

UNIVERSITY OF OKLAHOMA
GRADUATE COLLEGE

DEVELOPING A FRAMEWORK FOR SEAMLESS PREDICTION OF
SUBSEASONAL TO SEASONAL EXTREME PRECIPITATION EVENTS IN THE
UNITED STATES

A DISSERTATION
SUBMITTED TO THE GRADUATE FACULTY
in partial fulfillment of the requirements for the
degree of
DOCTOR OF PHILOSOPHY

By
TY A. DICKINSON
Norman, Oklahoma
2023

DEVELOPING A FRAMEWORK FOR SEAMLESS PREDICTION OF
SUBSEASONAL TO SEASONAL EXTREME PRECIPITATION EVENTS IN THE
UNITED STATES

A DISSERTATION APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Jason C. Furtado (Chair)

Dr. Michael B. Richman (Co-Chair)

Dr. Renee A. McPherson

Dr. Elinor M. Martin

Acknowledgments

Oh, what a journey it has been. I have been incredibly fortunate to be around so many amazing people who made grad school survivable. First and foremost, I would like to thank my co-advisors Drs. Jason Furtado and Michael Richman. Their guidance throughout my time was vital to my success and I could not have asked for better advisors, especially throughout a pandemic. I am also thankful for my other two committee members, Drs. Elinor Martin and Renee McPherson, for their expertise throughout the project and for always reminding me to keep the big picture in mind.

I moved to Oklahoma knowing basically nobody, but I formed a second family down here and I would not have made it without the help of countless friends. Working alongside Melanie Schroers throughout this project was so, so rewarding and I could not imagine a better research partner. I also owe so much to Kayla Wheeler, who I met and began dating during grad school. She helps push me to the best person I can be and I truly cannot imagine finishing school without her. Perhaps the beginning of my second family was meeting Ryan Bunker, who I cannot thank enough for helping me in virtually every aspect throughout both undergrad and grad school.

I would also like to shout out many other PRES2ⁱP team members: Greg Jennrich, Devin McAfee, Olivia VanBuskirk, Paulina Cwik, and Derek Rosendahl; y'all have been the best team. Lastly, I have been so lucky to have so many fun and kind people to share an office space with; Jacob Ohnstad, Matt Rogers, and Katie Giannakopoulos, thank you so much for everything.

Finally, the biggest thank you goes out to my family. I have been blessed with the best family and support system whom helped make all my dreams come true. I owe so much of my success to my parents and I can only hope to shine on others in the future in the same light that they have shined on me.

This work has been supported by the National Science Foundation's Resilience Against Extreme Events, Grant #1663840. I am eternally grateful for my time at OU in the School of Meteorology.

Table of Contents

Acknowledgments	iv
List Of Tables	vii
List Of Figures	viii
Abstract	xiii
1 Introduction	1
2 Subseasonal-to-Seasonal Extreme Precipitation Periods in the Contiguous United States: Generating an Observationally-Based Database	5
2.1 Introduction	5
2.2 Extreme period identification	6
2.2.1 Data	6
2.2.2 Extreme criteria	8
2.2.3 Finding an extreme region	12
2.3 Cataloging extreme periods	15
2.4 Statistical characteristics of 14-day extreme periods	18
2.4.1 Extreme period regionalization	18
2.4.2 Subseasonal extreme period frequency	24
2.5 Summary and conclusions	30
3 S2S Extreme Precipitation Periods: Extending Databases to Reanalysis and Model Datasets	33
3.1 Introduction	33
3.2 Data	34
3.3 Generating new databases	34
3.4 Climatological comparison	38
3.4.1 Criteria thresholds	38
3.4.2 Extreme period frequency	41
3.5 Summary and conclusions	47
4 Synoptic- and Hemispheric-Scale Characteristics of Subseasonal Extreme Precipitation Periods in the United States	50
4.1 Introduction	50
4.2 Data	52
4.3 Results	52
4.4 Summary and conclusions	60

5	Predicting U.S. Subseasonal Extreme Precipitation Periods Part I: Random Forests	62
5.1	Introduction	62
5.2	Data and methods	65
5.2.1	Data	65
5.2.2	Decision trees and random forests	66
5.2.3	Model validation	69
5.3	Results	71
5.3.1	Model skill	71
5.3.2	Variable importance	73
5.4	Summary and conclusions	78
6	Predicting U.S. Subseasonal Extreme Precipitation Periods Part II: Convolutional Neural Networks	81
6.1	Introduction	81
6.2	Data and Methods	83
6.2.1	Data	83
6.2.2	Convolutional neural network	86
6.2.3	Object-oriented verification	91
6.3	Results	93
6.3.1	Summary statistics	93
6.4	Case studies	101
6.4.1	May 2013 pluvial	101
6.4.2	February 2018	102
6.4.3	September 2008: Hurricanes Gustav and Ike	107
6.4.4	Realtime forecast: January 2023	108
6.5	Summary and conclusions	111
7	Broader Impacts: Incorporating Stakeholder Feedback	114
7.1	Introduction	114
7.2	Methods	115
7.2.1	Workshop 2	115
7.2.2	Workshop 3	115
7.3	Results	117
7.3.1	Fact sheets	117
7.3.2	Forecast uncertainty	118
7.3.3	Object-oriented verification	120
7.4	Conclusions	122
8	Summary and Conclusions	123
	Reference List	130
	Appendix	142
1	Appendix A	142
2	Appendix B	146

List Of Tables

2.1	A sample of events from version 1.0 of the 14-day extreme precipitation period database. $Precip_{AA}$ denotes areal-averaged precipitation in the region; Max_{total} is the greatest 14-day precipitation total from any point inside the region; Max_{1day} is the greatest single-day precipitation total from any point inside the region from any of the 14 days. Area and TOE are rounded to the nearest square kilometer and millimeter, respectively, and all other entries are rounded to 2 decimal places.	18
3.1	Values developed for each dataset during the database generation process. . .	35
5.1	Optimized hyperparameter configurations for each of the fit random forests. . .	69
5.2	2x2 contingency table used to assess skill in binary forecasts.	69
5.3	Variables used to train the cluster-specific random forests.	71
5.4	Performance for cluster-specific random forests within the training datasets. CSI and SEDS have been rounded to 3 significant figures.	72
5.5	As in Table 5.4 but over the testing datasets.	73
5.6	As in Table 5.5 but for a relaxed scheme. The change in scores relative to the strict verification scheme are added as the final two columns.	74
6.1	Variables used to train our U-Nets. n/a indicates a 2-dimensional variable. . .	84
A.1	Standardized Anomalies for Central West Coast.	147
A.2	Detrended Anomalies for Central West Coast.	147
A.3	Standardized Anomalies for Central Plains.	148
A.4	Detrended Anomalies for Central Plains.	148
A.5	Standardized Anomalies for Ohio River Valley.	149
A.6	Detrended Anomalies for Ohio River Valley.	149

List Of Figures

1.1	Figure taken from (White et al. 2017, Fig. 1). (a) Qualitative estimates of skill from short-range to long-range forecasts. (b) Schematic outlining the relationships among different forecast types and uses in the decision-making process.	3
2.1	Total number of stations used in processing monthly precipitation data from PRISM as a function of time.	7
2.2	(black) Raw 99 th percentile for each 14-day window and (green) smoothed percentiles using the first three Fourier harmonics for six selected grid points across the CONUS. The plotted date is the start date of each window.	9
2.3	The 99 th percentile of precipitation in the CONUS for one 14-day window in each meteorological season.	10
2.4	(a) Total recorded precipitation (mm) between 12 and 25 Jan 1937. (b) The 99 th percentile of precipitation (mm) derived for 12-25 Jan. (c) Number of days with above normal daily precipitation during 12-25 Jan 1937. Black contour outlines 7 days, which is the threshold for the second criterion in event selection (see text). (d) Points flagged as extreme per the algorithm (see text for details).	11
2.5	Flowchart describing the logic in determining if any given point is labeled extreme. See text for details.	13
2.6	(a) Full KDE normalized density field using the Epanechnikov kernel and 0.02 bandwidth for the 12-25 Jan 1937 window. (b) Extreme period polygon drawn using the 0.2710 contour.	14
2.7	Flowchart describing the logic during postprocessing of extreme precipitation events. See text for details.	16
2.8	Various statistics as a function of number of clusters for 14-day extreme precipitation periods: (a) intracluster sum of squared distances (i.e., inertia), (b) average silhouette score, (c) total number of clustered events, and (d) product of (b) and (c), (i.e., the HI). The dashed red line in (d) marks the HI maximum.	21
2.9	(colored contours) Average polygon for $k = 15$ clusters across the CONUS for 14-day extreme precipitation periods. Also shown are the regional boundaries defined by Jennrich et al. (2020) (gray outlines).	22
2.10	(black) Total yearly counts of recorded 14-day extreme precipitation periods throughout the CONUS as a function of time. QR models are fit on (green) $\tau=0.1$, (red) $\tau=0.5$ (i.e., median), and (blue) $\tau=0.9$ quantiles of the time series. Slopes are listed as the second number in the legend, and the 95% confidence intervals on the slope are shown in brackets thereafter. Confidence intervals were calculated using BCa bootstrapping. Slopes and confidence interval bounds were rounded to 3 decimal places.	25

2.11	As in Fig. 2.10, but for the total number of grid points inside extreme periods as a function of time.	26
2.12	Bar plots comparing the (solid) observed frequencies and (hatched) expected frequencies of subseasonal extreme precipitation periods across the 3 ENSO phases. All extreme periods in the database are considered in (a), whereas panels (b)-(p) are for specific clusters. Bold titles indicate statistical significance at $p < 0.5$	27
2.13	Total number of times a grid point is within an extreme period polygon over 1915–2018.	28
2.14	Trend (number per year; shading) in yearly frequency of grid points falling inside 14-day extreme period polygons using ordinary linear regression. Black contour delineates statistical significance of the slope at $p < 0.05$ using bootstrapping.	29
3.1	Smoothed 14-day 99 th percentiles from the (green) Livneh dataset, (orange) ERA5 dataset, and (purple) CNRM-CM6-1-HR historical run using the first three Fourier harmonics for six selected grid points across the CONUS. The plotted date is the start date of each window.	38
3.2	Mean daily precipitation from the (green) Livneh dataset, (orange) ERA5 dataset, and (purple) CNRM-CM6-1-HR historical run. The plotted date is the start date of each 14-day window used to calculate the long-term mean.	39
3.3	The 99 th percentile of 14-day precipitation across the CONUS region from the (top row) Livneh dataset, (middle) ERA5 dataset, and (bottom) CNRM-CM6-1-HR historical run for one window in each meteorological season.	40
3.4	(black) Total yearly counts of recorded 14-day extreme precipitation periods domain-wide as a function of time in the (a) Livneh, (b) ERA5, and (c) CNRM-CM6-1-HR historical run 14-day databases. QR models are fit on (green) $\tau=0.1$, (red) $\tau=0.5$ (i.e., median), and (blue) $\tau=0.9$ quantiles of the time series where solid lines were fit on the entire time series and dashed lines were fit using 1950–2014. Slopes are listed as the second number in the legend, and the 90% and 95% confidence intervals on the slope are shown in brackets thereafter. Confidence intervals were calculated using BCa bootstrapping. Slopes and confidence interval bounds were rounded to 4 decimal places.	41
3.5	As in Fig. 3.4 but only for CONUS-clustered extreme periods.	43
3.6	Total number of times inside an extreme period polygon from the (a) Livneh (1915–2018), (b) ERA5, and (c) CNRM-CM6-1-HR historical run databases.	44
3.7	(shaded contours) Trend (number per decade) in yearly frequency of grid points falling inside extreme period polygons using ordinary linear regression in the (a) ERA5 and (b) CNRM-CM6-1-HR historical run 14-day databases. Black contour delineates statistical significance of the slope at $p < 0.05$ using bootstrapping with 5000 iterations.	45
3.8	As in Fig. 3.7 but performing ordinary linear regression in the Livneh and CNRM databases beginning in 1950.	46

