GEV is generally preferred except under NSI = 0.1 or Scenario I conditions. If standard deviation is increasing with time along with mean, leading to Scenario II or III conditions, the preference for the NS-GEV dominates the parameter space.

The record length greatly affects preference choice as it has a strong influence on nRMSE ranges, seen in Figure 5. The preference grids for n = 30 and n = 100 are shown in S3. For the shortest record length and thus the least-informed fitting of the GEV, preference at the beginning of the project is very slight towards the S-GEV distribution and overall, it appears that the nRMSE of either mistaken application is similar to the other. By the end of the project, there is emerging preference for the NS-GEV when nonstationarity behavior is high and data variance is low. When record length is 100 years, at NSI 0.1 under Scenario I conditions, the S-GEV is slightly preferred when $\gamma > 1.25$; there is slight preference to the NS-GEV at $\gamma = 0.10$. At the end of a project lifespan built with a record length of 100 years, there is stronger preference for the NS-GEV across most of the parameter space. Thus the decision on using an S-GEV or NS-GEV model is dependent on the station record length; stations with shorter records will have higher error in model fits and resulting show preference to the S-GEV when NSI = 0.1. Longer records allow for models to be better fit, this serves to decrease the variance enough to encourage use of the NS-GEV only at low γ when NSI = 0.1. At other parameters, the S-GEV is still preferred.

3.4 Discussion and Conclusions

This paper aimed to examine the nRMSE, particularly in the role of bias and variance, in the application of an S-GEV and an NS-GEV when used in the incorrect environment. In the application of the NS-GEV model to S-GEV data, we find variance largely dominates error across the project lifespan. This variance is associated with adding a parameter into the distribution. Applying the S-GEV model to NS-GEV data on the other hand leads to varying ranges of variance fraction, where bias drives error under certain conditions such as high NSI, long record length, or in the increasing CV scenario. In this application, the variance of the estimated storms is constant with time while bias changes with time, leading to different changing variance fractions.

The trade-offs in bias and variance means recommendations of either model is not clear cut but depends most notably on record length, data γ, magnitude of NSI, and scenario choice. Conditions that lead to a smaller variance of estimated storms (a longer record length, lower CV and γ) or signal a more rapidly changing environment (higher NSI, Scen III conditions of increasing CV with time) are well-suited for the application of the NS-GEV model. Opposite conditions favor the application of the S-GEV model.

The ability to accurately discern γ, magnitude of NSI, and scenario choice is not guaranteed, adding further uncertainty to analyses when applied to real station data. Further, limitations of using a short record length are significant. When n = 30, the overall difference between nRMSE of misapplications is small, and model preference is weak. Overall, we found that there must be a large signal of nonstationarity in order for the application of the NS-GEV to be worthwhile.

References

Adamowski, 1989

K. Adamowski

A Monte Carlo comparison of parametric and nonparametric estimation of flood frequencies

Journal of Hydrology, 198 (1989), pp. 295 – 308

Cavalcanti, 2012

I. F. A. Cavalcanti

Large scale and synoptic features associated with extreme precipitation over South America: A review and case studies for the first decade of the 21st century

Atmospheric Research, 118, pp. 27 – 40

Cheng et al., 2014

L. Cheng, A. AghaKouchak

Nonstationary Precipitation Intensity-Duration-Frequency Curves for Infrastructure Design in a Changing Climate

Scientific Reports, 4 (2014), p. 7093

Cunderlik and Burn, 2003

J. M. Cunderlik, D. H. Burn

Non-stationary pooled flood frequency analysis

Journal of Hydrology, 276 (1-4), pp. 210 – 223

Easterling et al., 2000

D. R. Easterling, G. A. Meehl, C. Parmesan, S. A. Changnon, T. R. Karl, L. O. Mearns

Climate Extremes: Observations, Modeling, and Impacts

Science, 289 (5487), pp. 2068-2074.

Easterling et al., 2000

D. R. Easterling, J. L. Evans, P. Y. Groisman, T. R. Karl, K. E. Kunkel, P. Ambenje

Observed Variability and Trends in Extreme Climate Events: A Brief Review

Bulletin of the AMS, 81 (3), pp. 417 – 426

Easterling et al., 2017

D. R. Easterling, K. E. Kunkel, J. R. Arnold

Precipitation change in the United States

Climate Science Special Report: Fourth National Climate Assessment, 1 (2017), pp. 207 – 230

El Adlouni et al., 2007

S. El Adlouni, T. B. M. J. Ouarda, X. Zhang, R. Roy, B. Bobee

Generalized maximum likelihood estimators for the nonstationary generalized extreme value model

Water Resources Research, 43 (3)

Frich et al., 2002

P. Frich, L. V. Alexander, P. Della-Marta, B. Gleason, M. Haylock, A. M. G. Klein Tank, T. Peterson

Observed coherent changes in climatic extremes during the second half of the twentieth century

Climate Research, 19 (2002), pp. 193 – 212

Gilroy and McCuen, 2012

K. L. Gilroy, R. H. McCuen

A nonstationary flood frequency analysis method to adjust for future climate change and urbanization

Journal of Hydrology, 415-415, pp. 40 – 48

Groisman et al., 2005

P. Y. Groisman, R. W. Knight, D. R. Easterling, T. R. Karl, G. C. Hegerl, V. N. Razuvaev

Trends in Intense Precipitation in the Climate Record

Journal of Climate, 18 (9), pp. 1326 – 1350

Groisman et al., 2012

P. Y. Groisman, R. W. Knight, T. R. Karl

Changes in Intense Precipitation over the Central United States

Journal of Hydrometeorology, 13 (1), pp. 47 – 66

Karl et al., 1995

T. R. Karl, R. W. Knight, N. Plummer

Trends in high-frequency climate variability in the twentieth century

Nature, 377 (1995), pp. 217 - 220

Kunkel et al., 1999

K. E. Kunkel, K. Andsager, D. R. Easterling

Long-Term Trends in Extreme Precipitation Events over the Conterminous United States and Canada

Journal of Climate, 12 (8), pp. 2515 - 2527

Kunkel, 2003

K. E. Kunkel

North American Trends in Extreme Precipitation

Natural Hazards, 29 (2003), pp. 291 – 305.

Kunkel et al., 2003

K. E. Kunkel, D. R. Easterling, K. Redmond, K. Hubbard

Temporal variations of extreme precipitation events in the United States: 1895 – 2000

Geophysical Research Letters, 30 (17)

Kunkel et al., 2012

K. E. Kunkel, T. R. Karl, D. R. Easterling, K. Redmond, J. Young, X. Yin, P. Hennon

Probable maximum precipitation and climate change

Geophysical Research Letters, 40 (7), pp. 1402 – 1408

Landwehr et al., 1979

J. M. Landwehr, N. C. Matalas, J. R. Wallis

Probability weighted moments compared with some traditional techniques in estimating Gumbel Parameters and quantiles

Water Resources Research, 15 (5), pp. 1055 – 1064

Leclerc and Ouarda, 2007

M. Leclerc, T. B. M. J. Ouarda

Non-stationary regional flood frequency analysis at ungauged sites

Journal of Hydrology, 343 (3-4), pp. 254 – 265

Lopez-Cantu and Samaras, 2018

T. Lopez-Cantu, C. Samaras

Temporal and spatial evaluation of stormwater engineering standards reveal risks and priorities across the United States

Environmental Research Letters, 13 (7)

Lu et al., 2019

M. Lu, Y. Xu, N. Shan, Q. Wang, J. Yuan, J. Wang

Effect of urbanization on extreme precipitation based on nonstationary models in the Yangtze River Delta metropolitan region

Science of the Total Environment, 673, pp. 64 – 73

Milley et al., 2008

P. C. D. Milley, J. Betancourt, M. Falkenmark, R. Hirsch, Z. W. Kundzewicz, D. P. Lettenmaier, and R. J. Stouffer

Stationarity is Dead: Whither Water Management?

Science Magazine, 319, pp. 573 – 574

Min et al., 2011

S. Min, X. Zhang, F. W. Zwiers, G. C. Hegerl

Nature, 470 (2011), pp. 378 – 381

Mishra et al., 2012

V. Mishra, J. M. Wallace, D. P. Lettenmaier

Relationship between hourly extreme precipitation and local air temperature in the United States

Geophysical Research Letters, 39 (16)

Montanari and Koutsoyiannis 2014

A. Montanari, D. Koutsoyiannis

Modeling and mitigating natural hazards: Stationarity is immortal!

Water Resources Research, 50 (12), pp. 9748 – 9756

Nathan et al., 2003

R. J. Nathan, P. E. Weinmann, P. I. Hill

Use of Monte Carlo simulation to estimate the expected probability of large to extreme floods.

Proceedings of the 28th International Hydrology and Water Resources Symposium, 1, pp. 105 – 112

Ngongondo et al., 2011

C. Ngongondo, C. Xu, L. Gottschalk, B. Alemaw

Evaluation of spatial and temporal characteristics of rainfall in Malawi: a case of data scarce region

Theoretical and Applied Climatology, 106 (2011), pp. 79 – 93

Overeem et al., 2008

A. Overeem, A. Buishand, I. Holleman

Rainfall depth-duration-frequency curves and their uncertainties

Journal of Hydrology, 348 (1-2), pp. 124 – 134

Rahman et al., 2002

A. Rahman, P. E. Weinmann, T. M. T. Hoang, E. M. Laurenson

Monte Carlo simulation of flood frequency curves from rainfall

Journal of Hydrology, 256 (3–4), pp. 196 – 210

Serinaldi and Kilsby, 2015

F. Serinaldi, C. G. Kilsby

Stationarity is Undead: Uncertainty Dominates the Distribution of Extremes

Advances in Water Resources, 77, pp. 17 – 36

Smith and Matthews, 2015

A. B. Smith, J. L. Matthews

Quantifying uncertainty and variable sensitivity within the US billion-dollar weather and climate disaster cost estimates

Natural Hazards, 77 (2015), pp. 1829 – 1851.

Stedinger and Griffis, 2008

J. R. Stedinger, V. W. Griffis

Flood Frequency Analysis in the United States: Time to Update

Journal of Hydrologic Engineering, 13 (4), pp. 199 – 204

Svensson and Jones, 2010

C. Svensson, D. A. Jones

Review of rainfall frequency estimation methods

Journal of Flood Risk Management, 3 (14), pp. 296 – 313

Um et al., 2017

M. Um, Y. Kim, M. Markus, D. J. Wuebbles

Modeling nonstationary extreme value distributions with nonlinear functions: An application using multiple precipitation projections for U. S. cities

Journal of Hydrology, 552, pp. 396 – 406

USGCRP, 2018

D. R. Reidmiller, C. W. Avery, D. R. Easterling, K. E. Kunkel, K. L. M. Lewis, T. K. Maycock, B. C. Stewart

Impacts, Risks, and Adaptation in the United States: Fourth National Climate Assessment, Volume II

Report-in-brief

Vasiliades et al., 2015

L. Vasiliades, P. Galiatsatou, A. Loukas

Nonstationary Frequency Analysis of Annual Maximum Rainfall Using Climate Covariates

Water Resources Management, 29 (2015), pp. 339 – 358

Villarini et al., 2009

G. Villarini, F. Serinaldi, J. A. Smith, W. F. Krajewski

On the stationarity of annual flood peaks in the continental United States during the 20th century

Water Resources Research, 45 (8)

Westra et al., 2013

S. Westra, L. V. Alexander, F. W. Zwiers

Global Increasing Trends in Annual Maximum Daily Precipitation

Journal of Climate, 26 (11), pp. 3904 - 3918

Wilks, 1993

D. S. Wilks

Comparison of three-parameter probability distributions for representing annual extreme and partial duration precipitation series

Water Resources Research, 29 (10), pp. 3543 – 3549

Zhou et al., 2017

X. Zhou, Z. Bai, Y. Yang

Linking trends in urban extreme rainfall to urban flooding in china

International Journal of Climatology, 37 (13), pp. 4586 - 4593

Table 3.1: Parameter space for Monte Carlo experiments.

	Range (increment)	number of values
NSI	0.1 – 0.5 (0.1)	5
CV	0.25 – 0.65 (0.2)	3
γ	0.10; 1.25 – 3.75 (1.25)	4

Table 3.2 – Resulting error at example stations from mistakenly applying the NS-GEV model to S-GEV data by the end of a 50-year project. Example stations have the same record length (60 years). The differences in CV and particularly γ affect nRMSE while the mean dictates magnitude of RMSE.

Station Region	Northwest	Southeast	West
Coordinates	48.54º N, 121.45º W	35.22º N, 80.96º W	39.58º N, 120.37º W
Mean (mm/hr)	8.0	33.2	11.8
CV	0.58	0.34	0.60
γ	1.90	1.06	3.49
RMSE (mm/hr) of mistaken NS-GEV application			
Stationary environment	1.88	4.70	3.51
RMSE (mm/hr) of mistaken S-GEV application, Scenario I, NSI 0.1			
Scenario I	1.80	4.70	3.44
Scenario II	1.79	5.15	3.24
Scenario III	2.03	7.57	3.41
Recommendation	S-GEV, Use NS-GEV if S3	Either, Use NS-GEV if S2 or S3	Strong recommendation for S-GEV

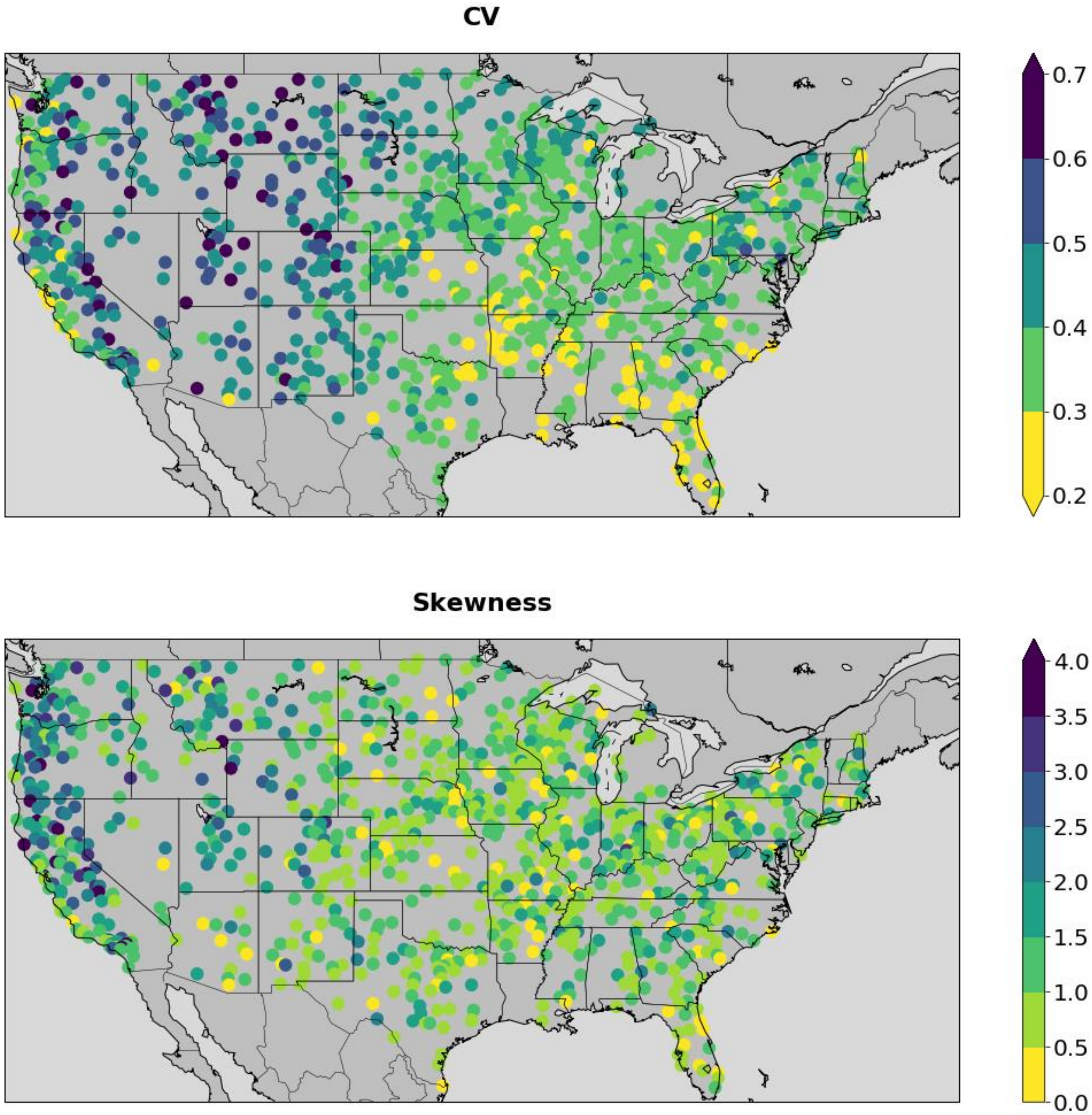


Figure 3.1: (Top) the CV and (bottom) γ of normalized 1-hour annual maxima precipitation series across CONUS. The highest values of both CV and γ are primarily along the west coast and in the Rocky Mountain region. Over half of the stations have a CV between 0.20 and 0.40; about three-quarters of the stations have a γ below 1.0.

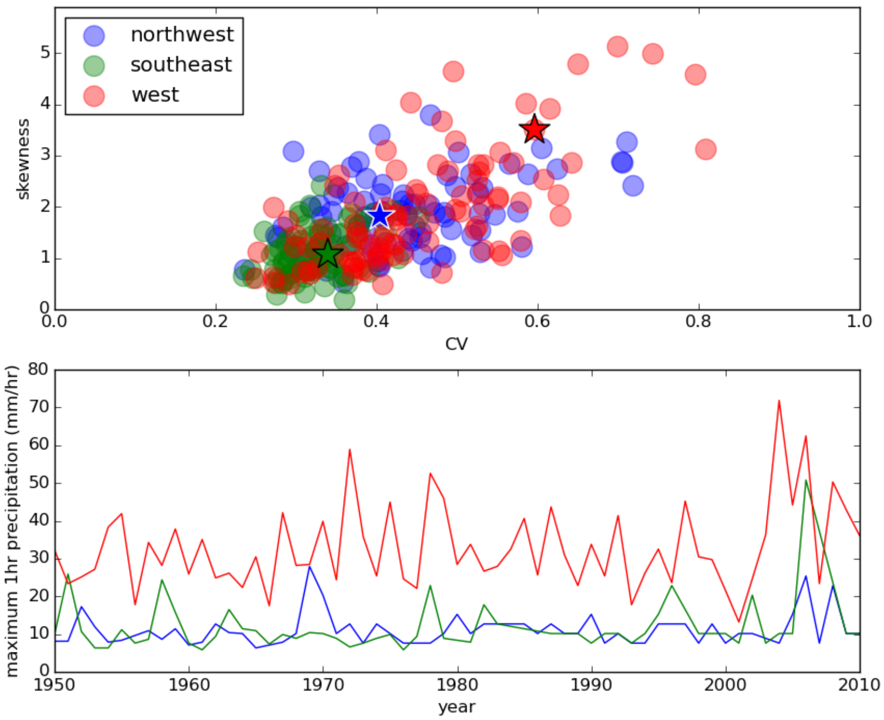


Figure 3.2: (Top) The γ versus CV of stations in northwestern, southeastern, and western regions from common region boundaries established in federal reports. (Bottom) The 1-hour annual maxima precipitation series of the three selected stations that are shown as outlined stars in the top panel.

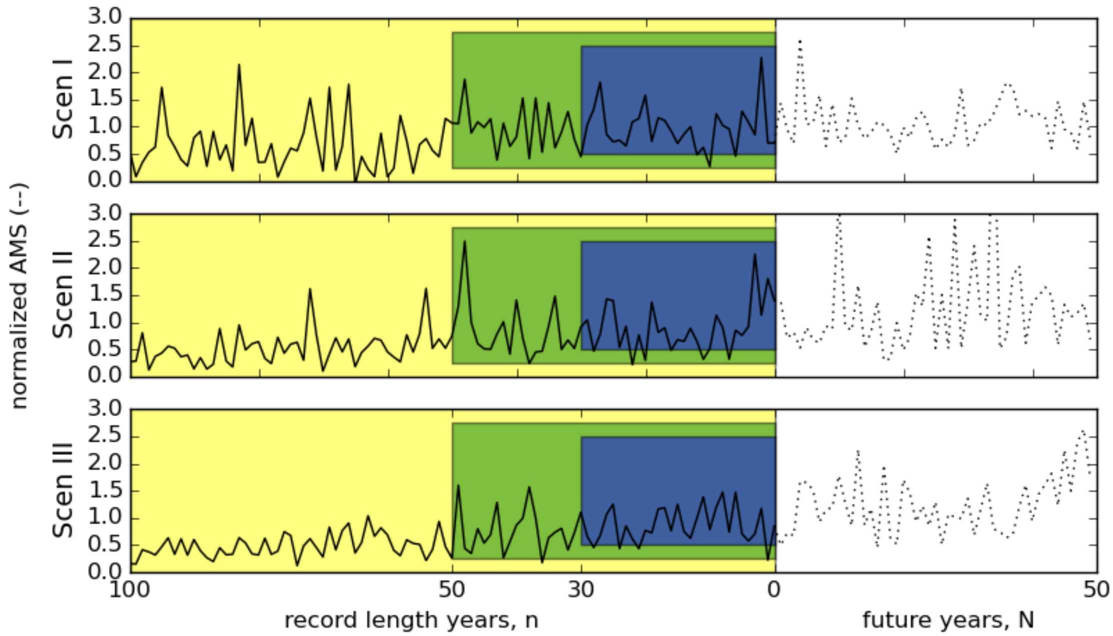


Figure 3.3: Illustration of how the relationship of the CV with time affects the AMS time series when in all cases, the mean is increasing at the same rate. In Scenario I, the CV decreases with time; in Scenario II, the CV remains constant with time; and in Scenario III, the CV increases with time at a constant rate of 0.1/century. Record lengths used are 30, 50, or 100 years, shown as overlapping rectangles (blue, green, and yellow respectively). The present time is represented by t = 0; n refers to historical years while N refers to future years ahead of present time, shown by a dotted line. We show on of 10,000 generated realizations of future data.

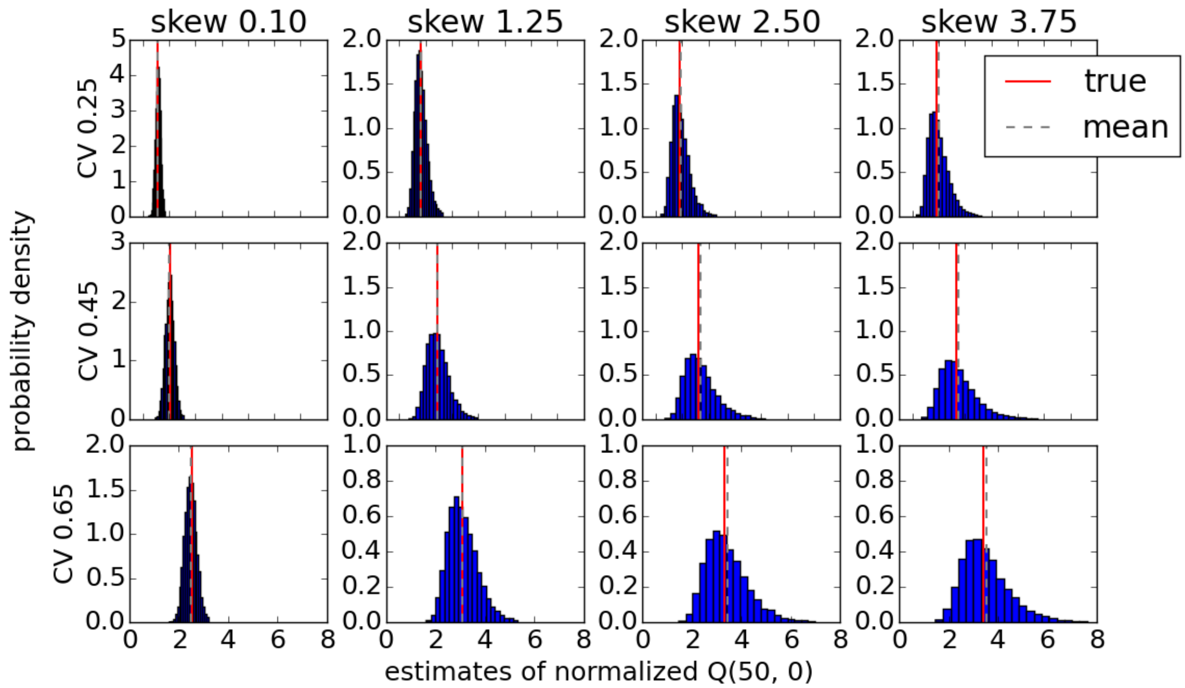


Figure 3.4: Histograms of the estimated Q_{100} across the parameter space at the beginning of the project length. The mean of each probability density function is indicated by a dashed grey line while the true storm event is shown by a red line. The shape of the distribution changes most noticeably with γ where increasing γ leads to a flatter distribution. This is more prominently seen when CV is larger.

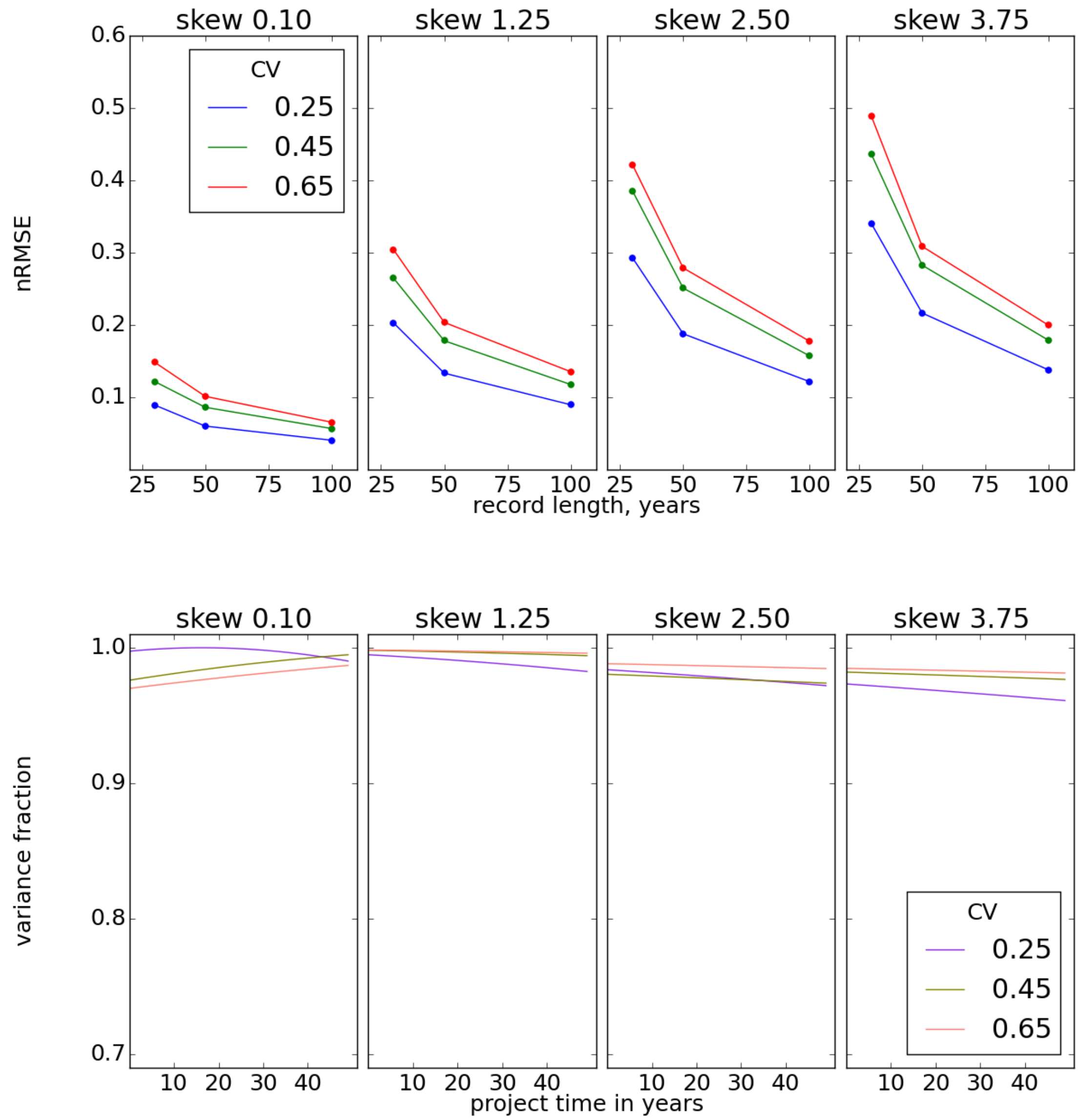


Figure 3.5: (top) nRMSE from fitting NS-GEV model to S-GEV data for Q_{100} at different values of γ, CV, and record length. nRMSE decreases as record length increases, CV decreases, or γ decreases. (bottom) Variance fraction at different values of γ and CV as project time (future period) increases for synthetic data with a record length of n = 50. Variance fraction decreases with project time as well as with increasing γ, largely due how bias increases with time once the estimated storm is moving away from the true storm.

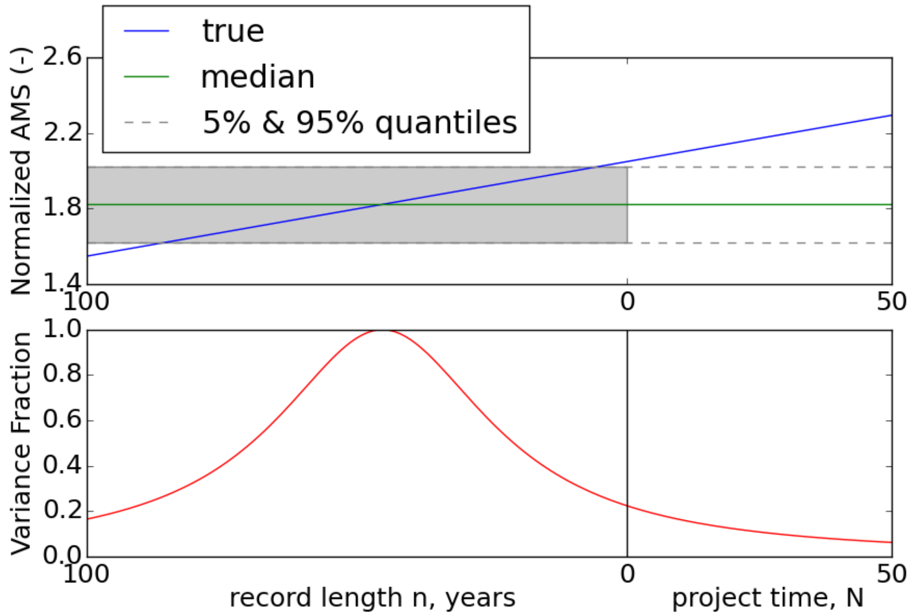


Figure 3.6: Using results from simulations set at CV = 0.45, γ = 0.10, and NSI = 0.5. The top panel shows the true Q_{100} in blue with the median estimated Q_{100} shown in green; dashed grey lines indicate the 95% and 5% estimated Q_{100}s. Grey shading denotes historical data with present time being 0. Bottom panel shows variance fraction increasing to 100% at the moment where the estimated and true storm intensities match. By the time of the project beginning (t = 0), variance fraction is low and continually decreasing with time.