4.1	(shading) 500 hPa standardized geopotential height anomaly composite from subseasonal extreme precipitation days in the Livneh database across various regions of the CONUS. Hatching denotes statistical significance at $p < 0.05$ after bootstrapping.	54
4.2	As in Fig. 4.1 but using CNRM simulated periods.	55
4.3	As in Fig. 4.1 but for 500 hPa vertical velocity.	56
4.4	As in Fig. 4.3 but using CNRM simulated periods.	57
4.5	Composite of (shading) 850 hPa specific humidity and (vector) wind standardized anomalies. Only significant regions of specific humidity and vectors are shown.	58
4.6	As in Fig. 4.5 but for 250 hPa.	59
5.1	Domain-total Gini impurity importances for the 7 predictors for the (black) CWC, (green) CP, and (red) ORV. Darker (lighter) shades are for the standardized (detrended) models.	75
5.2	(contours) Lag 0 500 hPa geopotential height standardized anomalies with contours beginning at ± 0.1 in increments of 0.1. (shading) Gini importance with lighter shades of red less important and darker shades representing higher importance. The 3 clusters are given in the columns while the top (bottom) row is a composite over model hits (false alarms).	76
5.3	As in Figure 5.2, but for a 7-day lag.	76
5.4	As in Figure 5.2, but for the zonal wind component.	77
5.5	As in Figure 5.2, but for the meridional wind component.	78
5.6	As in Figure 5.2, but for precipitable water.	78
6.1	Diagram of a multi-layer perceptron. Taken from https://medium.com/codex/introduction-to-how-an-multilayer-perceptron-works-but-without-complicated-math-a423979897ac	83
6.2	Frequency of subseasonal extreme precipitation from 1950-2018.	85
6.3	U-Net1 architectures used throughout the study. Dashed lines indicate optional steps depending on the experiment. Gating signals are only used for models that use attention gates. Numbers inside each feature map block indicate the number of latitudes, longitudes, and predictors/feature maps; in the decoder, the number of feature maps constitutes a sum of the maps coming from the encoder and the next-lowest decoder level.	86
6.4	Similarity functions for (a) centroid distance, (b) area ratio, and (c) aspect ratio difference between forecast-truth object pairs. Grey dashed lines indicate the e -folding distance.	91
6.5	Congruence coefficient boxplots for the different U-Net1 configurations for 2-week periods with an extreme somewhere in the CONUS (purple) and non-extreme periods (green). Nearest neighbor interpolation was used for models with “near” whereas models with “bilin” use bilinear interpolation. Notches represent a 95% confidence interval on the median calculated using bootstrapping with 5000 iterations.	94

6.6	Object-oriented performance diagram for various U-Net configurations (see legend) in the testing dataset. Forecast objects were created using the 50% probability contour. Error bars represent a 90% confidence interval and were calculated by bootstrapping with replacement using 5000 iterations.	96
6.7	(a) F_2 and (b) $F_{0.5}$ scores vs the area under a precision-recall curve for different U-Net configurations. Forecast objects were created using the 50% probability contour. Error bars represent a 90% confidence interval and were calculated by bootstrapping with replacement using 5000 iterations.	97
6.8	As in Fig. 6.6 but using the 60% contour.	99
6.9	As in Fig. 6.7 but using the 60% contour.	100
6.10	Histogram of probabilities across the CONUS in the testing set for each U-Net.	100
6.11	(a) Total precipitation (mm; from PRISM) and extreme period polygon (black outline), (b) total precipitation percentile, (c) liquid NCEI storm reports, and subseasonal extreme precipitation predictions from the (d) base-bilin, (e) DDC-conv-bilin, and (f) residual-bilin U-Nets for the 19 May - 1 June 2013 period.	102
6.12	(left) 500 hPa Geopotential height (contours) and standardized anomalies (shading), (middle) 850 hPa vector wind and (shading) specific humidity standardized anomalies, and (right) 500 hPa vertical velocity standardized anomalies on (top row) 20 May, (middle) 27 May, and (bottom) 31 May 2013.	103
6.13	As in Fig. 6.11 but for the 3-16 February 2018 period. The red polygon in (a) is an extreme period which starts 2 days later (see text). Both liquid and frozen storm reports are included in (c).	104
6.14	As in Fig. 6.12 but for (top) 5 February, (middle) 8 February, and (bottom) 15 February 2018.	105
6.15	As in Fig. 6.11 but for the 16 February - 1 March 2018 period.	105
6.16	As in Fig. 6.12 but for (top) 21 February, (middle) 25 February, and (bottom) 28 February 2018.	106
6.17	As in Fig. 6.11 but for the 1-14 September 2008 period. Note that total precipitation in (a) is from the Livneh dataset.	108
6.18	As in Fig. 6.12 but for (top) 3 September, (middle) 9 September, and (bottom) 13 September.	109
6.19	(a) Provisional PRISM total precipitation, (b) total precipitation percentile, and NCEI storm reports for (c) liquid and (d) frozen report types for the 1-14 January 2023 period.	110
6.20	Realtime forecasts initialized on 29 December 2022 from (a-c) ECMWF and (d-f) NCEP subseasonal models valid for 1-14 January 2023. The dynamical forecasts were input into the (a,c) base-bilin, (b,d) DDC-conv-bilin, and (c,e) residual-bilin U-Nets.	111
7.1	Fact sheet on defining a subseasonal extreme period.	116
7.2	Fact sheet on the 14-day extreme period climatology.	117
7.3	Schematic examples of hypothetical forecast-observation combinations. Taken from (Davis et al. 2019, Fig. 1)	120

7.4	Various pairings of forecasts (black) and truth (red). Colored dots represent the polygon centroid. Distances between centroids are about (a) 150 miles, (b) 220 miles, and (c) 350 miles.	121
7.5	As in Fig. 7.4 but for scenarios with similar centers but different spatial scales.	121
A.1	As in Fig. 2.2, but with using (dark blue) 2 harmonics, (green) 3 harmonics, (purple) 4 harmonics, (light blue) 5 harmonics, and (red) 6 harmonics. . . .	143
A.2	As in Figure A.1, but for 30-day events.	144
A.3	(a) Histogram of the area of every developed polygon in the database period of record. (b) Polygon area as a function of percentile from (a). Dashed red line in (a) and red dot in (b) correspond to our chosen 200,000 km ² threshold.	144
A.4	As in Fig. 2.9, but for different seed initializations. Panel (a) is the seed used to produce Figure 9.	145
A.5	As in Fig. 2.9, but using Ward's linkage.	146

Abstract

Extreme precipitation events are natural hazards that pose a significant risk to life and a commensurately large cost through property loss across multiple timescales. Unfortunately, extreme precipitation, especially extended-duration precipitation, remains as one of the most challenging hazards to forecast. The subseasonal to seasonal timescale, defined as the period between 2 weeks and 3 months, is characterized with low skill and a subsequent lack of informative forecast products for end-users. The present work constitutes a major step into the identification, characterization, prediction, and communication of subseasonal extreme precipitation. To identify an extreme precipitation period, thresholds for both total precipitation and the duration of the precipitation are used to identify events with sufficient length to accentuate the synoptic and subseasonal contributions. I then generate databases of 14-day extreme precipitation periods over the contiguous United States (CONUS) using observational datasets, reanalysis products, and climate model simulations, displaying the flexibility of my developed scheme. These datasets are used to quantify trends and synoptic-scale characteristics in subseasonal extreme precipitation. At present, the observational database is fully accessible in tabular format along with informational fact sheets describing the definition and climatology all available online. Rossby wave trains are shown to be present and co-located with anomalously strong moisture in several regions during both observed and simulated extreme periods. The observational database is also used to fit random forests and convolutional neural networks to understand the predictability of these extreme periods. The random forests show some skill above climatology in differentiating extreme and non-extreme days in three different regions of the U.S. with better predictability over the West Coast compared to the Central Great Plains. Although the critical success index (CSI) scores were relatively low, peaking at 0.23 along the West Coast, the random forests appear more skillful when using skill scores tailored towards extreme events. The convolutional neural networks produce probabilities of subseasonal extreme precipitation onto a map over the CONUS are verified using an object-oriented scheme (e.g., centroid offset, area ratio, etc.) and CSI scores reach 0.3. Although the skill is seemingly low, the networks are often identifying “close hits” and may still bring value to end users who are more tolerant to false alarms. The object-oriented scheme allows for flexibility in matching forecasts and observations that end-users can tolerate and gives a

better indication of forecast usability. Although much work lies ahead to develop an operational forecast product, the work herein lays the foundation for future works to continually make advances into the S2S timescale.

Chapter 1

Introduction

Extreme precipitation is a natural hazard that poses significant risk to life and property globally. Between 1980 and 2019, there were 49 flooding and winter storm events affecting much of the contiguous United States (CONUS) with losses exceeding \$1 billion in inflation-adjusted damages (NCEI 2020). These events totaled nearly \$200 billion in Consumer Price Index-adjusted damages and caused 1603 fatalities throughout the CONUS. Impacts from extreme precipitation affect a myriad of sectors, such as water resources (e.g., Piao et al. 2010), agriculture (e.g., Klemm and McPherson 2017), energy (e.g., Schaeffer et al. 2012), and ecosystems (e.g., Knapp et al. 2008). To date, there has been an abundance of studies examining the trends, dynamics, and thermodynamics of extreme precipitation on timescales ranging from hourly to about one week (e.g., Bradley and Smith 1994; Keim 1996; Schumacher and Johnson 2005; Grotjahn and Faure 2008; Kunkel et al. 2012; Moore et al. 2012; Warner et al. 2012; Moore et al. 2015; Barlow et al. 2019; Crossett et al. 2023). However, there remains a knowledge gap about the identification and dynamical characterization of extreme precipitation events on the subseasonal-to-seasonal (S2S) timescale, defined as the period ranging from 14 days to about 2 or 3 months (Vitart et al. 2017; NOAA 2018). Considerable research has been published on short-duration extreme precipitation events, as noted in a recent review by Barlow et al. (2019). Yet, 14 of the 49 aforementioned billion-dollar disasters occurred over periods of 14 days or longer, demonstrating the ability of longer-duration events to cause significant societal impacts. In some cases, these impacts are more detrimental compared to daily extremes because extended-duration events can affect larger areas and populations, ultimately making disaster response more difficult. The epitome of an extreme precipitation period comes from California this past

winter. Between 26 December and 17 January 2023, California saw numerous atmospheric rivers (ARs) make landfall, each bringing an abundance of moisture and forcing extreme flooding and snowfall totals depending on elevation. NCEI currently estimates there were about \$3.5 billion in damages with impacts lasting into mid-March.