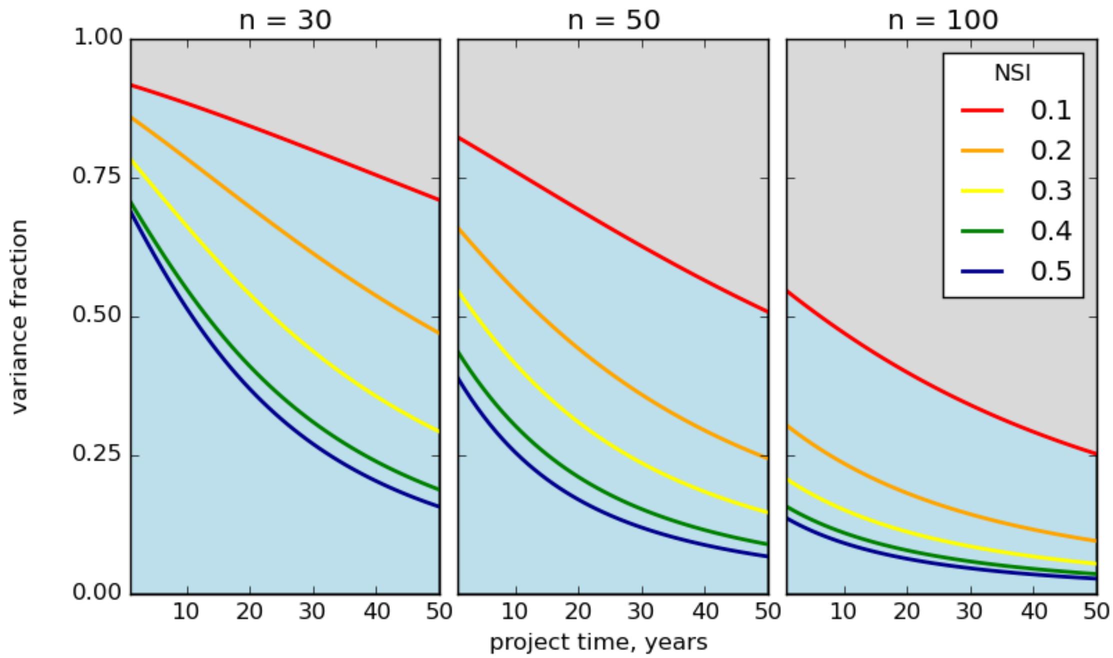


Figure 3.7: Variance fraction under Scenario I conditions for the range of NSIs for the distribution that begins with a CV of 0.25 and γ of 0.10. Blue shading indicates proportion of error attributed to variance for NSI = 0.1. The grey shading gives the complementing proportion of error attributed to bias, again at NSI = 0.1. Bias increases with time while the variance of estimated stationary storms is constant with time, leading to a decreasing variance fraction over project time. Lower NSIs and shorter record lengths have higher variance fractions.

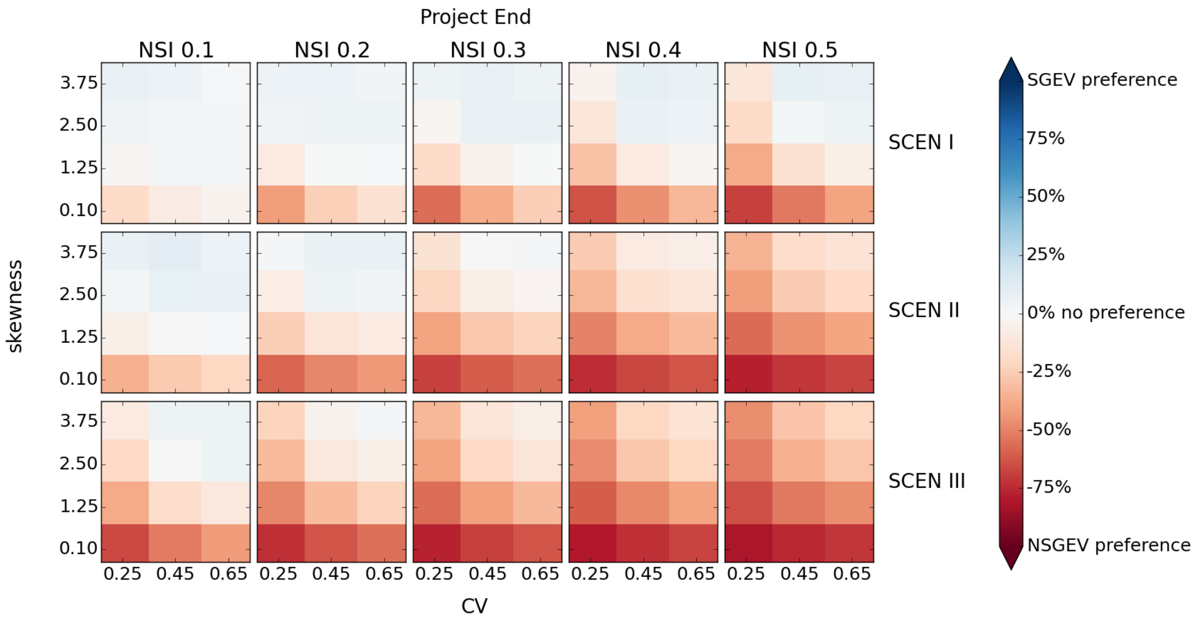


Figure 3.8: Preference grids across the γ -CV space based on a 50-year record length at the end of the project. Darker blue indicates stronger preference for S-GEV while darker red indicates stronger preference for NS-GEV. The preference is based on the percent change as expressed in Eq. 8 where a negative percent change means that $nRMSE_{NS}$ (error from applying the NS-GEV) is smaller than $nRMSE_S$, and therefore application of the NS-GEV is preferred. Positive percentages indicate the $nRMSE_{NS}$ is larger than $nRMSE_S$, and application of the S-GEV is preferred.

Chapter 4. Conclusions and recommendations for future work

4.1. Conclusions

In Chapter 1, I introduce two science questions. The first science question is *"What contributes to the differences in temperature datasets in California, and what is the significance of these differences in the context of the 2011 – 2014 period of drought*?*"*. Chapter 2 examined winter temperature in California in the context of modeling snowpack growth over the season and found that the statewide average minimum winter temperatures differed significantly between datasets. I take a detailed look into the creation of five separate gridded datasets (BEST, HL, PRISM, SWM, and VOSE) all based on station observations to find that the spatial variation of winter temperature trends is highly dependent on the interpolation method chosen to create a gridded dataset from station points while dependence on the station set themselves varies with dataset. The BEST dataset showed the most spatial homogeneity and used a kriging algorithm in its gridding process. Its gridded data showed the least dependence on long-term station observations. The PRISM dataset showed the largest spatial variation and used a linear climate-elevation function that was allowed to vary locally. The relationship between long-term station observations and gridded data is also much clearer than in BEST.

I then run a hydrologic model across the state of California to find how the modeled April 1 SWE is affected by the differences in temperature input of the five datasets. SWM produces the highest average SWE volume of 23.6 km^3 due to its low minimum winter temperatures. PRISM produces the second highest average volume with an estimate of 15.0 km^3. The BEST and VOSE datasets produce the lowest average volumes at 12.6 km^3 and 12.5 km^3 respectively. The high SWE estimate from the SWM runs more closely match SWE estimates from another product primarily based on snow reanalysis from snow model estimates from satellite imaging

(Margulis et al., 2016). The changes between average volume of SWE are significant between the highest to lowest estimate, meaning that the temperature dataset has a significant influence on secondary hydrological variables, and the comparison of resulting SWE estimates could be used to improve understanding of uncertainty associated with SWE calculations. Thus, in selecting a temperature dataset, the user should carefully consider the role of spatial variance necessary in the context of their scientific inquiry.

The second science question asked is *"When does the application of a nonstationary model to extreme precipitation outperform the application of a stationary model when the assumption of stationarity is incorrect?"* In Chapter 3, I used Monte Carlo simulations to find the error of the estimated 100-year storm from fitting a model to an environment with the incorrect assumption of stationarity or nonstationarity. I found that record length and at-site skewness are critical pieces of information that will influence the error magnitude. With a short record length or with high data skewness, either applied model will have poorer fits and more inconsistent estimates resulting in resulting in large variance in errors. In these cases, the potential benefit of having a more accurately modeled storm is overshadowed by the increase in variance of the estimated storms. However, when record length is sufficient and data skewness is low, the nonstationary model and stationary model errors are comparable, with neither model having a large advantage over the other.

The decision making of choosing a stationary versus nonstationary model also relies on an understanding of the behavior of nonstationarity, namely the magnitude of change in the mean and the magnitude of change in the standard deviation of the extreme precipitation data. Determining the magnitude of change and whether it exists or not is no trivial task, and likely cannot be answered by historical observations alone, which have been slow to show statistically significant trends. The signal for nonstationarity must be large before it is practical to use a nonstationary distribution.

4.2. Recommendations for future work

This dissertation only shows two specific examples of how temporal nonstationarity and spatial nonstationarity can affect decision making based on historical station observations when under climate change. In Chapter 2, I found that the choice of temperature dataset as input in a hydrological model significantly impacts the estimation of water resources in the western US. While I used winter temperature and modeled SWE, there are other possible applications of this kind of analysis. A similar study could use summer temperatures and modeled potential evapotranspiration to understand how station coverage and temperature's spatially nonstationary relations affect regional drought vulnerability (Zhou et al., 2020).

To expand on the analysis of Chapter 2, attribution studies could be conducted to understand the relative impact of the differences in the creation of the datasets such as the choice of interpolation in the gridding process, the selection of stations used, and the bias correction of station data where instrumentation had changed. Only two of the five temperature datasets included bias corrections for changes in instrumentation. A characteristic not included in Chapter 2 analyses that could be in future work is the time of observation at each site which is known to produce an observation bias if the schedules are not the same (Karl et al., 1986), which was corrected for in one out of five datasets.

Further one could use the datasets at their original spatial resolutions to study what spatial scale is needed for spatial nonstationarity to appear. Coarser resolutions, such as the original

BEST dataset, tend towards stationarity (Li et al., 2010), so a following question is then "how fine does the resolution need to be to pick up on spatial nonstationarity?". A stationarity index would serve as measure of spatial nonstationarity, and temperature predictors could include elevation, aspect, and distance to the coast. Defining the necessary spatial resolution of the temperature dataset could aid in more accurate modeling estimates of SWE.

The work done in Chapter 3 belongs to a broader and ongoing conversation about the assumption of stationarity or nonstationarity with climate change. One expansion on this work is based on the conclusion that record length is extremely important in deciding on a stationary versus nonstationary model. To optimize the usefulness of at-site data, records should be supplemented by whichever means necessary. A common pathway for this is regional frequency analysis, which substitutes spatial data for temporal data, and can reduce variance in the estimates of parameters in short record lengths (Hosking and Wallis, 1993). A similar analysis would follow the outline of Chapter 3 but incorporate regional frequency analysis where currently at-site estimates are used. This additional information can help decision makers in knowing if they have enough data to use a nonstationary model.

An additional benefit to regional frequency analysis is that regional trends may better represent the true magnitude of nonstationarity within a region (Kjeldsen and Prosdocimi, 2021). Understanding the behavior of nonstationarity is critical for decision-making in which model to use. The use of projection outputs can be useful in this context but is outside the scope of the work done in Chapter 3. However, Monte Carlo simulations can be used to show what minimum criteria are for different statistical trend tests to reliably find significant trends for different magnitudes and behaviors of nonstationarity, namely the change in the mean and standard deviation of extreme precipitation. This understanding would further our understanding about the risk of making type II errors in which there is a false negative result to a trend test.

In my work, I examine two instances of how the ability or lack thereof to incorporate nonstationarity into historical station observations can impact water resource management decisions. In comparing the gridded temperature datasets across California, I consider how station coverage as well as choice of regression model can lead to different estimates of temperature across a field with each product showing different spatial variety in examined trends. The Monte Carlo simulations of extreme precipitation directly expands the stationary GEV probability distribution into a nonstationary one, and I consider the cost of increased variance associated with adding a time-dependent component to represent temporal nonstationarity and ultimately make recommendations for when a nonstationary GEV model is most and least suitable. The goal of the proposed extended work is to further our understanding of the limits of station reliability in making water resource decisions under the context of a warming earth. There remains a tremendous amount of work to do in understanding how hydrologic changes translates to changes in water resource management practices.

References

Hosking, J. R. M., J. R. Wallis (1993), Some statistics useful in regional frequency analysis. *Water Resources Research*, 29(2), 271 – 281.

Karl, T. R., C. N. Williams Jr, P. J. Young, W. M. Wendland (1986), A Model to Estimate the Time of Observation Bias Associated with Monthly Mean Maximum, Minimum and Mean Temperatures for the United States. *Journal of Applied Meteorology and Climatology*, 25(2), 145 – 160.

Kjeldsen, T. R., I. Prosdocimi (2021), Assessment of trends in hydrological extremes using regional magnification factors. *Advancements in Water Resources*, 149.

Li, A., Z. Zhao, X. Miaomiao, Y. Wang (2010), Investigating spatial non-stationary and scale-dependent relationships between urban surface temperature and environmental factors using geographically weighted regression. *Environmental Modelling and Software*, 25(12), 1789 – 1800.

Zhou, J. et al. (2020), Choice of potential evapotranspiration formulas influences drought assessment: A case study in China. *Atmospheric Research*, 242.

Appendix A

How much have California winters warmed over the last century? – Supplemental Materials

Appendix A provides supporting information for Chapter 2, which has been published in its current form in the Geophysical Research Letters © American Geophysical Union. Modifications for formatting.

Wang, K. J., A. P. Williams, D. P. Lettenmaier (2017), How much have California winters warmed over the last century? Geophysical Research Letters, 44(17), 8893 – 8900. https://doi.org/10.1002/2017GL075002

Introduction

In this supplementary document, we provide more details as to the datasets, methods, and results in the paper. We also detail the procedures used to produce the five gridded datasets, and the relevant station lists. In addition, we provide more comprehensive trends statistics at the grid cell level for each of the gridded dataset.