There has been growing demand for usable forecasts between medium-range weather and seasonal forecasts as decision-makers attempt to increase resilience to extreme events or optimize resource management decisions. The need for advances in S2S extremes also crosses into the geopolitical and national security realms, seen in the Weather Research and Forecasting Innovation Act of 2017 (HR 313 2017) , which aims to improve S2S forecasts and speed up research to operations. Indeed, this Act specifically directs the National Oceanic and Atmospheric Administration to “collect and utilize information in order to make usable, reliable, and timely foundational forecasts of subseasonal and seasonal temperature and precipitation,” (pg. 131, stat. 98) and to deliver impact assessments from forecasts on the severity and number of floods, among other natural hazards. Additionally, the Act addresses improving numerical weather prediction and forecasting, establishes a program to improve tornado warnings, and provides essential funding to research dedicated to improving forecasting capabilities. In total, the Act serves to greatly benefit socioeconomic welfare by prioritizing research to operations to directly protect life and property in the United States. However, there currently exists a so-called “predictability-desert” between the weather and seasonal scales 1.1. Forecasts on the subseasonal scale generally have lower quality because the influence of the initial conditions is often lost yet forecast lead-time is still too short for long-term boundary forcings (e.g., oceanic variability; Vitart et al. 2017). Nevertheless, prior studies have linked increased predictability to a number of precursors including soil moisture (e.g., Orth and Seneviratne 2014), stratosphere-troposphere coupling (e.g., Domeisen et al. 2019; Domeisen and Butler 2020), and large-scale teleconnections, such as the MJO (Rodney et al. 2013; Grimm et al. 2021) and ENSO (e.g., Domeisen et al. 2015; Yan and Guo 2023).

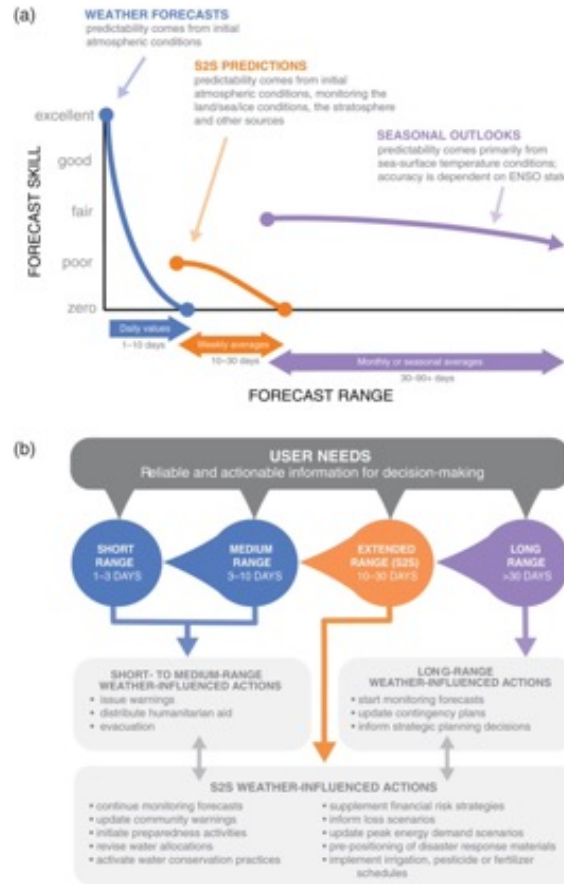


Figure 1.1: Figure taken from (White et al. 2017, Fig. 1). (a) Qualitative estimates of skill from short-range to long-range forecasts. (b) Schematic outlining the relationships among different forecast types and uses in the decision-making process.

This dissertation is part of the Prediction of Rainfall Extremes at Subseasonal to Seasonal Periods (PRES²iP) project funded through the National Science Foundation Prediction of and Resilience against Extreme Events. The PRES²iP project aims to increase societal resilience to S2S extreme precipitation by quantifying the large-scale characteristics that force these extremes and improve our predictive capabilities. Accordingly, my work has been driven by three major goals:

1. Develop a robust framework to define subseasonal extreme precipitation
2. Advance the predictability of subseasonal extreme precipitation
3. Facilitate co-production of knowledge with end-users

Chapter 2 describes my algorithm used to define S2S extreme precipitation using a high-resolution observationally-based dataset. The framework uses thresholds for both total accumulation throughout the chosen length (e.g., 14 days) and the frequency of above normal daily precipitation to ensure I am identifying events that are truly subseasonal (or longer) in nature. Using the developed database, I objectively define regions of subseasonal extreme precipitation across the CONUS and show linkages to ENSO, particularly in the Southwest US. There is also an increasing trend in subseasonal extreme precipitation, particularly in the High Plains and Northeast US. This chapter has been peer-reviewed and accepted in the *Journal of Climate*, accessible at <https://doi.org/10.1175/JCLI-D-20-0580.1>. In Chapter 3, I outline the methodological adaptations needed to translate the observationally-derived algorithm to other datasets. The process is illustrated with both reanalysis and climate model datasets and analogous 14-day databases are generated. Chapter 4 quantifies the large-scale mechanisms driving subseasonal extreme precipitation and their regional differences. Across the US, subseasonal extreme precipitation is driven by Rossby wave trains, which act as a forcing mechanism for ascent, co-located with strong moisture and momentum transport into the region. Chapters 5 and 6 investigate the predictability of these extreme periods. First, I use random forests to predict the likelihood of subseasonal extreme precipitation days within specific regions. I also employ convolutional neural networks that are verified with an object-oriented scheme. The best convolutional neural networks have an object-oriented critical success index of about 0.3, although many of the false alarms arise from forecast periods that were close to meeting my extreme event criteria. Chapter 7 outlines the interactions between PRES²iP scientists and stakeholders at two team-run workshops and the impacts of these workshops on both my immediate research and long-term career goals. Taken together, this work constitutes substantial headway toward improving our capability to identify and predict subseasonal extreme precipitation.

Chapter 2

Subseasonal-to-Seasonal Extreme Precipitation Periods in the Contiguous United States: Generating an Observationally-Based Database

2.1 Introduction

The development of a methodology to catalog S2S extreme precipitation events is the first step in extending forecast skill and ultimately improve societal responses to extreme precipitation events. These S2S extreme precipitation databases can then be used to identify forecasts of opportunity, calibrate and constrain S2S dynamical prediction systems, and to develop statistical models leading to more skillful forecasts of such events on the S2S timescale. What is considered “extreme” is often subject to the spatial and temporal scale being analyzed; Barlow et al. (2019) document 25 different definitions of extreme precipitation in various prior studies. Here, I introduce a method for identifying extreme events that can be consistently applied across time and space.

One of the first definitions of an extreme precipitation event on the S2S timescale came from Jennrich et al. (2020) where their extreme threshold was the 95th percentile of all 14-day windows between 1981 and 2010. After dividing the CONUS into six regions, using geopolitical boundaries and latitude/longitude lines, they defined an event by considering the areal extent within the region of grid boxes that exceed the 95th percentile, ensuring the areal-averaged precipitation was greater than 10 mm day⁻¹ for five or more days, and requiring the day of heaviest precipitation and surrounding days to not be greater than 50% of the total precipitation. Schroers and Martin (2022) extended this work through 2018 and

split the Plains into northern and southern halves because Flanagan et al. (2018) showed distinct drivers of pluvial years in these two regions. However, the division of the CONUS into regions that are not defined by the statistical properties of precipitation causes events that span across regional borders to potentially be missed, distorting the shapes of extreme events. Additionally, the continuity of grid boxes that exceed the 95th percentile was not considered, potentially merging two groups meeting their extreme criteria into one extreme event despite several hundreds of kilometers or more of geographical separation. The dynamics driving the two groups to be extreme would likely be distinct, thereby confounding the linkages in subsequent composite analysis.

The present study builds on the work by Jennrich et al. (2020) by relaxing the assumption of political boundary regions within the CONUS and extending the temporal study period. To illustrate the details and implementation of the framework, I develop a database of 14-day extreme precipitation periods and present a brief climatology of that database. Section 2 presents the algorithm for defining an extreme period with an example. Section 3 focuses on processing and aggregating all potential events into a single database. The temporal and spatial characteristics of the database are discussed in Section 4. A summary and conclusions follow.

2.2 Extreme period identification

2.2.1 Data

I used daily precipitation data from the Livneh (Livneh et al. 2013) and the Parameter-elevation Regressions on Independent Slopes Model (PRISM; Daly et al. 1994) datasets throughout the analysis. Because the Livneh dataset ends in 2011, it was augmented with PRISM through 2018 to complete the record to near-present. The PRISM precipitation data were bilinearly interpolated using the Python library *xESMF* (Zhuang et al. 2020) from a native 4 km grid onto Livneh’s grid of about 6 km. Cubic spline interpolation was tested

and found to offer essentially the same precipitation field as bilinear interpolation when visually compared (not shown), hence the simpler method was adopted. To examine the interpolation error, I took PRISM's daily precipitation between 1981 and 2010, interpolated to Livneh's grid, interpolated back to PRISM's grid, and divided the difference between PRISM's original value and its twice-interpolated value by 2. The mean absolute error over the CONUS was $O(10^{-2})$ mm or $O(10^{-3})$ mm depending on the month. However, errors approached $O(10^1)$ mm in the Sierra Nevada and Cascade Range mountains during boreal winter. Although both the interpolation error and difference between the Livneh and PRISM datasets may become relatively large, particularly in the western third of the CONUS, the Livneh dataset, at least partially, is a function of the 1981-2010 PRISM climatology. This dependence makes PRISM the best choice of precipitation datasets to append to the Livneh dataset. The mean error between the Livneh and PRISM datasets is $O(10^{-3})$ mm (not shown) with Livneh generally being larger than PRISM, albeit by small amounts. Accordingly, the database extends through 31 December 2018, giving a total of 104 years for the analysis.

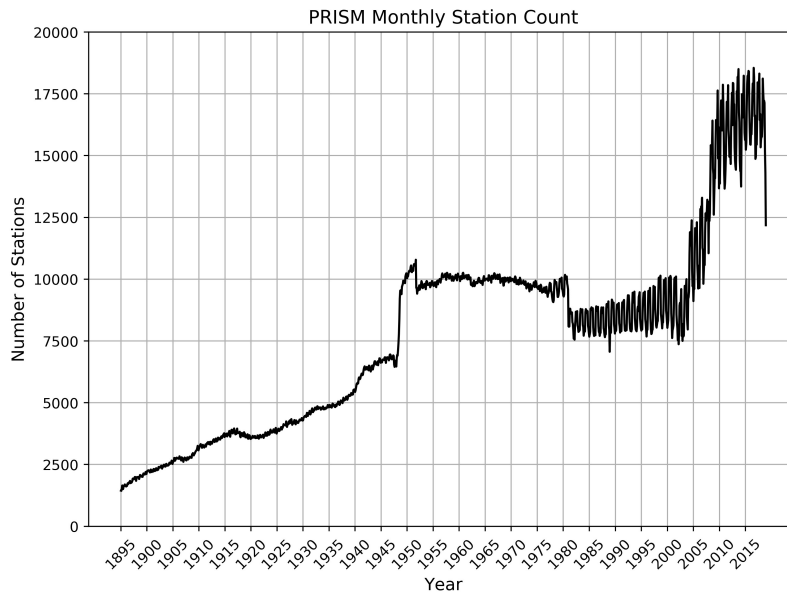


Figure 2.1: Total number of stations used in processing monthly precipitation data from PRISM as a function of time.

Precipitation in the Livneh and PRISM datasets, and thus the calculated extreme value statistics, are affected by the number of rain gauges used by the models. Figure 2.1 depicts a time series of the number of stations employed in calculating monthly precipitation in PRISM; the Livneh dataset does not come with a metafield showing the number of gauges used in processing a given day's precipitation field. The primary precipitation data source for both datasets is the National Weather Service Cooperative Observer Program (Daly et al. 2008; Livneh et al. 2013; Henn et al. 2018), thus it is very likely that magnitude and trend in Fig. 2.1 is similar for the Livneh dataset. Over time, the number of rain gauges in the network increased from under 2500 stations to about 18 000 stations. Lopez et al. (2015) found that having fewer rain gauges led to decreased accuracy in estimating higher rainfall rates. Therefore, in the early 20th century, the decreased spatial density of the rain gauges likely led to many areas having underestimated precipitation amounts relative to later in the 20th century and beyond. Thus, a positive trend in the number of identified events, when using a stationary threshold, in the database may be due, at least to some degree, to station density.

The multivariate ENSO index (Wolter and Timlin 2011) was also used to assess connections between ENSO phase and bi-monthly extreme period frequency. The MEI is defined in two chunks: the MEI extension, which spans 1915–1949, and MEI version 1 which I use through 2018. La Niña (LN) conditions were defined as the bottom (i.e., coldest) 32 years for each bi-monthly season and, similarly, El Niño (EN) seasons were the warmest 32 years. I use a χ^2 test with an expected frequency of 30.8% for EN/LN (32 seasons divided by 104 years) and the remaining 38.4% is the expected frequency of the neutral (N) phase.

2.2.2 Extreme criteria

I began by finding all grid points within the CONUS which experienced extreme precipitation conditions over a given 14-day window. A grid point was flagged as having experienced extreme precipitation if it passed two criteria. First, total precipitation must meet or

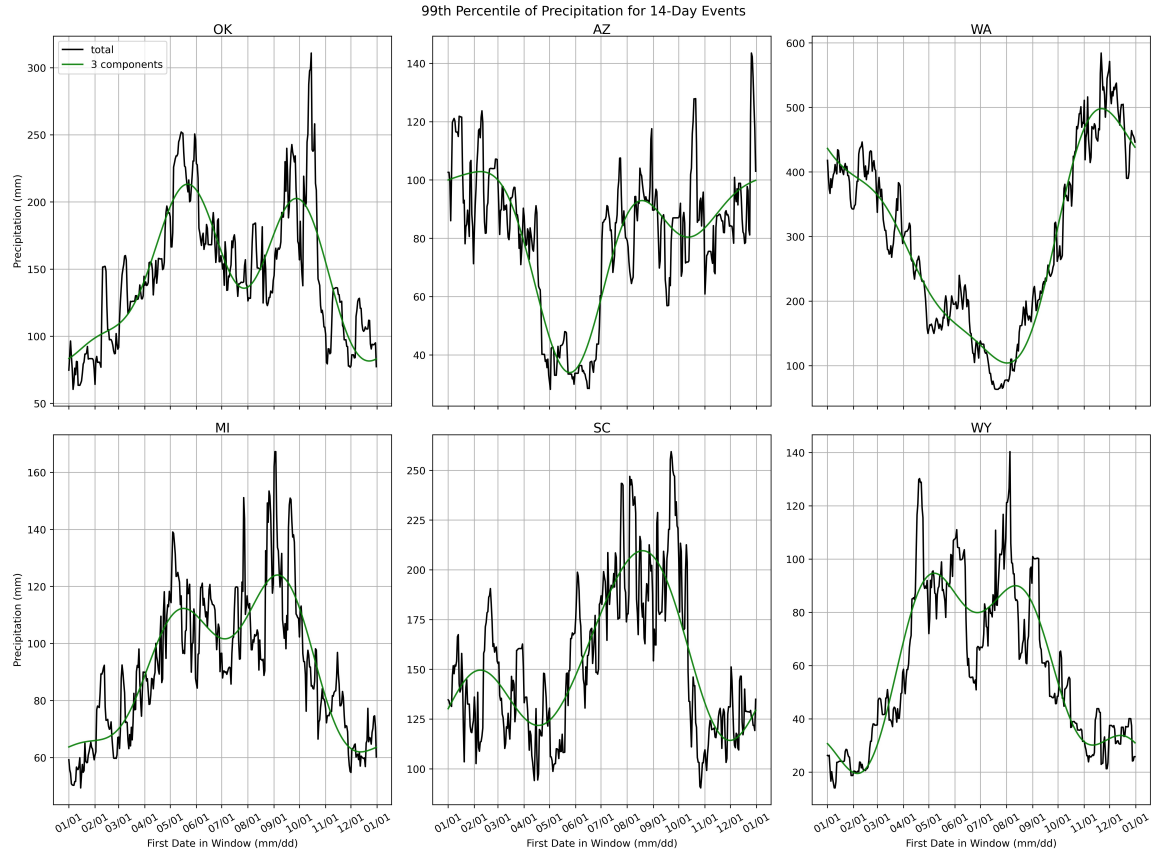


Figure 2.2: (black) Raw 99th percentile for each 14-day window and (green) smoothed percentiles using the first three Fourier harmonics for six selected grid points across the CONUS. The plotted date is the start date of each window.

exceed the 99th percentile. I selected the 99th percentile to identify extreme events in the far-right tail of the precipitation distribution as these events are very likely to be coincident with major socioeconomic impacts. The 99th percentile was calculated using each 14-day sum over all 104 years for each grid point and each overlapping 14-day window (i.e., 1-14 January, 2-15 January, etc.). The black line in Figure 2.2 depicts the derived percentiles as a function of calendar day for six points across the CONUS: Norman, Oklahoma (OK); Sedona, Arizona (AZ); 1° longitude east of Seattle, Washington (WA); 1° longitude west of Port Huron, Michigan (MI); Colombia, South Carolina (SC); and Cheyenne, Wyoming (WY). The large degree of window-to-window noise in precipitation is apparent in the corresponding percentiles with oscillations of ≥ 50 mm seen in the span of two weeks at every grid point. Therefore, I smoothed the raw percentiles by decomposing each time series into

its Fourier coefficients. I tested reconstructed time series by retaining harmonics ranging from two to six and subjectively determined that three harmonics led to the best smoothed signal (see green line on Fig. 2.2; see Figs. A.1, A.2 for results of sensitivity testing). Furthermore, numerous studies have examined precipitation seasonality and shown that the first three harmonics generally explain the majority of the variance in the precipitation distribution (e.g., Scott and Shulman 1979; Kirkyla and Hameed 1989; Epstein and Barnston 1990). Figure 2.3 displays the total thresholds for one day in each season. Each window and each grid point in the CONUS will have its own threshold criterion. To illustrate, consider a precipitation field over the example window of 12-25 January 1937 (Figure 2.4a). Points that pass the first criterion are where values in Fig. 2.4a exceed the extreme threshold values in Fig. 2.4b (e.g., parts of Arkansas, Tennessee, Kentucky, Indiana, etc.).

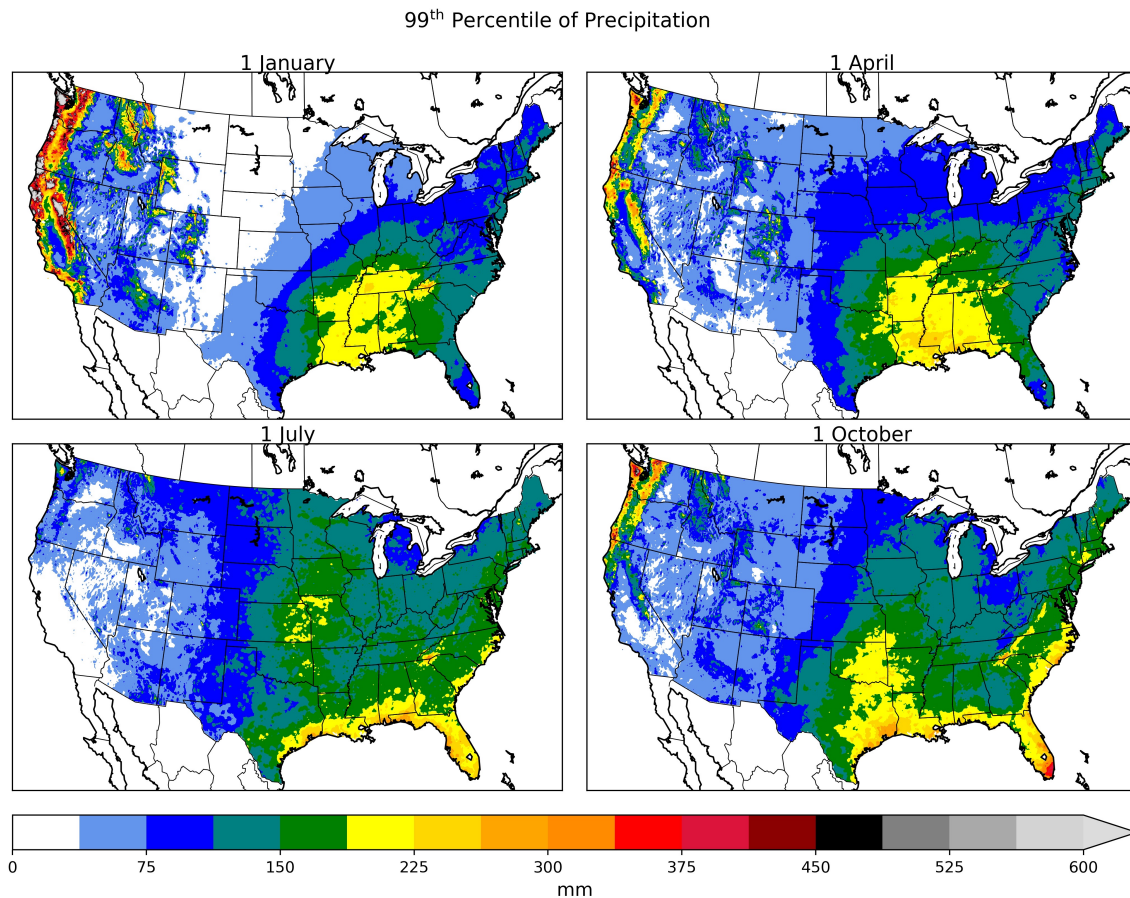


Figure 2.3: The 99th percentile of precipitation in the CONUS for one 14-day window in each meteorological season.