A.1 Dataset Summaries

The five gridded datasets we utilize in the paper are the National Climatic Data Center's gridded National Temperature Index data (VOSE; Vose et al. 2014), the Berkeley Earth Surface Temperatures (BEST; Rohde et al. 2013), the Precipitation Regression on Independent Slopes

Method (PRISM; Di Luzio et al. 2008), UCLA's California and Nevada Drought Monitoring System (SWM; Xiao et al. 2016), and the Hamlet and Lettenmaier (HL; Hamlet and Lettenmaier, 2005) gridded dataset.

a) VOSE

VOSE draws on over 6,000 stations within the Global Historical Climatology Network within the United States, of which 336 are in California. Both recorded and unrecorded changes in instrumentation are detected by using a pairwise comparison method, and the station data are adjusted for changes in time of observation, location, instrumentation, and site condition. The minimum length of record requirement is 10 years of valid monthly data since 1950 [Vose et al. 2014].

Monthly climatology values are simple averages over the base period 1981 – 2010. Missing values are estimated by using least absolute deviation regression with input from up to five neighboring stations. To create a grid with a resolution of 1/24th degree, Vose et al. [2014] use a thin-plate smoothing spline with an objective function that creates a continuous surface that minimizes a function that accounts for roughness via a smoothing parameter and includes a roughness penalty function as well as station data and station location. The smoothing parameter is obtained by minimizing a generalized cross validation using a subset of stations. Anomalies are calculated by subtracting the station's observed temperature from the climate normal for that month. Large (compared to neighboring) anomalies are smoothed by using a residual-based index and weighting process. The anomalies are interpolated to a grid using inverse-distance weighting, using input from 15 to 25 neighboring stations. The final daily temperature product is created by adding the anomaly grid to the climate normal grid. The dataset is updated monthly, including statewide monthly averages. The VOSE dataset is publicly available at ftp://ftp.ncdc.noaa.gov/pub/data/climgrid/.

b) PRISM

The PRISM dataset uses stations in the Cooperative Observer Program (COOP), Historical Climatology Network (HCN), and Snow Telemetry (SNOTEL) networks. The stations used in PRISM allow shorter minimum periods of record than other gridded sets like VOSE and SWM. Use of SNOTEL station observations allows for better determination of temperature lapse rates in mountainous areas than using only long-term stations, which are mostly at low elevations, but introduce their own measurement biases [Oyler et al. 2015]. Missing values are replaced with estimates based on neighboring stations that have been spatially interpolated using various seasonally and geographically dependent methods so station records are serially complete. PRISM uses a decomposition-interpolation-composition approach along with climatologically-aided-interpolation to create the gridded dataset. PRISM uses station data and a simple linear regression model to establish a climate-elevation relationship for each grid cell. Several factors other than elevation are considered, such as coastal influence and inversions [Di Luzio et al. 2007]. Depending on how each factor affects temperature over the base period 1961 – 1990, it is given a certain weight. Each station then has a combined weight as a result.

Daily anomaly-average ratios are calculated by subtracting the station's daily observation from the PRISM monthly value, and the anomaly is then divided by the PRISM monthly value. These ratios are gridded by interpolating via inverse-distance weighting using input from up to 12 neighboring stations. The final daily temperature product is the gridded ratio times the PRISM monthly value added to the PRISM monthly value. The PRISM dataset is publicly available at http://prism.nacse.org/.

c) BEST

The BEST dataset is global and has the coarsest native resolution of all the datasets we utilized. It uses 657 stations in California with temperature observations. Unlike other datasets, which use bias adjustments to ensure a continuous record, the BEST method splits the records into segments whenever there is a gap in the record longer than a year, a change in station location, time of observation, or other breakpoints, which are automatically detected [Rodhe et al. 2013]. Stations with record lengths under one year or with missing metadata were not used.

To create the gridded dataset, a continuous temperature field is first created for each month's normal values by using kriging to interpolate the station data. The kriging process uses correlations between the stations' temperature and distances between the stations to estimate temperature at any intermediate point while avoiding bias from station clusters. The monthly average, calculated as a simple arithmetic average, of the results from kriging are compared to the station observations' average monthly temperature. Observations that strongly differ from the kriging results are down-weighted so their effect as an outlier is reduced. The process is iterated, using the new weight adjusted station data to create an updated field via kriging until the least-squares differences between the adjusted station data and the kriging results are minimized. BEST incorporates a spatially varying lapse rate dependent on nearby stations to model temperature dependence on elevation as a function primarily used to establish climatology.

The publicly available BEST product consists of temperature anomalies rather than actual temperature values. We used the BEST anomalies to estimate actual monthly temperatures by using mean fields from the Oyler et al. [2015] dataset. The BEST anomalies were first bilinearly interpolated to 1/24th degree resolution and the Oyler et al. [2015] dataset (with a native resolution of 1/120 degrees) was aggregated to 1/24th degree resolution. For each grid cell and month, the BEST time series of annual values were linearly adjusted to match the mean and standard deviation of the Oyler et al. values for the base period 1961-2010. The BEST dataset is available at http://www.berkeleyearth.org/ and the Oyler et al. dataset is available at https://github.com/jaredwo/topowx.

d) SWM

The SWM dataset is produced as part of a near -real- time drought monitoring system that relates current (for model-derived soil moisture and snow water equivalent) to historic conditions. Thus the SWM production requires relatively long climatologies. Stations with long records and real-time reporting are best suited for this dataset. SWM uses 102 COOP stations in California. We used a serially complete dataset created by Livneh et al. [2013] gridded at 1/16th degree spatial resolution to find an average annual temperature climatology over the base period 1961 – 2010. The Livneh dataset has 20-year minimum record requirement and is gridded using the SYMAP algorithm. Erroneous values are identified based on their monthly coefficient of variation and are removed. Temperatures are orographically adjusted using a constant lapse rate of 6.5°C/km.

In addition to these climatology values, SWM computes annual temperature averages over the same base period for each station. Anomalies are calculated by subtracting the station's daily observation from the station's climatology. These anomalies are gridded to the 1/16th degree spatial resolution by inverse-distance weighting with input from up to five neighboring stations. The final temperature product is computed by adding the climatological values based on the external (Livneh) dataset to the gridded anomaly values.

e) Hamlet and Lettenmaier

The HL dataset [Hamlet and Lettenmaier, 2005] first creates two smoothed products from different networks, both of which are used to create a final adjusted product. One product is created from the Global Historical Climatology Network - Daily product (GHCN-D), which includes many stations from the NOAA Cooperative Observer network, and the second is from HCN data. Stations are first screened for a minimum of one continuous year of data, or five years total of data, and the observations are quality checked. The station data are then regridded using the SYMAP algorithm, as in SWM, at 1/16th degree. The daily time steps for GHCN-D stations observations are then aggregated to monthly values, then the interpolated datasets for GHCN-D and HCN data are matched to each other. An adjusted product is created based on the raw data using the smoothed data as adjustments, and a daily time series can be scaled based on the monthly value [Hamlet et al. 2007]. The PRISM temperature climatology values (1971 – 2000) are used to bias correct the final product.

A.2 Inferred Trends for 1971 – 2015 Time Period

We aggregated temperatures from each dataset to form statewide averages (on a monthly basis) for all datasets and computed the Theil-Sens estimators at the 95% confidence level over the shorter time period starting from the winter of 1970/71 to the end of the record for each gridded product. This time period was chosen since it is known for its significant warming [e.g., Williams et al. 2015]. The trend magnitudes, reported as degrees C per century, are generally larger over the post-1970 period compared to the post-1920 period. In Figure S1 below, long-term trends are shown in grey over the statewide temperature time series, in blue.

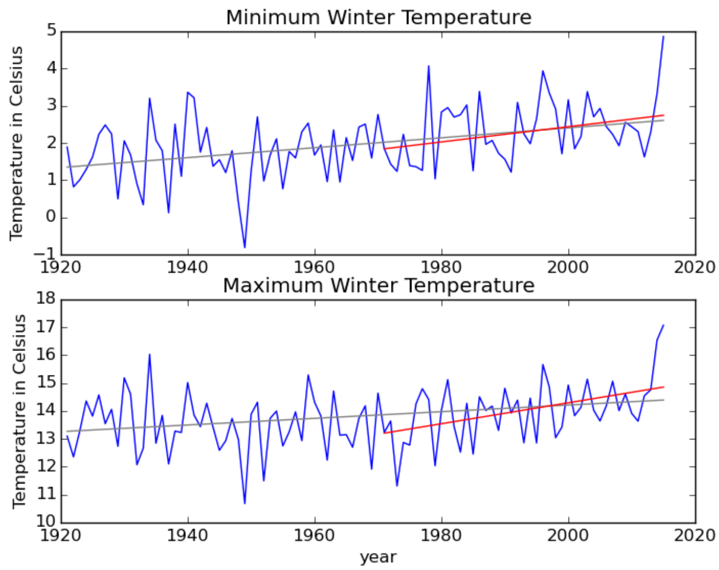


Figure A1: (a) Theil-Sens estimators at the 95% confidence level for the BEST dataset averaged over the state with the 1920/21 to 2014/15 trend in grey and 1970/71 to 2014/15 in red

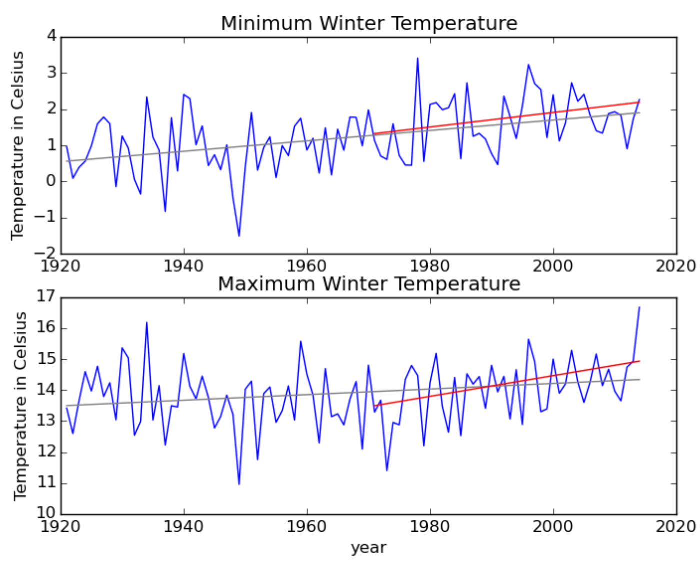


Figure A1: (b) Theil-Sens estimators at the 95% confidence level for the HL dataset averaged over the state with the 1920/21 to 2013/14 trend in grey and 1970/71 to 2013/14 in red.

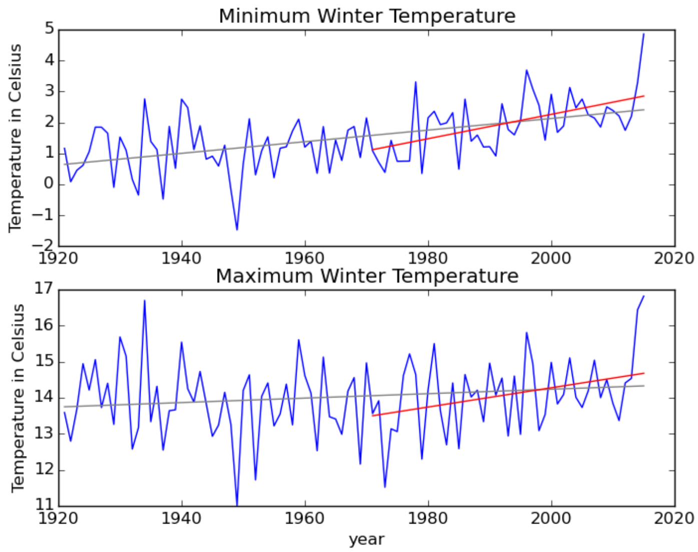


Figure A1: (c) Theil-Sens estimators at the 95% confidence level for the PRISM dataset averaged over the state with the 1920/21 to 2014/15 trend in grey and 1970/71 to 2014/15 in red.

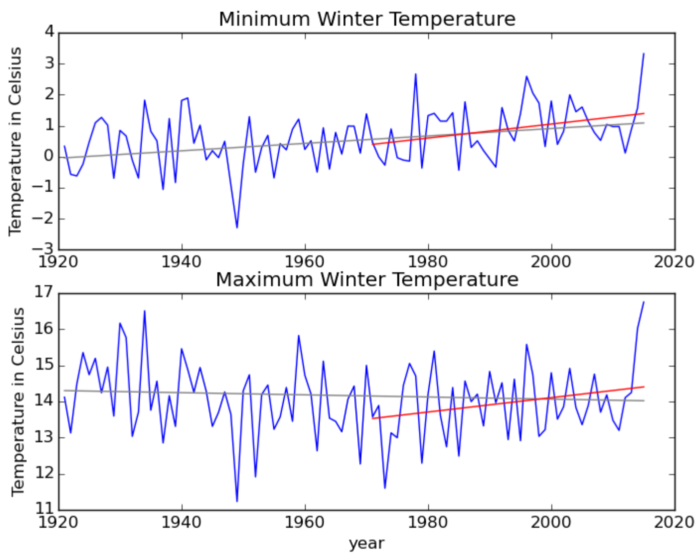


Figure A1: (d) Theil-Sens estimators at the 95% confidence level for the SWM dataset averaged over the state with the 1920/21 to 2014/15 trend in grey and 1970/71 to 2014/15 in red.

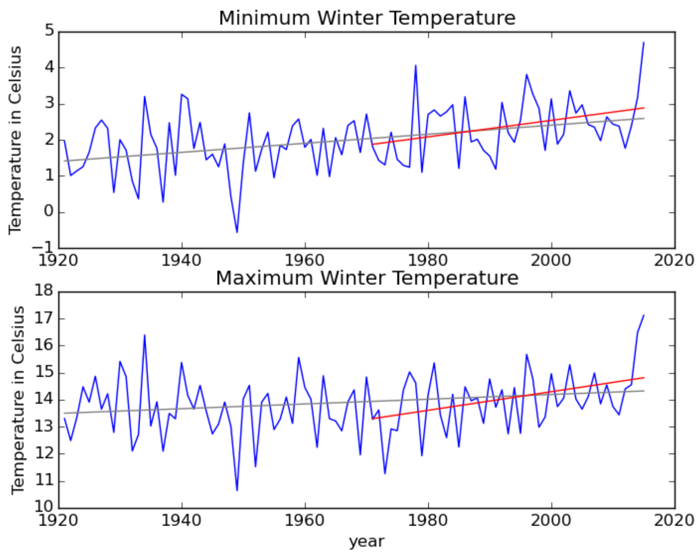


Figure A1: (e) Theil-Sens estimators at the 95% confidence level for the VOSE dataset averaged over the state with the 1920/21 to 2014/15 trend in grey and 1970/71 to 2014/15 in red.

A.3 HCN analysis

The Historical Climatic Network (HCN) is a network of a subset of NOAA Cooperative Observer stations with lengthy records, of which there are 54 in California. The station data have been subjected to rigorous data quality checks as described in Menne et al. [2009] and adjustments have been made where necessary for station moves and changes in instrumentation. There is no gridded product that uses only the HCN network, so we limited our analysis of HCN to averages of the 54 California HCN stations.

We aggregated the HCN station data to statewide averages of winter temperatures and replaced missing winter seasons by first identifying the five nearest stations that have records for each of the missing years of the station of interest. Next, we used up to thirty years to establish a climatology for each station. We chose the thirty-year period such that all nearby stations and the station of interest have minimum missing values within the time span. We found the anomalies of the nearby stations by subtracting the station's winter temperature of the missing year in question from the climatology of that station. We took the average of the five anomalies and subtracted it from the station of interest's climatology to replace the missing year's value.

The results are shown in Figure S2. The averaged HCN product leads to a consistently higher minimum and maximum temperature than the ensemble of all the gridded products (because the stations are predominately at low elevations). The HCN minimum product is on average 4.9°C higher than the ensemble and has a trend of 1.80°C/century, which is the second highest trend after PRISM. The maximum product is 2.2°C higher than the ensemble and has a trend of 0.205°C/century, which is the second lowest after SWM.

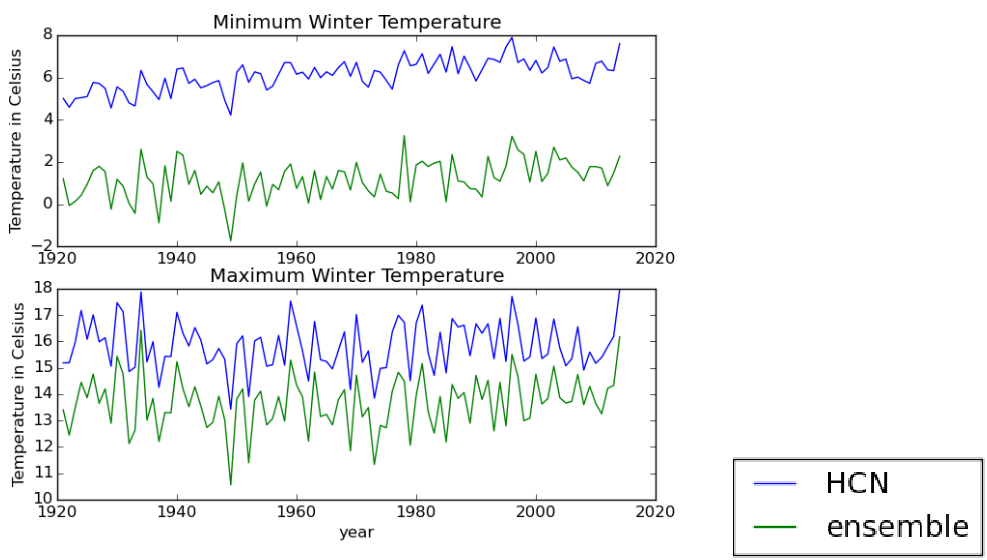


Figure A2: State-wide averaged station observations of minimum and maximum temperature for HCN

A.4 Station Analysis

a) Station Lists

To calculate temperature trends for the station data, we first used two screens to filter out stations with substantial missing data. The first established winter averages for each year based on having all five winter months reported from the monthly datasets, and 90% of winter days reported from the daily datasets. Non-leap year winters have 151 days, so a station must report for at least 136 of those days. If there were not enough reporting days, the winter for that year was considered incomplete and the entire winter was disregarded in the analysis. The second screen was for stations with at least 90% of winters reporting based on length of record. Specifically, for the 95-year period 1920/21 to 2014/15, at least 86 years must have had valid data according to the first pass screens. Rather than attempting to fill in missing data, we applied the Theil-Sen slope estimator, which is designed so that it is robust to a few missing years.

To illustrate the overlap of long-term stations within each product's station list, Figure S3 shows the percent of stations that each product has in common with the station list for PRISM. PRISM, by construct, is at 100% for Tmax and Tmin. BEST has the fewest stations in common in either Tmax or Tmin networks. For SWM, 12 out of 15 stations in the SWM long-term Tmin station list are also in the PRISM long-term Tmin station list, and 12 out of 17 stations in the SWM Tmax list are also in the PRISM Tmax list. For HCN's Tmin list, 23 out of 31 stations are in common with PRISM while 24 out of 33 stations are in common for Tmax. The entirety of the

VOSE Tmin and Tmax long-term stations are within the PRISM long term station networks.

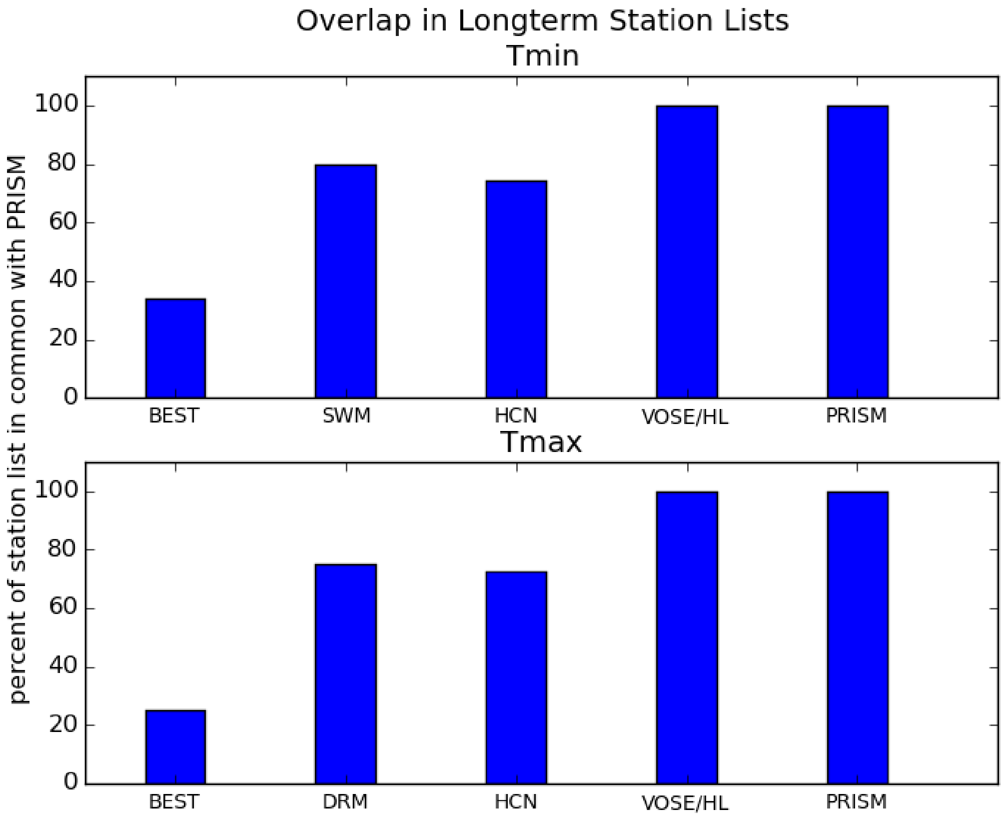


Figure A3: Percent of long term station list each product has in common with PRISM for Tmin (top panel) and Tmax (bottom panel). By construct, PRISM plots at 100% for both measures.

b) Station Instrumentation Changes

As noted in Section 2.1 of the main text, the primary instrumentation shift over the last century in the Cooperative Observer network was in the mid-1980s from liquid-in-glass thermometers hosted in cotton regional shelters to a thermistor-based maximum-minimum temperature system (MMTS). Metadata from the Historical Observing Metadata Repository enabled us to find specific dates of instrumentation change for many (though not all) stations. In particular, we examined the station network for SWM, although we know that BEST, PRISM, and HL also do not make adjustments for this change.

Of the 102 SWM stations in California, 62 had sufficient documentation to identify a switch in instrumentation to MMTS, mostly in the mid-1980s (although some changes to MMTS occurred later). Quayle et al. [1991] found that daily Tmax changed by roughly 0.3°C after the instrumentation switch while daily Tmin changed by roughly minus 0.4°C. We applied the bias correction suggested by Quayle et al. [1991] to each of the 62 stations on the date of the documented switch. Then we re-gridded using the same algorithm as the original SWM dataset. Figure S4 compares the modified SWM dataset with the original when aggregating to a statewide and winter average. For minimum winter temperatures, the adjustment causes a slight cooling effect following the period when many of the instrument switches occurred. Maximum winter temperatures are overall similar to the original dataset. Applying Theil Sens estimators over the 1920/21 to 2014/15 winters, we find that minimum statewide winter temperature trends are reduced to 0.99°C/century (compared to 1.20°C/century), and the maximum statewide winter temperature trends have increased to -0.052 °C/century (from -0.297°C/century). Notably, the SWM dataset is unique in that it is the only one to indicate a negative trend in statewide winter maximum temperature.

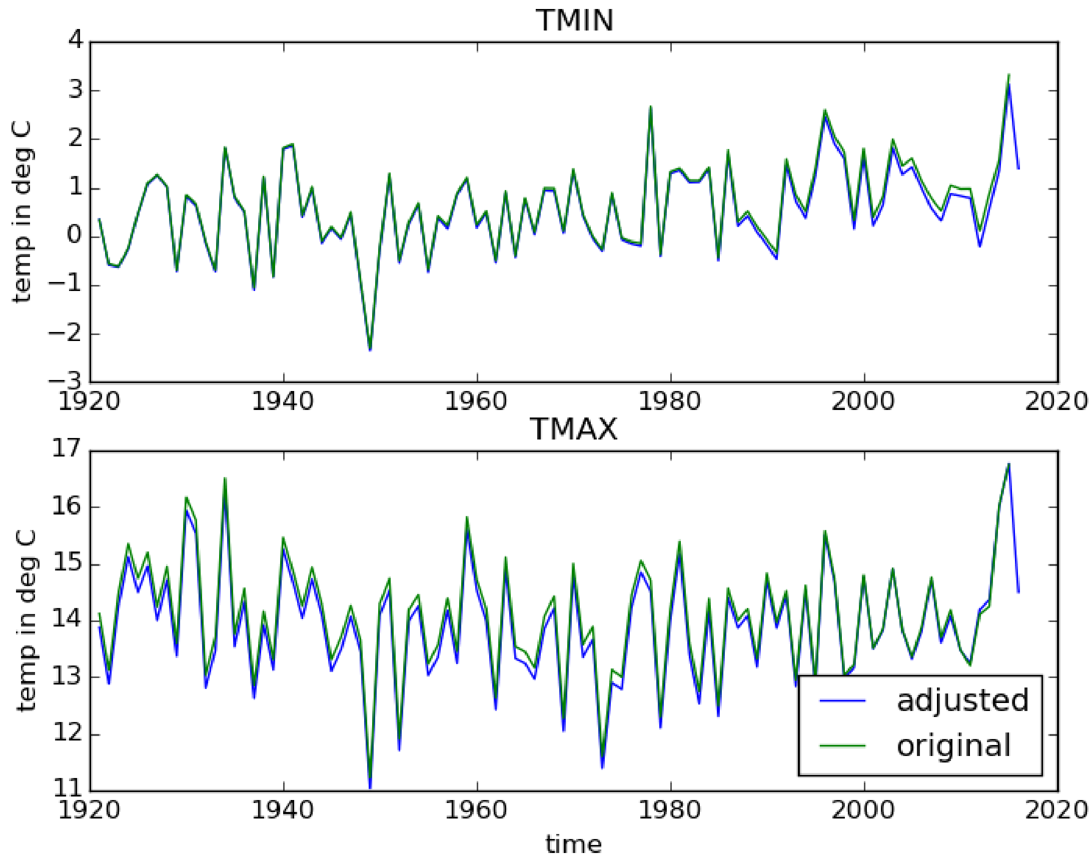


Figure A4: Time series of statewide average winter temperatures. Minimum temperatures are shown on top and maximum are on bottom. The original SWM dataset averages are shown in green, while the adjusted dataset (created to take out the bias of COOP system-wide instrument change largely in the 1980s) winter temperatures are shown in blue. Minimum temperatures are similar with a separation occurring in the mid-1980s after which the adjusted dataset is slightly cooler. Maximum temperatures see slight disagreement with the adjusted set being cooler in early 1900s.

Evaluation of temperature trends on the grid cell level reveals that the overall spatial pattern and magnitudes look similar to the original SWM dataset (Fig. S5). Some basic statistics are shown in Table S1. As was the case for the statewide trends, the bias correction generally causes winter Tmin trends to decrease and winter Tmax trends to increase.

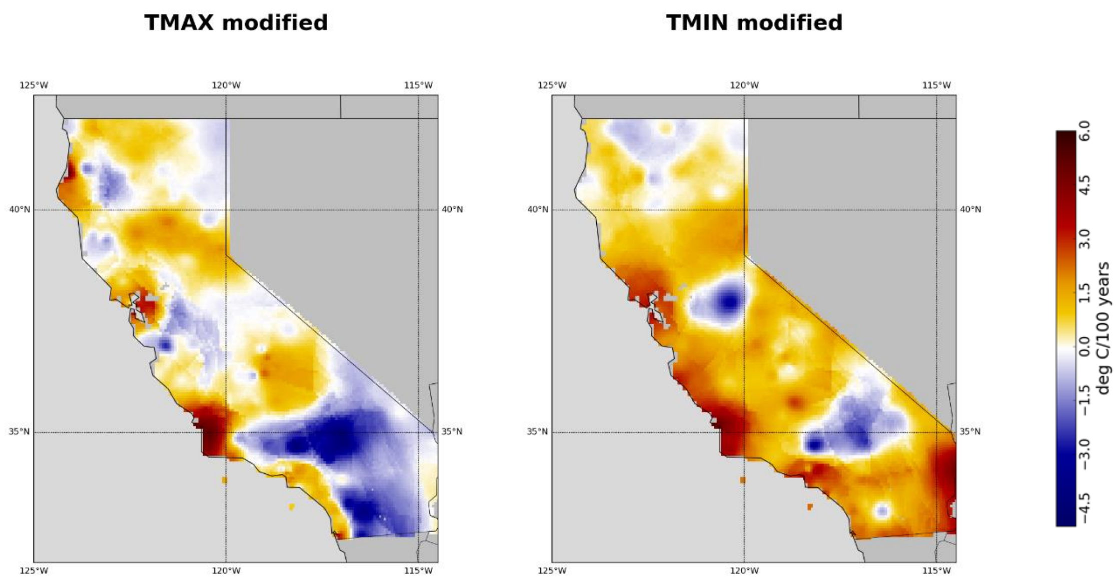


Figure A5: Grid cell temperature trends in degrees C per century for the modified SWM set. The spatial pattern and magnitude of trends generally is similar to that shown in Figure S2 in the paper.

Table A1: Selected statistics with temperature trends shown in degrees Celsius per century. These statistics are for the modified SWM dataset and can be compared to the other gridded products' trend statistics shown in Table S3. Overall, the modified SWM set has higher maximum trends and lower minimum trends. Changes in maximum trends are more noticeable though variation is similar for both sets.

		10% Quantile	90% Quantile	Mean	Median	Standard Deviation
SWM*	Tmax	-1.62	1.45	-0.00101	0.439	1.31
	Tmin	-0.460	2.42	0.973	0.993	1.19

A.5 Linear Regression

We compare the performance of least-squares regression to the Theil-Sen estimator for calculating the trend of the state-average Tmin and Tmax. The results are similar, but least-squares regression in all six cases for Tmin leads to slightly higher trends. The largest difference is a 0.10°C/century increase in BEST. For Tmax, least-squares regression leads to higher trends for all gridded datasets, but not for the HCN network. The largest difference is a 0.07°C/century increase in SWM. Trend lines from both methods, least-squares regression as well as Theil-Sens estimators, are shown in Figure S6. The red lines represent the results of least-squares regression while Theil-Sens results are in gray. Table S2 shows the least-squares regression results for Tmin and Tmax for each dataset.