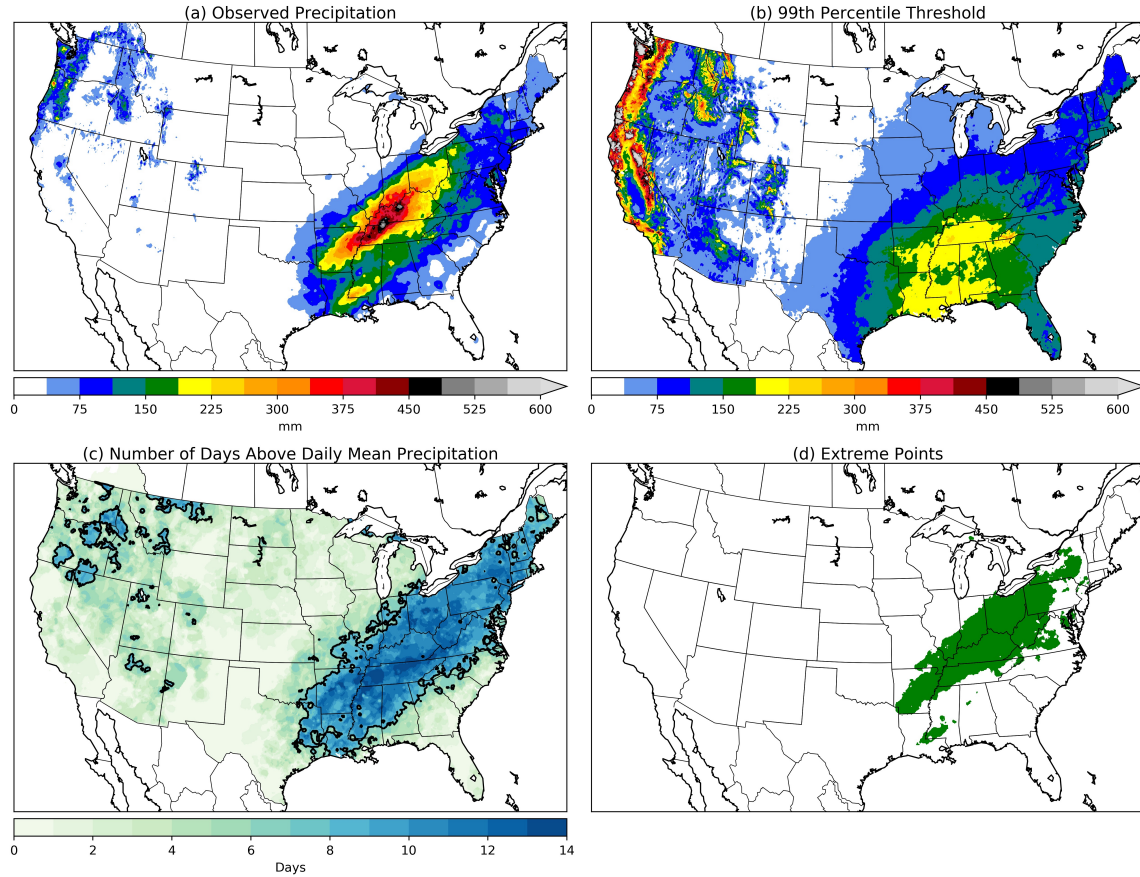


Figure 2.4: (a) Total recorded precipitation (mm) between 12 and 25 Jan 1937. (b) The 99th percentile of precipitation (mm) derived for 12-25 Jan. (c) Number of days with above normal daily precipitation during 12-25 Jan 1937. Black contour outlines 7 days, which is the threshold for the second criterion in event selection (see text). (d) Points flagged as extreme per the algorithm (see text for details).

The next criterion considers the duration of the precipitation during the window. Although a duration check may eliminate short-duration events with exceptionally heavy precipitation totals, I aim to identify events with persistent synoptic-scale forcing to maximize potential predictability at meaningful (i.e., S2S) lead times to decision-makers. Given the importance of defining persistently-forced extreme precipitation events, for a grid point to be considered extreme, at least 7 of the days (i.e., half the window length) must receive greater than or equal to the long-term mean daily precipitation. I defined the climatological daily precipitation as the long-term mean of all days in a given 14-day window over all 104 years. In other words, I calculated the mean daily precipitation using all 1456 days in

the period of record. The relatively large sample size used in calculating the climatological mean led to more stable intraseasonal variability. The mean and median difference between one calendar day's long-term mean and the next day's mean is $O(10^{-5})$ mm and the maximum value is a mere 2.3 mm. Accordingly, further smoothing using Fourier harmonics was unnecessary. Grid points that meet or exceed the duration criterion are shown as medium and dark blue in Fig. 2.4c and are also outlined by the solid black contour.

2.2.3 Finding an extreme region

The two aforementioned criteria were applied to each grid point in the domain independently, and the grid points that met both criteria were flagged as extreme (e.g., Fig. 2.4d). See Figure 2.5 for a summary of the logic. Once grid points were identified as extreme, the next step was to consider the spatial continuity of these points in defining an extreme region. To objectively identify extreme regions, kernel density estimation (KDE; Rosenblatt 1956), a nonparametric technique that fits a probability density function (PDF) to a given set of observations, was employed. Here, the geospatial coordinates of all points flagged as extreme were used to define a 3-D PDF over the CONUS where the vertical coordinate (i.e., density) is greater where more extreme points are in close proximity. The density of a given point y with observations $x_i, i = 1, 2, \dots, n$, is given by:

$$\rho(y) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{y - x_i}{h}\right) \quad (2.1)$$

where K is the kernel, a prescribed weighting function, and h is the bandwidth (i.e., a smoother for the resulting PDF). After considering the advantages and disadvantages of a number of kernels, I concluded the Epanechnikov kernel was best suited for both this study and future studies with globally-gridded precipitation data. The Epanechnikov kernel for a random variable u is given by:

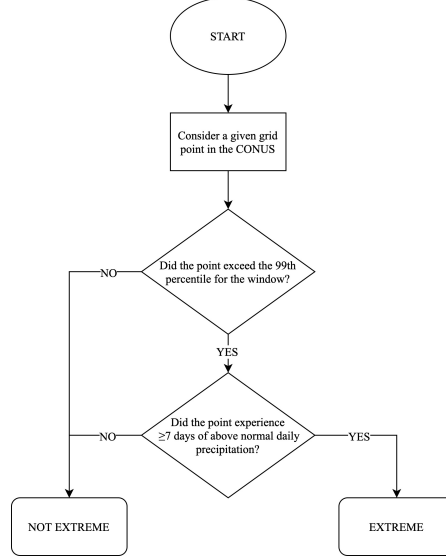


Figure 2.5: Flowchart describing the logic in determining if any given point is labeled extreme. See text for details.

$$K(u) = \begin{cases} \frac{3}{4}(1 - u^2) & |u| \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (2.2)$$

The Epanechnikov kernel was selected for its parabolic shape and the assignment of zero density outside of $|u| \leq 1$ (i.e., its support). Other kernel options, such as Gaussian and exponential, do not impose a support, which leads to a smoother density field. One benefit of Epanechnikov's more rigid kernel is its ability to effectively outline areas outside the CONUS where Livneh and PRISM lack data. Hence, polygons tend to follow the boundary and have little extent into the oceans. A Gaussian kernel, for instance, will still lack significant extent over oceans, but its smoother density field would produce a lesser gradient; however, I have no data over the oceans to support such a gradient. Cross-validation tests were conducted to determine the optimal bandwidth on a number of test events, with varying times and locations. These tests consistently gave bandwidths from 0.01 to 0.03. Therefore, I used a bandwidth of 0.02.

The resulting density was fit on a grid with 0.1° resolution in the domain 24°N - 50°N ; 128°W - 66°W , where the great circle distance calculated using the Haversine formula is

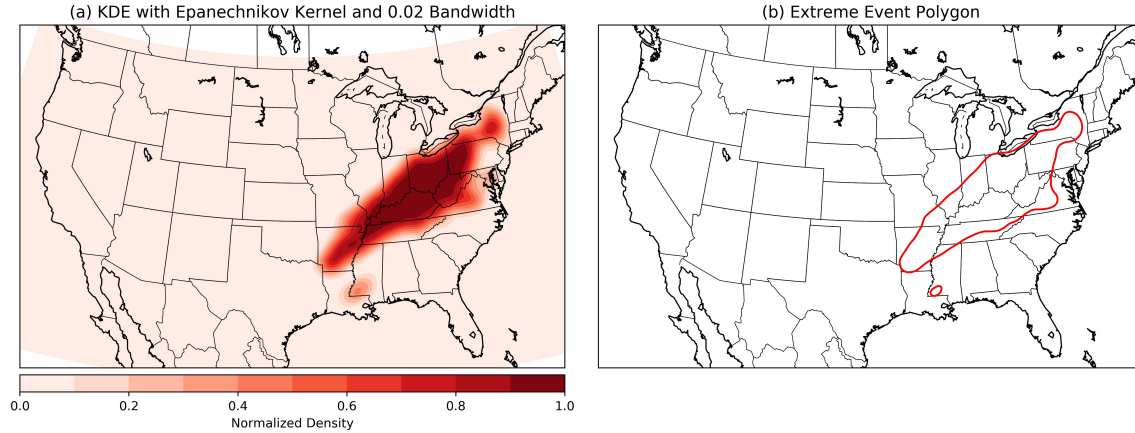


Figure 2.6: (a) Full KDE normalized density field using the Epanechnikov kernel and 0.02 bandwidth for the 12-25 Jan 1937 window. (b) Extreme period polygon drawn using the 0.2710 contour.

used as the distance metric. In the midlatitudes, a bandwidth of 0.02 equates to a point on the grid having nonzero density if it is within roughly 1.25° of an extreme point. Next, the density was normalized by dividing by the maximum value to facilitate direct comparison of potential events from different regions of the CONUS and during different times of the year (Figure 2.6a). Subsequently, an extreme precipitation period region was developed using a discrete normalized density value. I found a density value of 0.2710, the 99th percentile of all density fields between 1915 and 2018, produced representative and consistent polygons across both time and space. Fig. 2.6b displays the polygon drawn using the single contour, outlining the area defined as extreme for the 14-day window. A 14-day extreme precipitation period was then defined to have occurred if the area inside the polygon, calculated using Green's Theorem, exceeds $200,000 \text{ km}^2$. I calculated the areas of every KDE-derived polygon between 1915 and 2018 and examined the distribution of areal extents. An areal extent of $200,000 \text{ km}^2$ records the largest 5% of all areas captured and represents a “critical” point where the derivative of areas as a function of percentiles changes rapidly (Fig. A.3).

2.3 Cataloging extreme periods

Section 2.2 outlined how to identify an extreme precipitation period given a single window. The database was generated by considering all windows in the period of record and aggregating all identified events. I developed a Python program¹ that constructs a unique object based on input day, month, and year. The program is available for users to explore any date; e.g., other researchers may want to define their own regional database of events using the same methodology but with a smaller areal extent. The database was generated by inputting every possible start date from 1 January 1915 to 31 December 2018 (excluding leap days) and archiving identified extreme events to a text file. However, a consequence of simply moving a window in increments of a single day is the database contained extreme events with polygons over virtually the same area for several consecutive days. Thus, I developed an objective post-processing algorithm to group “repeat” events, label one of the events in the group as the extreme period, and discard all others.

The post-processing algorithm that discards all “repeats” began by considering the 2-wk window in the first row of the pre-processed text file (denoted as E_0 , see Figure 2.7). A smaller, subset file was constructed where only the events with starting dates on or before the ending date of E_0 are considered. For example, if E_0 is 1-14 January, then I found all events that have a start date on or before 14 January. Then, events were grouped by the location and shape of their polygons. A binary matrix with the same domain and resolution as the KDE density (see Section 2.2.3) was constructed for all events in the subset file where “1” denotes a grid point inside the event polygon and “0” denotes a grid point outside the polygon. The correlation coefficient r was calculated between all events in the subset file with respect to E_0 ; extreme precipitation events are considered to be representative of the same extreme period if $r \geq 0.5$. The events are then placed into a group of similar events.

¹<https://github.com/tydickinson29/PRES2iPpy/blob/master/pres2ippy/analysis/rainpy.py>

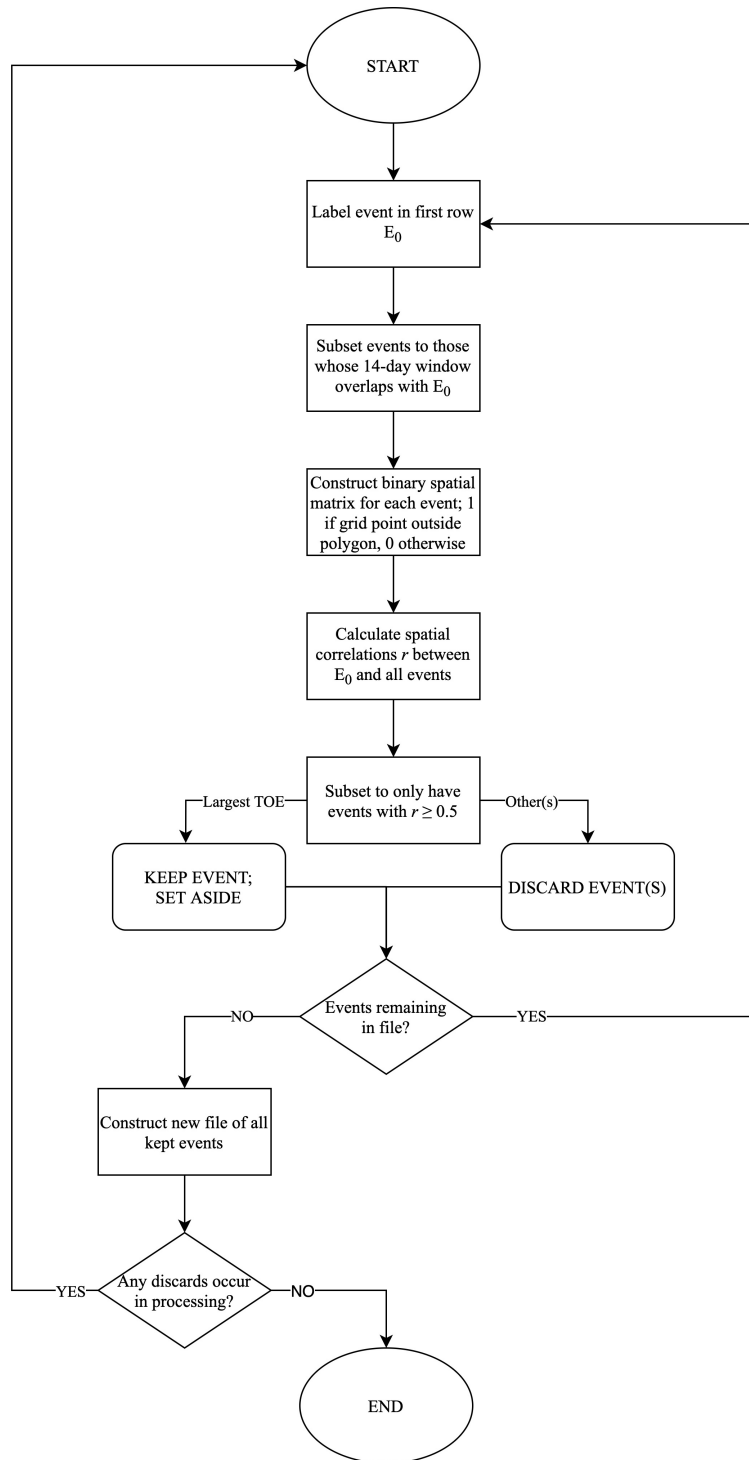


Figure 2.7: Flowchart describing the logic during postprocessing of extreme precipitation events. See text for details.

Within this subsetting group, the “most extreme” precipitation period was chosen for the database. To make this choice, I define a total over extreme (TOE) as:

$$\text{TOE} = \sum_{i=1}^n (P_{\text{total}}^i - P_{q99}^i) \quad (2.3)$$

where i denotes a particular grid point that was flagged as extreme, n is the total number of flagged grid points within the period’s polygon, P_{total} is the grid point’s total 14-day precipitation, and P_{q99} is the grid point’s 14-day 99th percentile threshold. TOE measures the extremity of the period by employing anomalies relative to the 99th percentile while indirectly considering the areal extent since larger polygons tend to have more points inside flagged as extreme. The period with the largest TOE within the group was determined to be “extreme” and is retained for the database whereas all other events in the group are discarded. This process of forming groups of similar events, based on the $r < 0.5$ criterion, was recursive until all events in the file either do not overlap temporally or do not have a correlation coefficient above 0.5. After post-processing, the 14-day database contains 851 unique events.

Table 2.1 lists a few events identified in the database with various statistics describing each period. A full table can be accessed at <http://pres2ip.com/extreme-event-tables> where users can apply filters to find specific events based on time and/or location. Additionally, the database in .csv and .shp formats, and the scripts used to generate and post-process the database are available at <https://github.com/tydickinson29/PRES2iPpy/tree/master/pres2ippy/databases>.

Table 2.1: A sample of events from version 1.0 of the 14-day extreme precipitation period database. $Precip_{AA}$ denotes areal-averaged precipitation in the region; Max_{total} is the greatest 14-day precipitation total from any point inside the region; Max_{1day} is the greatest single-day precipitation total from any point inside the region from any of the 14 days. Area and TOE are rounded to the nearest square kilometer and millimeter, respectively, and all other entries are rounded to 2 decimal places.

Begin date	End date	Area (km ²)	TOE (mm)	$Precip_{AA}$ (mm)	Max_{total} (mm)	Max_{1day} (mm)	Lon_C (°E)	Lat_C (°N)	Cluster
12 Jan 1937	25 Jan 1937	881 375	1 558 392	188.45	541.00	138.07	276.01	38.52	12
19 Nov 1998	2 Dec 1998	244 172	298 550	312.32	1366.15	286.08	237.52	44.28	3
7 May 2015	20 May 2015	234 242	277 190	223.28	458.09	228.37	263.34	34.59	13

2.4 Statistical characteristics of 14-day extreme periods

2.4.1 Extreme period regionalization

Another research goal is to objectively define regions of extreme 14-day precipitation. Classifying the events in the database into geographically similar clusters allows for the identification of typical extreme period shapes across the CONUS, as well as labelling specific extreme events as being shaped exceptionally anomalous. In addition, characteristic patterns associated with the events can be diagnosed for each region and is the focus of Chapter 4.

Two commonly applied types of clustering algorithms that are implemented on gridded data are hierarchical and nonhierarchical. Hierarchical clustering tends to use an agglomerative approach where each observation begins as its own cluster and observations recursively are merged together with the end result being visualized as a tree. Nonhierarchical methods utilize other recursive algorithms that do not form a hierarchy. In a rigorous examination of clustering techniques applied on precipitation data in the eastern two-thirds of United States, Gong and Richman (1995) found nonhierarchical methods to outperform hierarchical methods. Therefore, I apply the k -means clustering (Forgy 1965; Lloyd 1982)

implementation in the Python library *scikit-learn* (Pedregosa et al. 2011). The k -means algorithm clusters data into k groups of equal variance, attempting to minimize the intra-cluster sum-of-squares (i.e., inertia) using Euclidean distance. After the extreme periods are clustered, the silhouette score (Rousseeuw 1987) is calculated as a metric for how well each extreme period fits its assigned cluster. The silhouette score for a given sample is given by:

$$s = \frac{b - a}{\max(a, b)} \quad (2.4)$$

where a is the mean distance between a sample and all other samples in the cluster, b is the mean distance between a sample and all other samples in the next nearest cluster, and $\max(a, b)$ is a function which returns the higher value of a and b . From Eq. (2.4), the silhouette score can range from -1 to +1, where higher positive numbers indicate a good match to the assigned cluster. Events with negative silhouette scores are removed as they indicate a sample was likely misclassified and/or is highly anomalous relative to all other events in the cluster (Rousseeuw 1987). The silhouette score has been applied to meteorological applications to remove misclassified events (e.g., Richman and Adrianto 2010; Mercer et al. 2012).

For each extreme period in the database, a vector consisting of zeros and ones is constructed, where a grid point inside the polygon is assigned a one, with the same resolution as the KDE density field. K -means clustering is then applied iteratively in the following manner:

1. Cluster n events into k clusters;
2. Calculate the silhouette score, s , for all n events;
3. Find the number of events with negative silhouette scores (n^*);
4. If $n^* \neq 0$, remove the events with $s < 0$ and repeat from Step #1, otherwise end.

Choosing the optimal value for k is the most difficult and subjective decision when implementing any clustering algorithm. The elbow method (Thorndike 1953) and the silhouette method (Rousseeuw 1987) are two popular tools that are used as guidance to determine the best value for k . In both methods, a range of values for k are tested; for the work herein, I test $k = 5, 6, \dots, 30$. However, neither method depicts a clear signal as a tool in deciding which k to choose (Figures 2.8a and 2.8b). Although the inertia continues to decrease with increasing k , there is no cusp present. Similarly, the average silhouette scores generally increase with increasing k , though a major reason for the positive correlation is from the lower number of events being clustered. Comparing Figs. 2.8b and 2.8c, there is an inverse relationship between the average silhouette and the number of clustered events. Consequently, I define an index coined the Hybrid Index (HI), comprised of the product of the average silhouette and number of clustered events. The best value for k is determined to be the one where the maximum HI value occurs; for the 14-day database, the optimal number of clusters is $k = 15$ (Fig. 2.8d).

Figure 2.9 depicts the “average” polygons for each cluster. The average polygon is created using all events in each cluster; the average number of events per cluster is about 50. The polygons in the cluster are rasterized to a grid with 0.01° spacing where the grid bounds are the minimum and maximum longitudes and latitudes from all polygons in the cluster. The rasterized grid is normalized to represent a probability that any given grid point is within a polygon, ranging from 0 (no coverage) to 1 (total coverage). This grid is then convolved with a Gaussian kernel with a radius of 100 grid points to smooth the normalized grid. The final smoothed polygon is the quantile threshold of 0.2, thus depicting the area that is covered by $\geq 20\%$ of the events assigned to the cluster. Although there are some gaps in areas of cluster coverage, the majority of the CONUS is described with the chosen number of clusters, with the gaps likely arising from geographical limitations of employing KDE on a dataset only defined over land. Maps with a smaller value for k than chosen herein leave large parts of the CONUS outside a cluster or are too general in their shapes whereas larger values for k yield clusters that have extensive spatial overlap (i.e.,