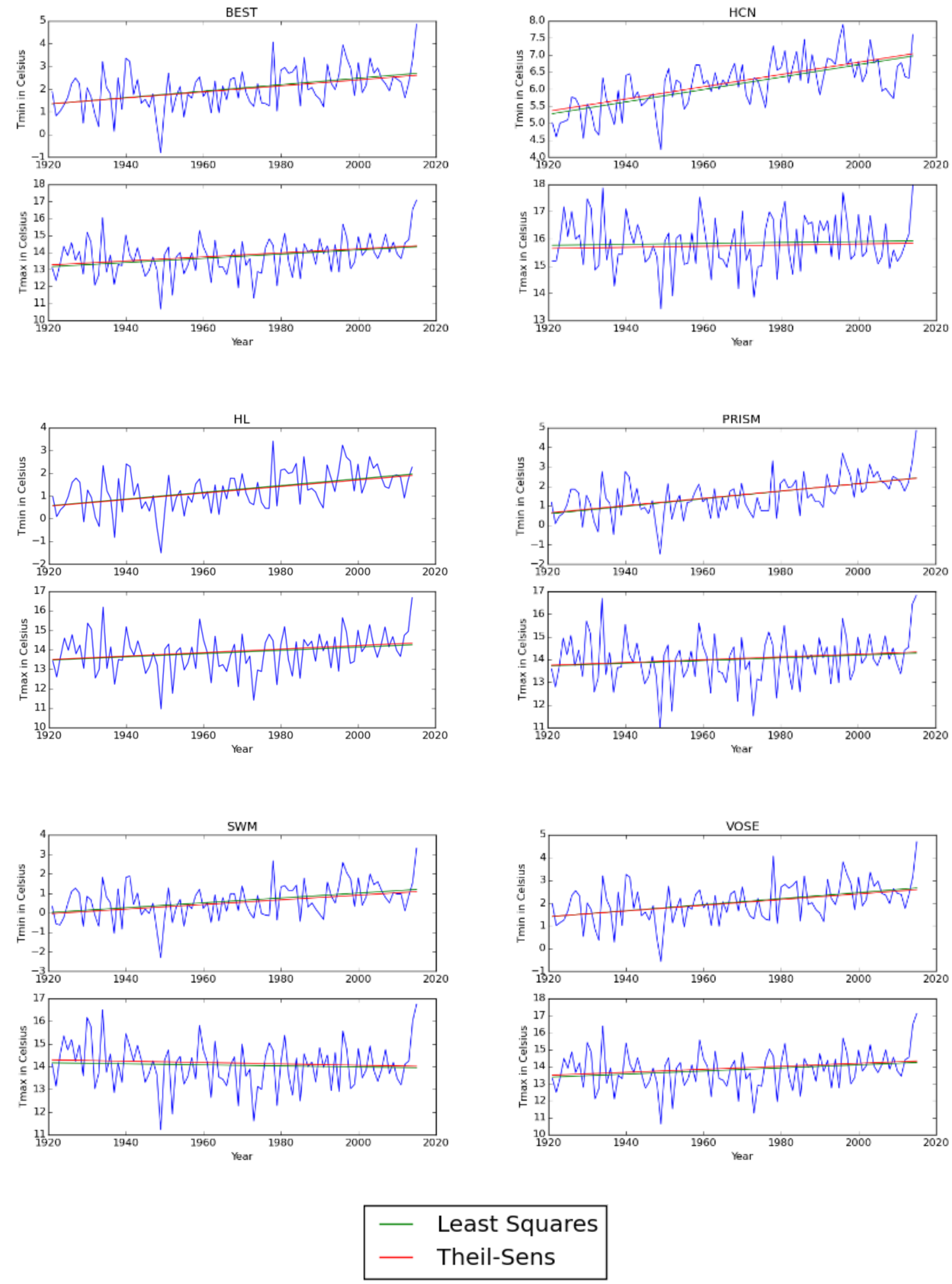


Figure A6: Comparison of least squares linear regression versus Theil-Sens estimators.

Table A2: Inferred statewide trends (in degrees Celsius per century) for minimum and maximum winter temperatures via least-squares regression as compared with Theil-Sen estimates

Dataset	Tmin (LSE)	Tmin (Theil-Sen)	Tmax (LSE)	Tmax (Theil-Sen)
BEST	1.43	1.25	1.24	0.87
H&L	1.48	1.44	0.83	0.90
HCN	1.82	1.80	0.17	0.21
PRISM	1.94	1.88	0.59	0.62
SWM	1.25	1.20	-0.23	-0.30
VOSE	1.33	1.33	0.91	1.19

a) Grid Cell Trend Statistics

We qualitatively reference some grid cell level trend statistics in the main text. Here we report more details of further comparisons among the datasets. As shown in Table S3, based on standard deviations, BEST has the smallest spread for both Tmin and Tmax, while SWM has the largest spread. SWM is the only dataset to have a negative Tmax mean and median, though Tmax HL, PRISM, and SWM all have small negative trends at the tenth percentile. All descriptor statistics for Tmin trends are positive. Further, Tmin trends all have a mean and median greater than 1.0 °C/century and are larger than the mean and median Tmax.

Table A3: Table of temperature trend (degrees Celsius per century) statistics for trends calculated at the grid cell level, which are shown in Figure 3 in the main text. Minimum temperature trends are consistently larger than maximum temperature trends within the same dataset. BEST has the smallest standard deviation while SWM has the largest.

		10% Quantile	90% Quantile	Mean	Median	Standard Deviation
BEST	Tmax	0.74	0.99	0.86	0.85	0.10
	Tmin	0.92	1.34	1.12	1.11	0.16
HL	Tmax	-0.09	1.84	0.86	0.86	0.71
	Tmin	0.26	3.00	1.42	1.23	1.06
PRISM	Tmax	-0.08	1.26	0.58	0.56	0.55
	Tmin	0.89	3.02	1.86	1.78	0.85
SWM	Tmax	-1.87	1.21	-0.24	-0.24	1.30
	Tmin	-0.36	2.68	1.17	1.21	1.21
VOSE	Tmax	0.36	1.91	1.22	1.28	0.64
	Tmin	0.44	2.11	1.32	1.34	0.66

b) Pattern correlation at grid cell level

We also considered pattern correlations between the winter temperature trend maps in a pair-wise fashion. To that end, we used the Pearson product-moment coefficient of linear correlation (equation S-1). The coefficient, R, measures correlation between two variables at the same location but on different maps – which in our case are two different datasets. One dataset has value x at grid cell index, i, while the second dataset has value y at the same location, i. The grid cell index ranges from 1 to n=10703, which is the number of cells representing California at our 1/16th degree gridded resolution. Because the variables must represent the same location but our datasets have different spatial resolutions, we first resampled using the nearest neighbor technique to the coarser 1/16th degree resolution. The resulting correlation coefficients for each pair of datasets are shown in Table S4. For Tmax and Tmin, the highest correlation is between HL and VOSE, which have fairly similar spatial patterns and textures. BEST and PRISM, which contrast perhaps the most in smoothness have a near zero correlation for Tmax and a slight negative correlation for Tmin. Overall, SWM has similar texture to HL and VOSE but a differing spatial trend, particularly for Tmax where SWM is unique in its large regions of cooling trends, which lead to negative correlations with most other datasets where the cooling trends are not present. For Tmin, SWM is most strongly correlated with VOSE; the two maps for Tmin trends have similar texture and general spatial pattern but again SWM has more regions of cooling trends that are not seen in VOSE and thus lower correlation. While HL and SWM have similar texture and both show cooling trends, the regions where cooling trends occur are different, leading to low correlation between the two trend maps.

We applied the same analysis to the winter climatology (mean fields) for each pair of datasets. Not surprisingly, the pattern correlations are much higher (between 0.92 and 0.99). This confirms that while the patterns of the climatology spatial fields are quite similar for the different datasets, the trend fields differ substantially.

$$R = \frac{\sum_{i=1}^{n}((x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad \text{Eqn. A1}$$

Table A4: Table of pattern correlation coefficients, *R*, between each pair of datasets for Tmax (top) and Tmin (bottom) winter temperature trends.

Tmax	**SWM**	**HL**	**PRISM**	**VOSE**
BEST	0.11	0.13	0.028	0.24
SWM	---	-0.21	-0.12	-0.19
HL	---	---	0.13	0.66
PRISM	---	---	---	0.16

Tmin	**SWM**	**HL**	**PRISM**	**VOSE**
BEST	0.056	-0.24	-0.12	0.045
SWM	---	0.038	0.15	0.33
HL	---	---	0.43	0.65
PRISM	---	---	---	0.36

A.7 Snow Water Equivalent (SWE) Analyses

We performed model simulations in which we changed only the temperature inputs to assess the effect of temperature on modeled SWE as reported in Section 3.3 of the main text. We used the Variable Infiltration Capacity (VIC) macroscale hydrologic model, version 4.0 (Liang et al., 1994), with the full energy balance option (as described in Liang et al., 1999) enabled. VIC inputs include daily minimum and maximum temperatures, wind speed, and precipitation. We used the SWM dataset to provide wind speed and precipitation data. We resampled all datasets to match the SWM 1/16th degree resolution as needed using the nearest neighbor approach. Since the datasets for BEST, PRISM, and VOSE are in a monthly timestep, we also needed to disaggregate to the daily timestep. We did so by adjusting the daily SWM data such that the monthly average would match the BEST, PRISM, or VOSE monthly average. Thus the daily anomalies are based on SWM, but the monthly averages still match the original dataset.

We ran VIC from 1920 through 2015 (2014 for HL) using the temperature data from all five sets while maintaining the wind speed and precipitation data as well as model parameters as constants. The result provides modeled SWE corresponding to each temperature dataset. We show the resulting time series of modeled April 1 SWE for each dataset in Figure S7. The Theil-Sen estimates of trend in cubic kilometers of SWE per century are shown in Table S5. Using the trend estimates, we de-trended the SWE time series to reflect 2015 conditions. The resulting cumulative distribution functions are shown in Figure S8.

The time series in Figure S7 show that SWM always has the highest SWE while the other temperature datasets result in similar, but considerably lower SWE accumulations. This is especially apparent in the cumulative distribution functions in Figure S8. The SWM CDF stands apart with higher snow accumulations and notably has the lowest statewide Tmin of all the datasets. It also has median Apr 1 SWE of around 20 km^3, which is roughly similar to estimates from other methods (e.g. Margulis et al., 2016). SWE for the other four datasets are relatively similar (but much lower), particularly BEST and VOSE, which also have similar statewide Tmin (see Figure 1 in the main text). HL and PRISM have slightly higher accumulations than BEST and VOSE, but are still distinctly below SWM accumulations. The differences in distribution between SWM and the other datasets are clear across all percentiles and nearly consistent; the gap between SWM and other dataset snow accumulations increases as the percentile increases.

The trends in SWE, as estimated using the Theil-Sen estimator, are all decreasing. HL is decreases the least (-5.0 km^3/century), while PRISM has the highest rate of loss (-7.6 km^3/century) followed closely by SWM (-7.1 km^3/century). However, the unrealistically low SWE estimates for most of the datasets could be acting as a constraint on the trend estimates since SWE estimates at a minimum can be zero. The overall low SWE estimates would decrease the range between high years and low year and fail to capture an accurate loss of SWE over time.

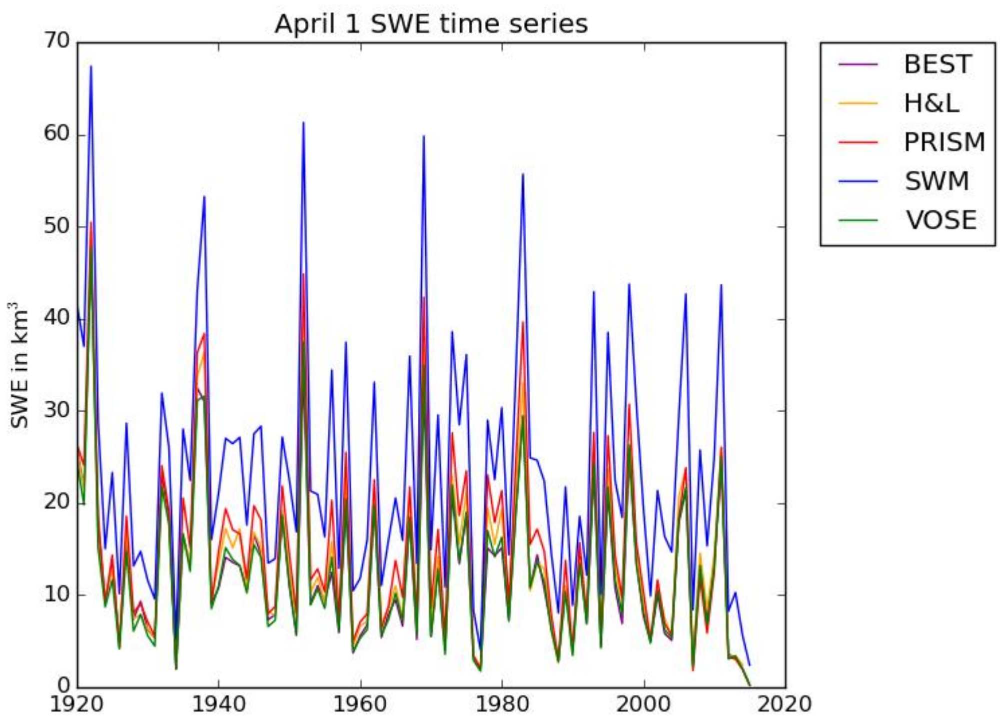


Figure A7: Time series of modeled April 1 SWE for five datasets. BEST, PRISM, and VOSE were resampled to match 1/16th degree resolution and disaggregated to a daily time step. Wind and precipitation conditions are the same across all model runs with only temperature (max and min) changing.

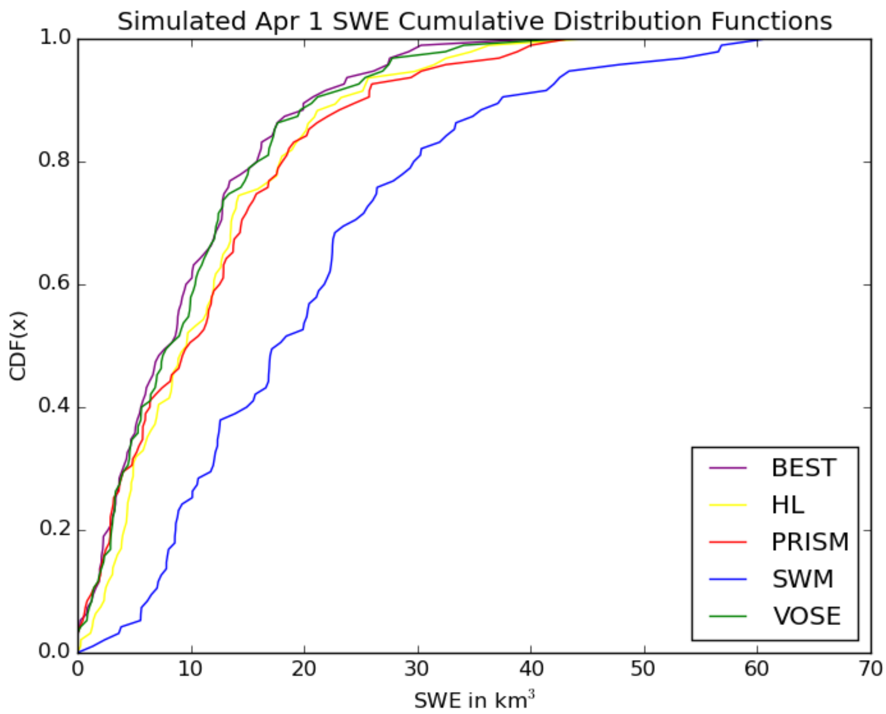


Figure A8: The cumulative distribution functions for modeled April 1 SWE for all five datasets that have been detrended using their respective Theil-Sen trend estimate to reflect current (2015) conditions.