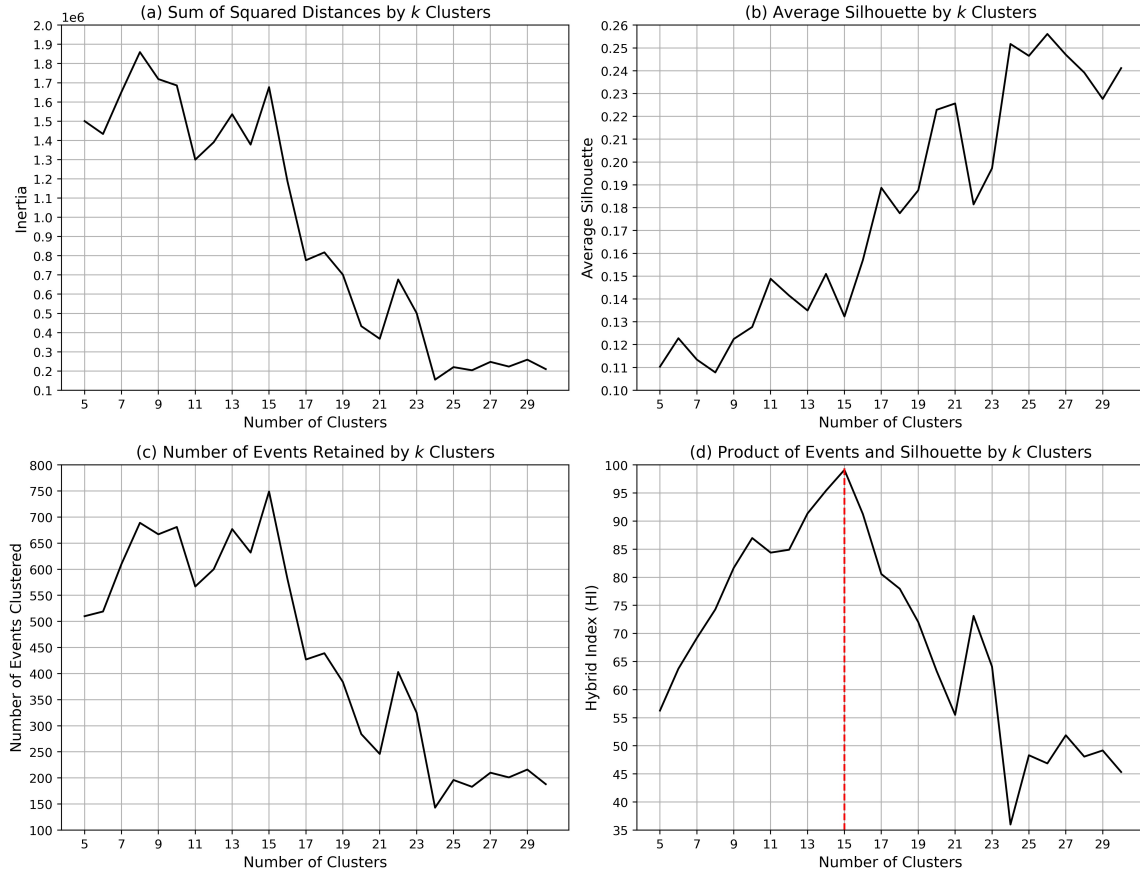


Figure 2.8: Various statistics as a function of number of clusters for 14-day extreme precipitation periods: (a) intracluster sum of squared distances (i.e., inertia), (b) average silhouette score, (c) total number of clustered events, and (d) product of (b) and (c), (i.e., the HI). The dashed red line in (d) marks the HI maximum.

large cluster intercorrelations). Overall, 102 of 851 events ($\sim 12.0\%$) are not assigned to a cluster.

Gong and Richman (1995) found the biggest limitation to the physical validity of k -means results is its sensitivity to the initial cluster center locations. To minimize such a problem, cluster centers are initialized using the “ k -means++” method (Arthur and Vassilvitskii 2006). This method begins by placing a random cluster center, but then uses a probabilistic approach to effectively disperse the other centers among attributes to yield improved cluster assignments. Nevertheless, after choosing k for both small and large events, I test 10 different seed initializations to examine the spatial stability and/or reliability in each of the clusters where the final assignments are from the seed with the largest

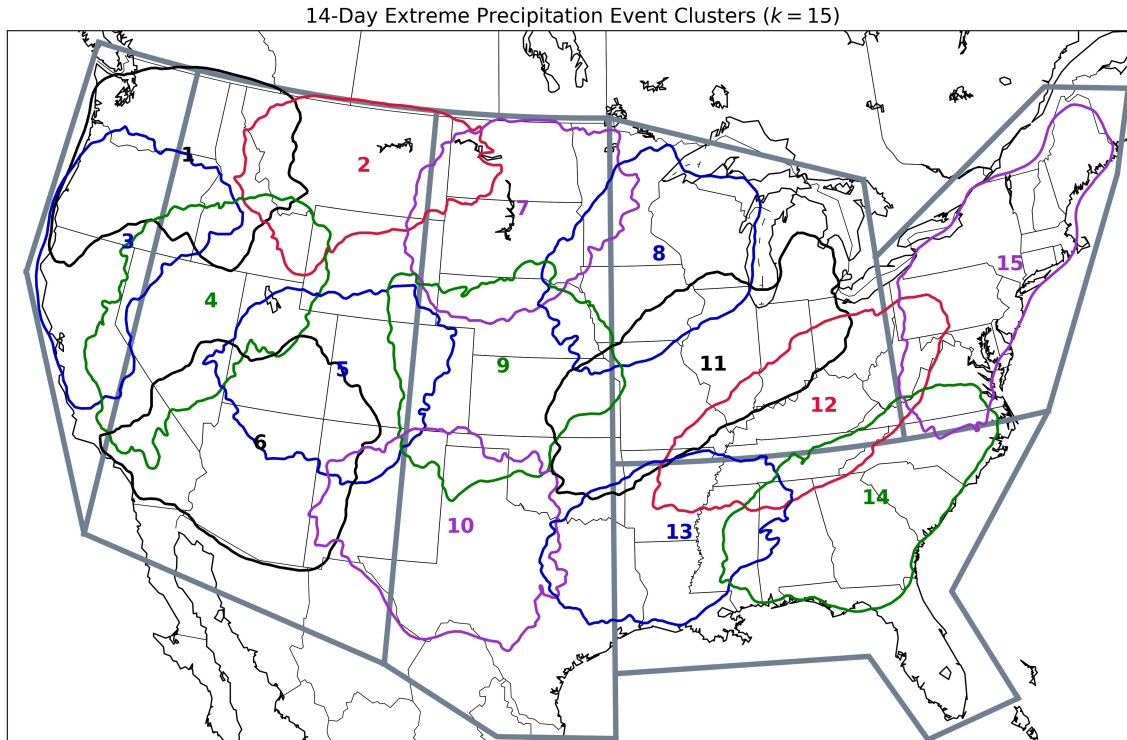


Figure 2.9: (colored contours) Average polygon for $k = 15$ clusters across the CONUS for 14-day extreme precipitation periods. Also shown are the regional boundaries defined by Jennrich et al. (2020) (gray outlines).

product between events clustered and average silhouette. In general, each of these seeds produced maps similar to those in Figure 2.9 (Fig. A.4). Although 3 seeds moved a cluster from the Mountain West to the Southeast (Figs. S4c, S4e, and S4g), the remaining seeds yielding consistent results.

To examine cluster stability further, I also test hierarchical clustering methods using Ward's linkage and average linkage. Ward's linkage produces clusters that are mostly similar to k -means but with a few differences (Fig. A.5). Cluster 3 is centered further north and extends further east than the k -means solution, and Cluster 10 stretches close to the southernmost point in Texas. Cluster 5 is the most different (see the orange region in Fig. S5) between Ward's linkage and k -means with the Ward's cluster shifted northwest with a nearly zonal major axis compared to a nearly circular region in the k -means case. The remaining clusters are very similar between the two algorithms. Conversely, average linkage produced chained results; 92% of clustered events are assigned to 1 cluster, a

consequence of its minimization criterion. Such behavior was also noted in Gong and Richman (1995) for certain average linkage implementations. Seasonality is also not expected to significantly affect cluster assignments since accumulation thresholds are different for each window throughout the year and events are recorded in all clusters' warm and cool seasons.

Fig. 2.9 displays potential weaknesses in defining regions before identifying events. Several clusters have significant portions of the average polygon in two or more regions used by Jennrich et al. (2020)(e.g., clusters 1, 11, 12, and 13). Extreme periods in these clusters may have been completely missed, owing to the period's areal extent being distributed among several regions, or counted twice but in different regions. Although the event depicted in Fig. 2.6b occurred before the analysis of Jennrich et al. (2020), the event polygon lies in the Plains, Southeast, Great Lakes, and Northeast. Additionally, some of the regions used in Jennrich et al. (2020) seem overly extensive, depicted by several of the clusters inside of a single region. For instance, clusters 2, 4, 5, 6 and a portion of 10 are inside the Mountain West, while several clusters cross into the Plains region. The developed clusters suggest that, as in Schroers and Martin (2022), the Plains should be divided into sub-regions, likely owing to contrasting atmospheric drivers for the northern and southern Great Plains (Flanagan et al. 2018). Conversely, cluster 15 is nearly identical to the Jennrich et al. (2020) Northeast region.

Although the framework here was able to eliminate most of the geopolitical boundaries within the CONUS, extreme periods are still bounded by the physical (i.e., oceans) and international borders of the CONUS. Nevertheless, the geographical locations and orientations of the clusters suggest favorable general synoptic drivers. Clusters 1, 2, and 7 likely have similar synoptic characteristics, such as the region being centered near an anomalous trough-ridge inflection point, with slightly differing phase and amplitude of the upper-level wave; Konrad (2001) showed 500 hPa cyclone centers were commonly observed upstream from extreme precipitation regions when considering 2-day extreme events. Additionally, Jennrich et al. (2020) and Zhao et al. (2017) showed similar trough-ridge dipoles across

nearby area, such as the Great Plains, Great Lakes, and Northeastern United States. Clusters 8, 11, 12, 13, and 15 have a southwest–northeast orientation and have less polygon variability in the clusters, signaling the synoptic influence of a longwave trough and strong surface baroclinicity associated with extratropical cyclones. Conversely, cluster 9 has larger variability in the shape of its extreme periods, suggesting a larger mesoscale influence within the synoptic signal (e.g., upslope flow, mesoscale convective systems; Schumacher and Johnson 2006). The broad shapes seen in clusters 6 and 10 likely have monsoonal-type flow with strong vapor transport from the south and southwest, supported by strong integrated vapor transport from the southwest into area (Zhao et al. 2017; Jennrich et al. 2020). Similarly, clusters in the southeast CONUS presumably rely on strong moisture transport from the Gulf of Mexico and Atlantic Ocean and are influenced by the subtropical jet (Jennrich et al. 2020).