Table A5: Theil-Sen estimates of trend in modeled April 1 SWE shown per dataset in cubic kilometers per century. All datasets show negative SWE, indicating loss of SWE with time.

Dataset	**SWE Trend in km^3/century**
BEST	-6.32
HL	-5.04
PRISM	-7.64
SWM	-7.07
VOSE	-5.36

References

Casola, J.H., L. Cuo, B. Livneh, D.P. Lettenmaier, M. Stoelinga, P.W. Mote and J. M. Wallace (2009), Assessing the Impacts of Global Warming on Snowpack in the Washington Cascades, *Journal of Climate*, 22(10), 2758-2772.

Cooper, Matthew G., Anne W. Nolin, and Mohammed Safeeq (2016), Testing the recent snow drought as an analog for climate warming sensitivity of Cascades snowpacks, *Environmental Research Letters*, 11(8), doi: 10.1088/1748-9326/11/8/084009.

Di Luzio, et al. (2008), Constructing restrospective gridded daily precipitation and temperature datasets for the conterminous United States, *Journal of Applied Meteorology and Climatology*, 47(2), 475-497.

Hamlet A.F., Lettenmaier D.P. (2005, Production of temporally consistent gridded precipitation and temperature fields for the continental U.S, *Journal of Hydrometeorology*, 6(3), 330-336.

Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J. Burges (1994), A Simple hydrologically Based Model of Land Surface Water and Energy Fluxes for GSMs, Journal of Geophysical Research, 99(D7), 14,415-14,428.

Liang, X., E. F. Wood, and D. P. Lettenmaier (1999), Modeling ground heat flux in land surface parameterization schemes, Journal of Geophysical Research, 104(D8), 9581-9600.

Livneh B, E.A. Rosenberg, C. Lin, B. Nijssen, V. Mishra, K. Andreadis, E.P. Maurer, and D.P. Lettenmaier (2013), A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States: Updates and extensions, *Journal of Climate*, 26(23) 9384-9392.

Mao, Yixin, Bart Nijssen, and Dennis P. Lettenmaier (2015), Is climate change implicated in the 2013-2014 California dought? A hydrologic perspective, *Geophysical Research Letters,* 42(8), 2805–2813.

Margulis, Steve and Gonzalos Cortes, Manuela Girotto, and Michael Durand (2016), A Landsat-Era Sierra Nevada Snow Reanalysis (1985-2015), *American Meteorological Society*, 17, 1203-1221.

Menne, Matthew J, Claude N Williams, Russel S Vose (2009), The US historical climatology network monthly temperature data, version 2, *Bulletin of the American Meteorological Society* 90(7), 993-1007.

Oyler, J. W., A. Ballantyne, K. Jencso, M. Sweet, and S. W. Running (2015), Creating a topoclimatic daily air temperature dataset for the conterminous United States using homogenized station data and remotely sensed land skin temperature, *International Journal of Climatology*, 35(9), 2258-2279, doi:10.1002/joc.4127.

Rohde, Robert, R Muller, R Jacobsen, S Perlmutter, A Rosenfeld, et al. (2013), Berkeley Earth Temperature Averaging Process, *Geoinformatics and Geostatics*, 1(2).

Vose, RS, et al. (2014), Improved historical temperature and precipitation time series for the US climate divisions, *Journal of Applied Meteorology and Climatology*, 53(5) 1232-1251.

Williams, A. P., R. Seager, J. T. Abatzoglou, B. I. Cook, J. E. Smerdon, and E. R. Cook (2015), Contribution of anthropogenic warming to California drought during 2012-2014, *Geophysical Research Letters*, 42(16), 6819-6828, doi:10.1002/2015GL064924.

APPENDIX B

The bias-variance tradeoff applied to nonstationary modeling of extreme precipitation evaluated using Monte Carlo simulations – Supplementary Materials

Appendix B provides supporting information for Chapter 3 and will be submitted to the *Journal of Hydrology* as the supplemental material of the following article.

Wang, K., K. McKinnon, D. P. Lettenmaier (2022), The bias-variance trade-off applied to nonstationary modeling of extreme precipitation evaluated using Monte Carlo simulations. *Journal of Hydrology* (in prep).

Introduction

The supplemental material for this paper is organized into three main sections, each meant to elaborate on a concept introduced in the main paper but that was truncated for brevity. We include here additional analyses that mirror the main paper and go more in-depth as to how record length and nonstationary scenarios affect results.

B.1 Histograms of Storm Estimates

a) Alternative Record Lengths

We show here the histograms analogous to Figure 3.4 but with respect to Monte Carlo simulations set up with synthetic time series of 30 years, Figure B1, and 100 years, Figure B2. The pattern discerned for record length of 50 years is seen here as well and is more discernable for n = 100 years and somewhat less distinct for n = 30 years. This is due to the fit of the distribution to data with more data leading to more information and a better fit whereas with less data, the overall variance clouds the signal of the pattern discussed.

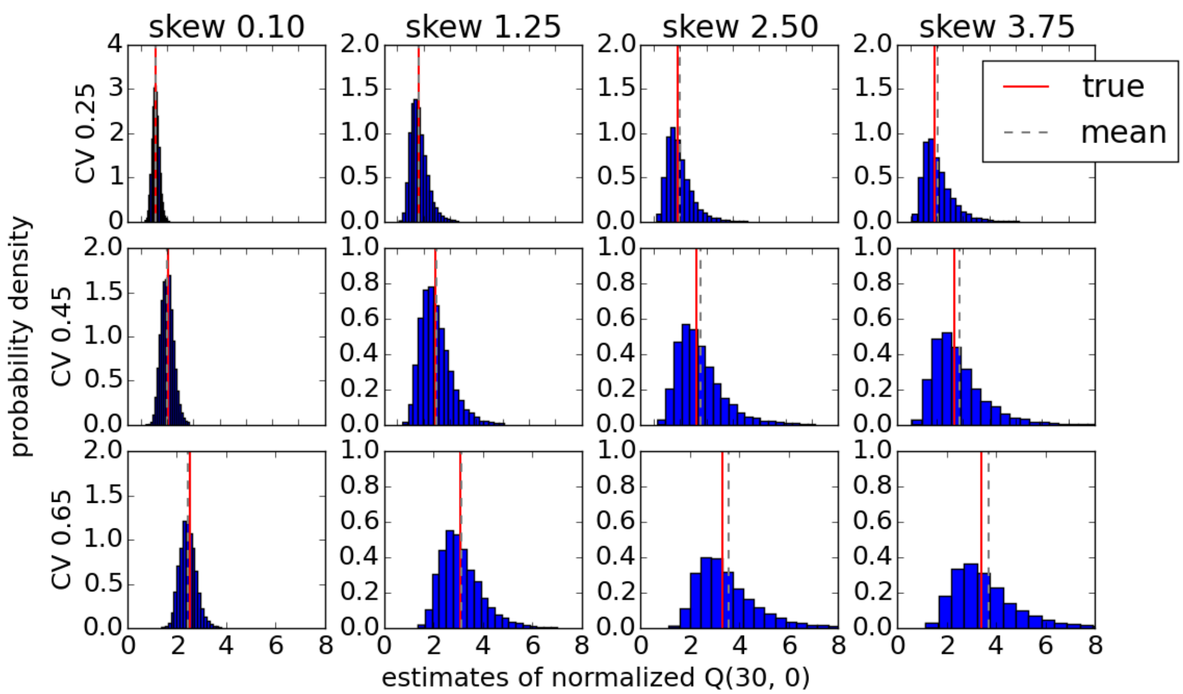


Figure B1: Histograms of the estimated Q_{100} across the parameter space at the beginning of the project length when record length is 30 years. The mean of each probability density function is indicated by a dashed grey line while the true storm event is shown by a red line.

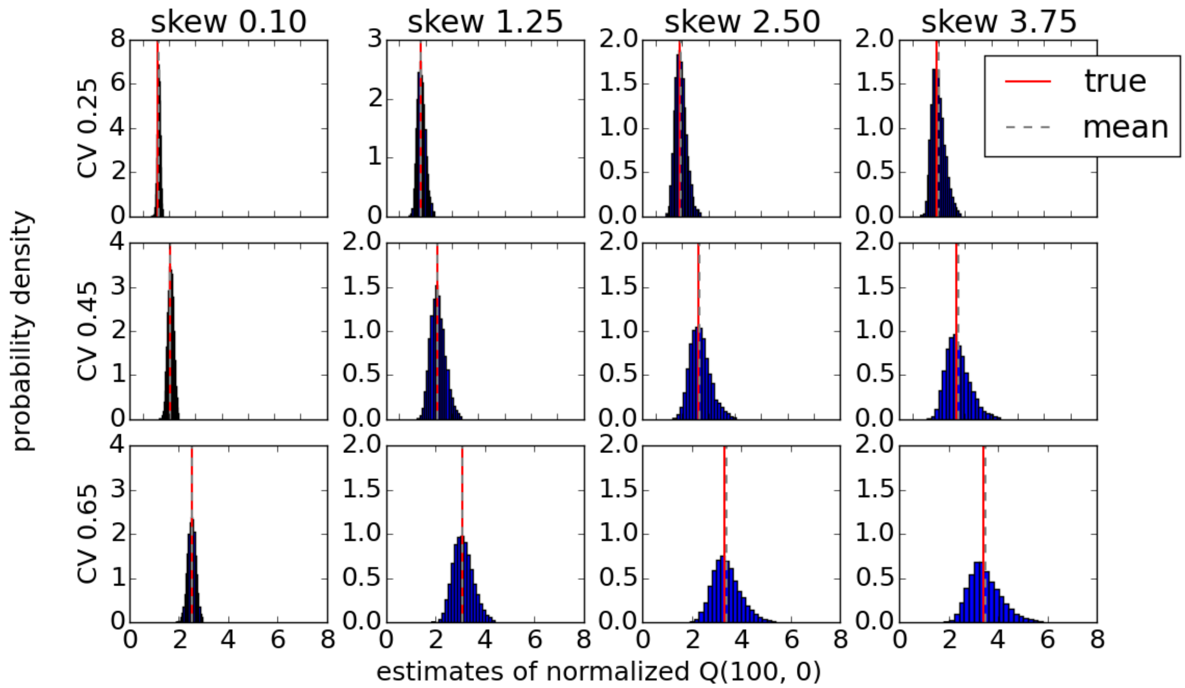


Figure B2: Histograms of the estimated Q_{100} across the parameter space at the beginning of the project length when record length is 100 years. The mean of each probability density function is indicated by a dashed grey line while the true storm event is shown by a red line.

b) End of Project Lifespan

The histogram of storm estimates at the end of the project lifespan (N = 50 years) is not significantly different from the beginning of the project lifespan, indicating that while the NS-GEV model is being used, the estimated time-dependent parameters are fairly small and do not contribute to a large difference over a fifty-year timespan.

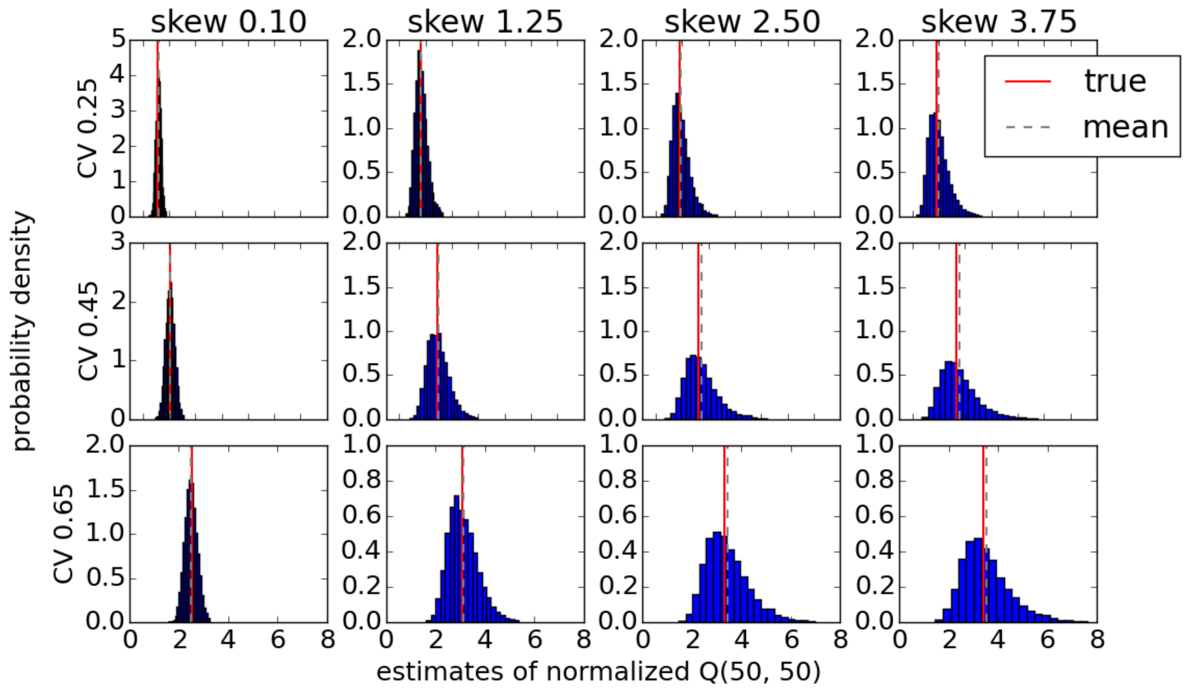


Figure B3: Histograms of the estimated Q_{100} across the parameter space at the end of the project length when record length is 50 years. The mean of each probability density function is indicated by a dashed grey line while the true storm event is shown by a red line.

B.2 Variance Fraction

Whether the estimated storm under- or over-estimates the true storm, there is a pattern to the partitioning of bias and variance of error where the bias fraction approaches zero as the true and estimated storm meet. This holds true whether the environment is stationary or nonstationary as long as it is a mis-fitted situation. Before this time, bias fraction is decreasing to zero; after this time, bias fraction is increasing from zero. The year in which zero bias is achieved varies with data CV and skewness, record length, and scenario condition being used. This year is at times within the observation period, typical when the true environment is stationary, or it could be in the future or projected time. There are conditions where this zero bias point is not achieved within the project lifespan but will occur at a farther point in time. We include the plots analogous to Figure 3.5 for alternative record lengths and for when the true environment is stationary (Figure B4).

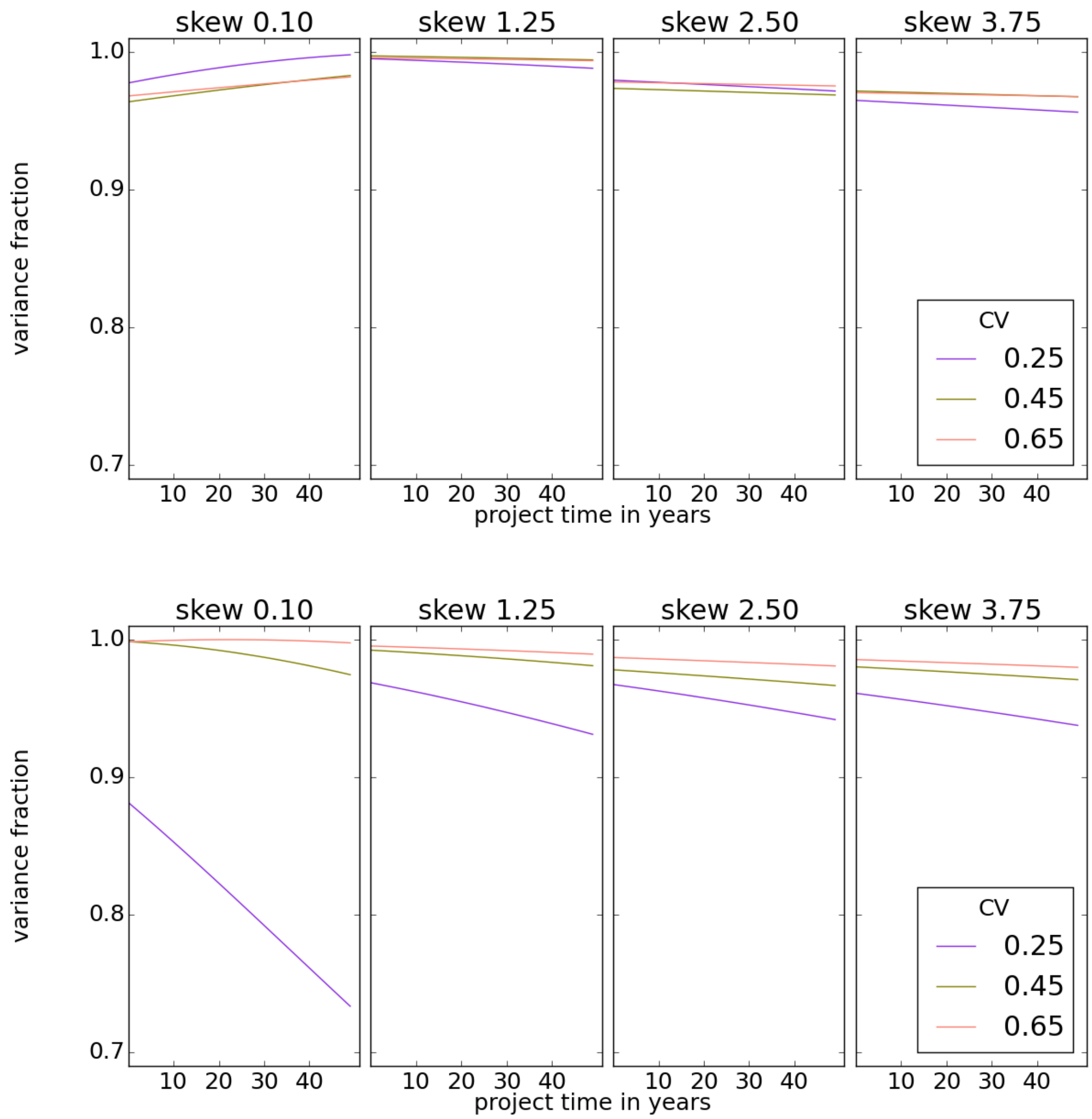


Figure B4: Variance fraction at different values of γ and CV as project time (future period) increases for synthetic data with a record length of (top) n = 30 and (bottom) n = 100.

Further, we provide figures analogous to Figure 3.7 showing variance fraction when the true environment is nonstationary in Figure B5 and B6 for Scenario II and III (respectively).

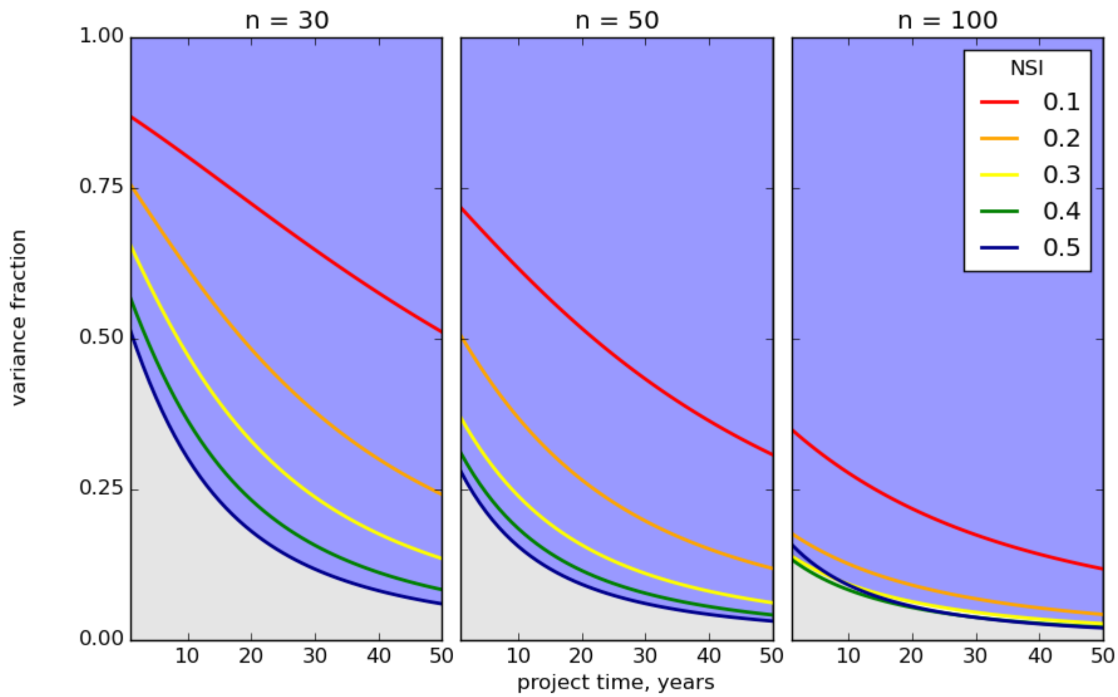


Figure B5: Variance fraction under Scenario II conditions for the range of NSIs for the distribution that begins with a CV of 0.25 and γ of 0.10. Blue shading indicates the largest possible proportion of error attributed to bias across NSI values. The grey shading accordingly gives the complementing proportion of error attributed to variance.

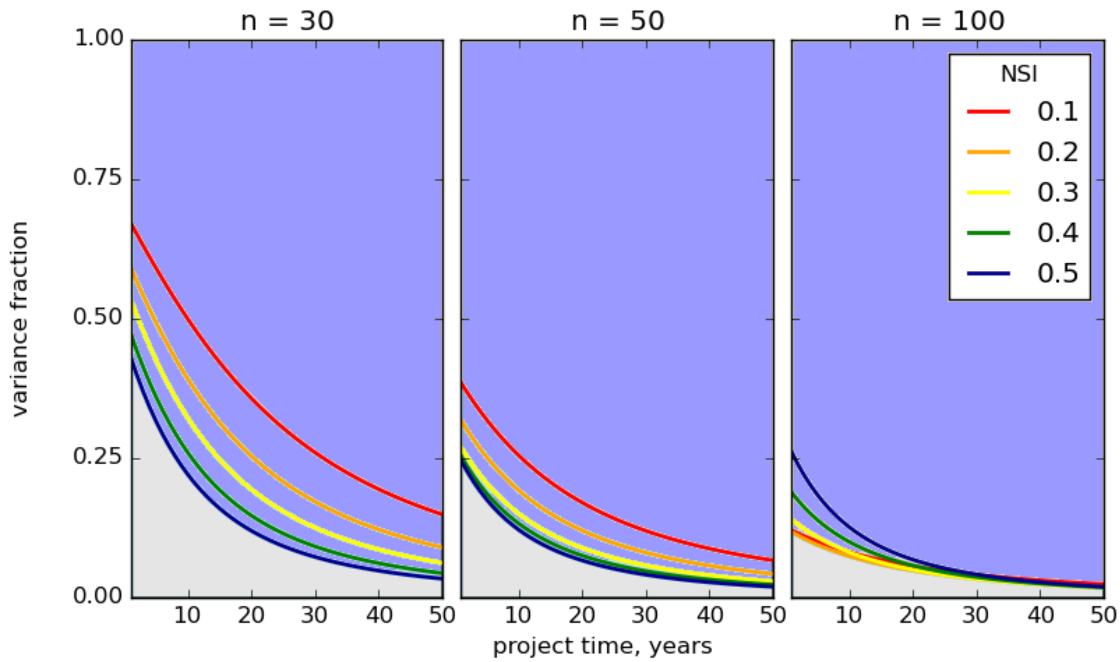
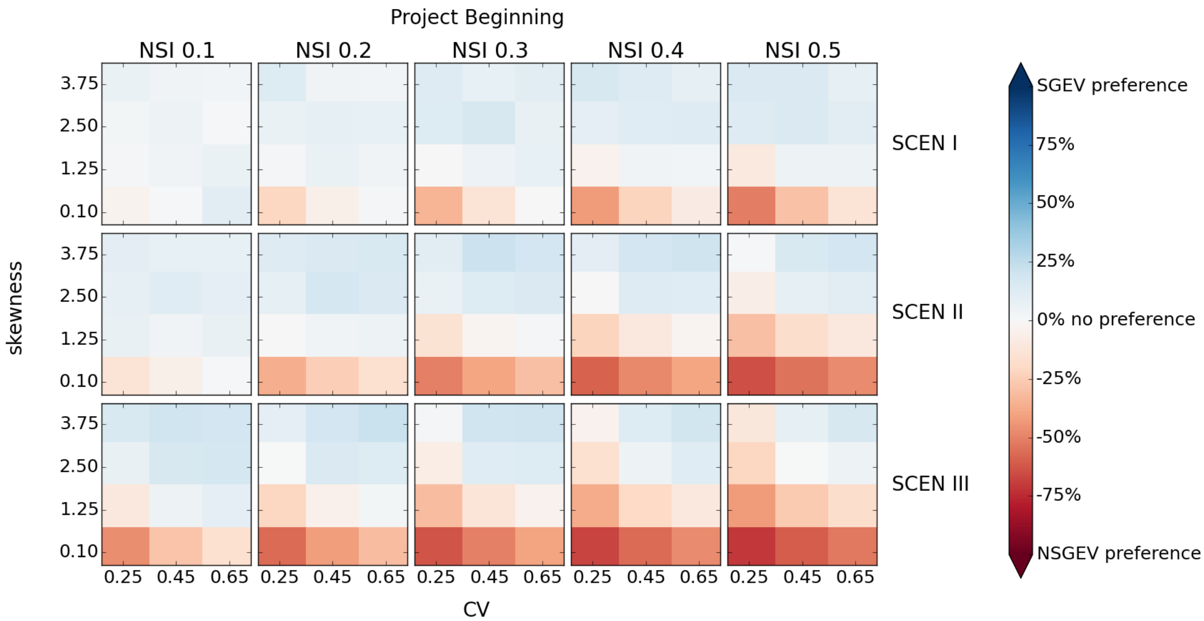


Figure B6: Variance fraction under Scenario III conditions for the range of NSIs for the distribution that begins with a CV of 0.25 and γ of 0.10. Blue shading indicates the largest possible proportion of error attributed to bias across NSI values. The grey shading accordingly gives the complementing proportion of error attributed to variance.

B.3 Preference Grids

Here we show preference grids analogous to Figure 3.8 for alternative record lengths considered.



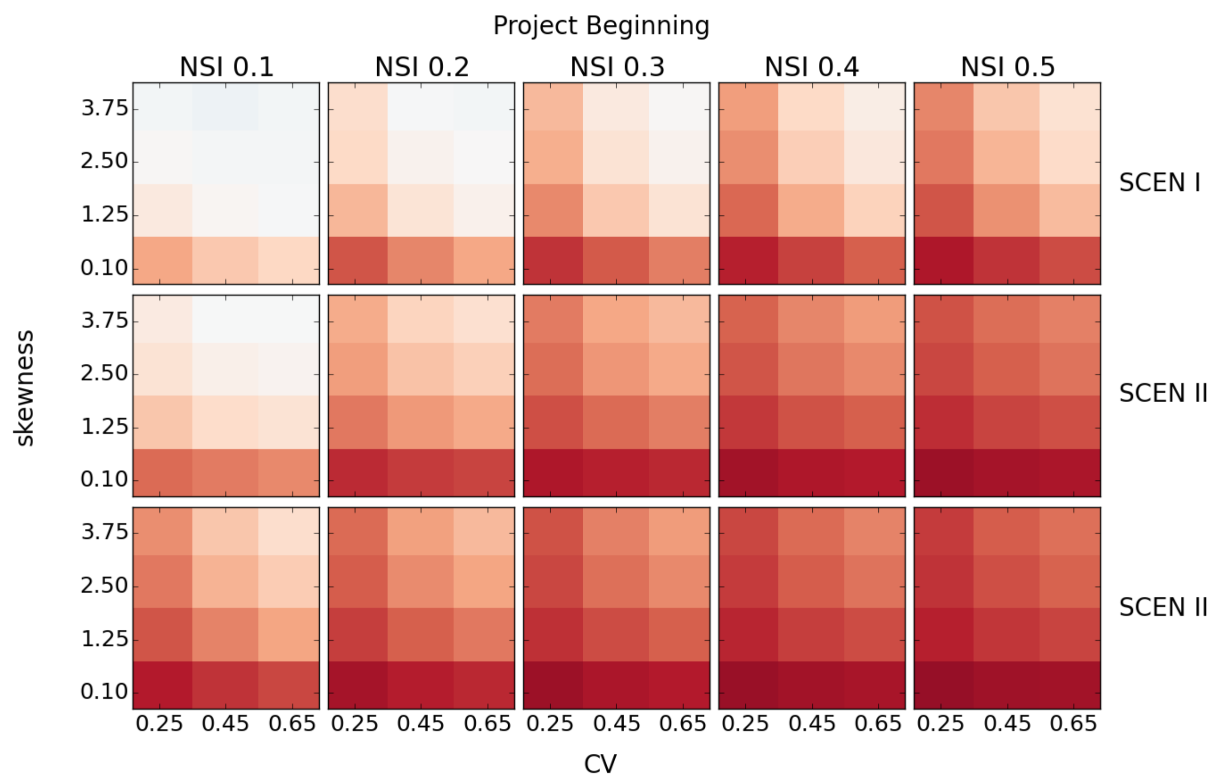


Figure 3.8: Preference grids across the γ -CV space based on a (top) 30-year record length and (bottom) 100-year record length at the end of the project. Darker blue indicates stronger

preference for S-GEV while darker red indicates stronger preference for NS-GEV. The preference is based on the percent change as expressed in Eq. 8 where a negative percent change means that $nRMSE_{NS}$ (error from applying the NS-GEV) is smaller than $nRMSE_{S}$, and therefore application of the NS-GEV is preferred. Positive percentages indicate the $nRMSE_{NS}$ is larger than $nRMSE_{S}$, and application of the S-GEV is preferred.