2.4.2 Subseasonal extreme period frequency

The temporal characteristics of the 14-day extreme precipitation periods are next examined. The median number of events per year within the CONUS is 8, and the mean number events is about the same (8.2). However, these statistics are not time invariant. Figure 2.10 depicts the number of extreme events over time in the database and the trends in the 10th, 50th, and 90th percentiles (i.e., $\tau = 0.1$, $\tau = 0.5$, $\tau = 0.9$ quantiles) calculated using quantile regression (QR; Koenker and Bassett Jr. 1978). QR is an extension of ordinary linear regression that incorporates a loss function such that the regression line can be fit to any part of the probability distribution. Furthermore, QR makes no distributional assumption of the predictand and inherits the same assumptions on the error terms as ordinary linear regression. All 3 quantiles have a positive trend with the median number of observed events increasing at a rate of about 1 period per 50 years. Statistical significance of the slopes was tested using bias-corrected and accelerated (BCa) bootstrapping (Efron 1987) with 10 000 replications at $p < 0.05$ and a null hypothesis of no trend. The 95% confidence intervals

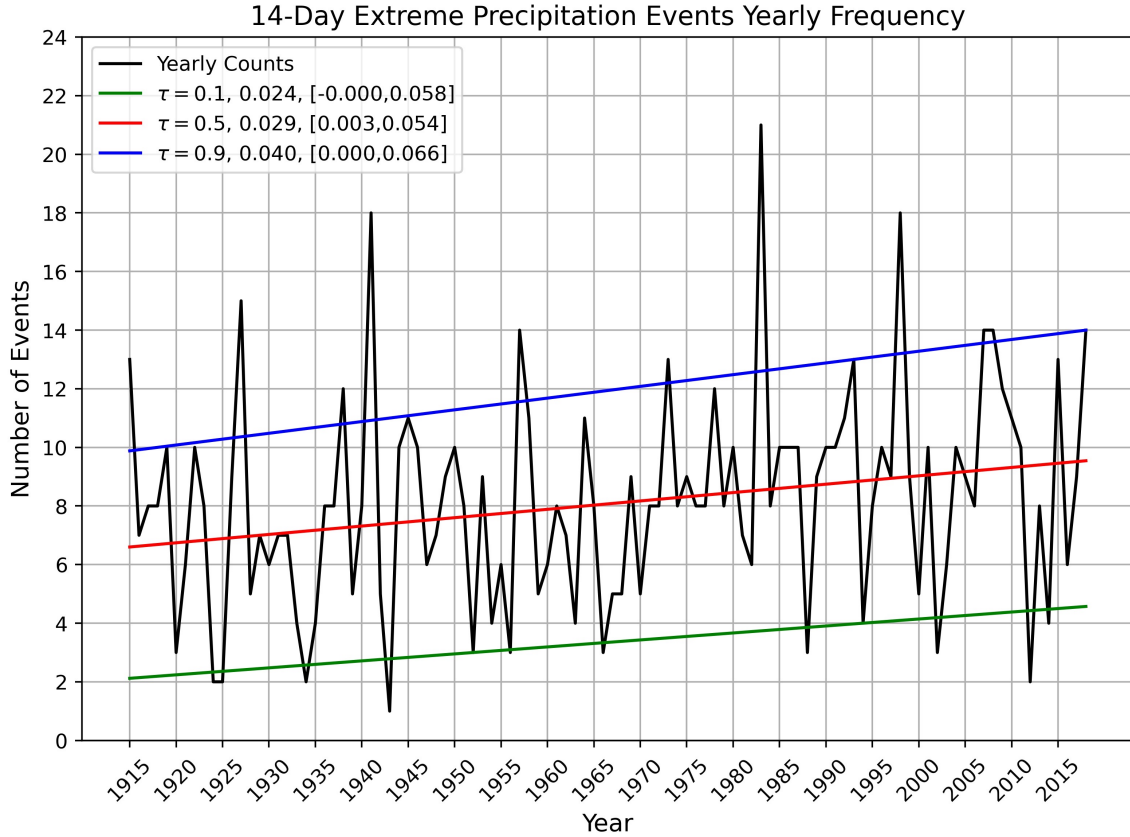


Figure 2.10: (black) Total yearly counts of recorded 14-day extreme precipitation periods throughout the CONUS as a function of time. QR models are fit on (green) $\tau=0.1$, (red) $\tau=0.5$ (i.e., median), and (blue) $\tau=0.9$ quantiles of the time series. Slopes are listed as the second number in the legend, and the 95% confidence intervals on the slope are shown in brackets thereafter. Confidence intervals were calculated using BCa bootstrapping. Slopes and confidence interval bounds were rounded to 3 decimal places.

on the slopes are given in the caption of Fig. 2.10. Although the null hypothesis cannot be rejected for the 10th percentile, it can be rejected for the 50th and 90th regression models as the associated lower bounds are positive. Furthermore, a permutation test on the means of the first and second halves of the time series yielded $p \approx 0.02$, suggesting there was a significant increase in the mean number of events CONUS-wide between 1967 and 2018 compared to 1915 and 1966. The median number of grid points within extreme event polygons has also increased over time and is significant at $p < 0.05$ (Figure 2.11). Although the 10th percentile is also increasing at a statistically significant rate, a permutation test on the areas of events in the first and second halves of the period of record yielded $p \approx 0.78$. Thus,

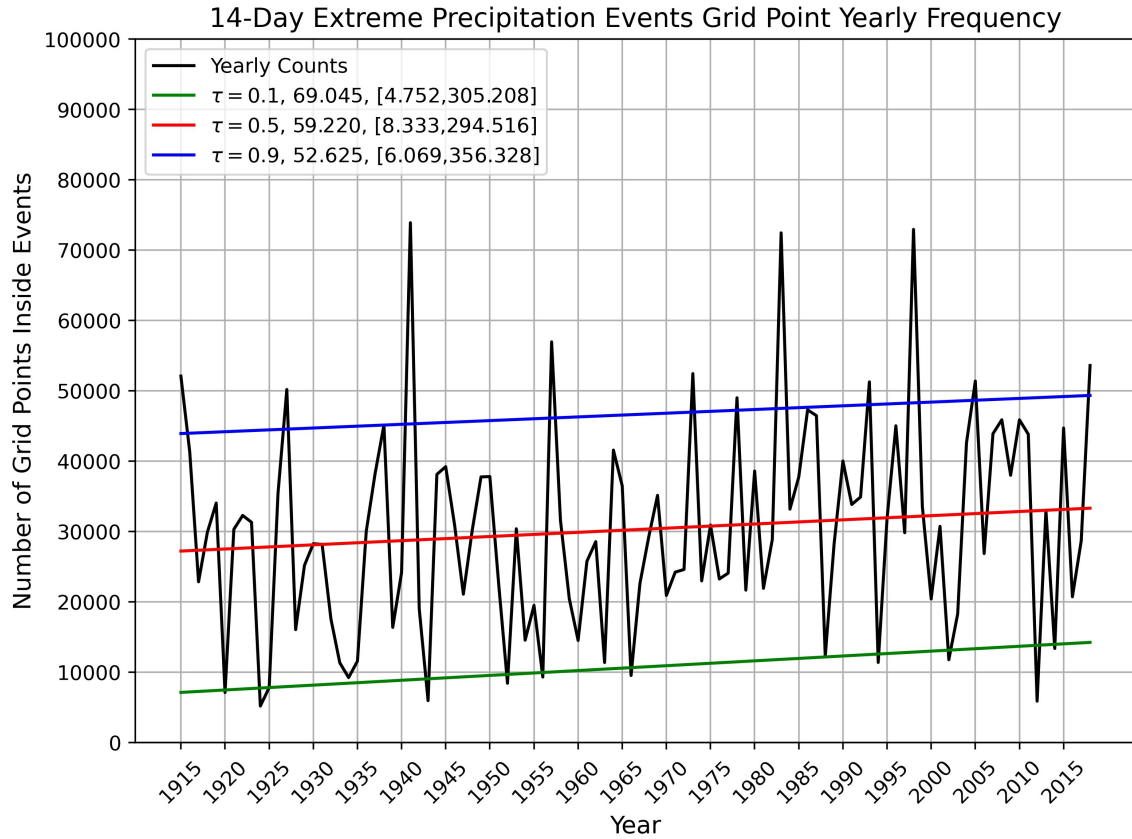


Figure 2.11: As in Fig. 2.10, but for the total number of grid points inside extreme periods as a function of time.

the increasing number of grid points in extreme events cannot be attributed to increasing extreme period sizes and corroborates the finding that the frequency of extreme precipitation events, defined by the 99th percentile, in the CONUS is increasing. The increased station density noted in Fig. 2.1 may also be a factor in the increasing frequency of extreme precipitation events, a possibility left for future research.

It should also be noted that 2 out of the 3 years with the highest number of events occurred in 1983 and 1998, corresponding to 2 of the strongest ENs on record. The strongest ENs on record occurred in 2015/16 and although less apparent as the aforementioned years, 2015 was another active year for extreme precipitation events as it lies exactly on the $\tau = 0.9$ QR model. Indeed, subseasonal extreme precipitation periods across the CONUS occur significantly more frequently during EN phases (Fig. 2.12a). Clusters 4, 6, and 10 also see significant increases when EN is active (Fig. 2.12e,g,k) with roughly 3 times more

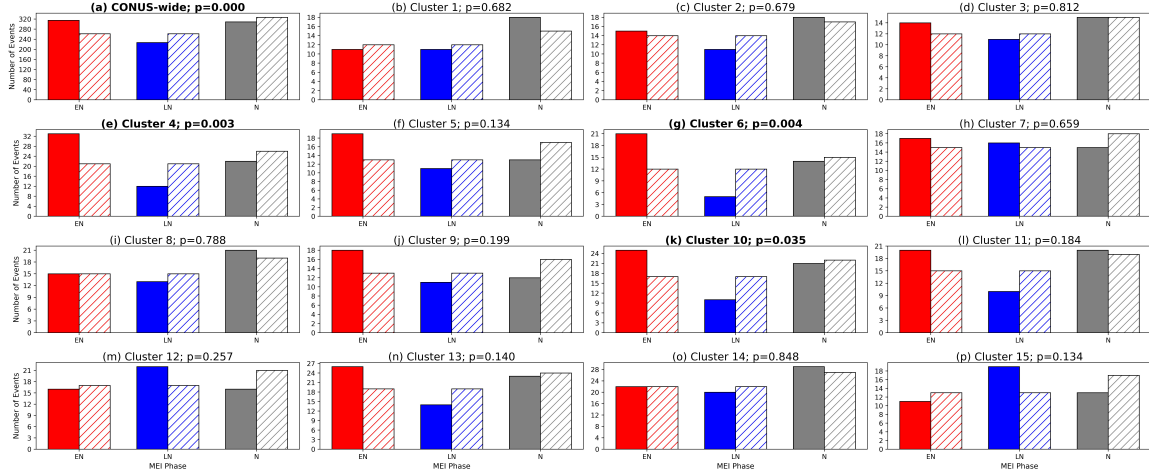


Figure 2.12: Bar plots comparing the (solid) observed frequencies and (hatched) expected frequencies of subseasonal extreme precipitation periods across the 3 ENSO phases. All extreme periods in the database are considered in (a), whereas panels (b)-(p) are for specific clusters. Bold titles indicate statistical significance at $p < 0.5$.

extreme periods during EN compared to LN. Jennrich et al. (2020) also reported connections between subseasonal extreme precipitation in the CONUS to other climate modes such as the Pacific-North American pattern (Wallace and Gutzler 1981) and the North Pacific Oscillation (Wallace and Gutzler 1981; Baxter and Nigam 2015), and comparisons with this newly created database are a subject of future work.

Spatially, there are relative maxima in extreme period frequency in Idaho, Oregon, and northern Nevada and across Arkansas and western Tennessee and Kentucky (Figure 2.13). There is a minimum in frequency in central Wyoming which is co-located with the largest gap in the interior CONUS for the developed clusters (Fig. 2.9). There are no discontinuities in the frequencies throughout the CONUS, which would not be the case if geopolitical or latitudinal/longitudinal delineations were chosen before identifying events. The sharp gradient near the coasts and international borders arises due to both Livneh and PRISM being masked outside of the CONUS and partially from the choice of the Epanechnikov kernel (see Section 2.2.3). The use of a globally-gridded dataset, such as a reanalysis product, would improve this aspect of the frequency analysis along the borders of the CONUS. However, reanalysis precipitation rates are largely model-derived and are

thus considered to have more uncertainty (e.g., Kalnay et al. 1996). Furthermore, the much larger horizontal grid spacing means these products will be unable to resolve small-scale features as accurately compared to Livneh and PRISM (Slinsky et al. 2019).

Total Number of Times in a 14-Day Extreme Precipitation Event (1915 - 2018)

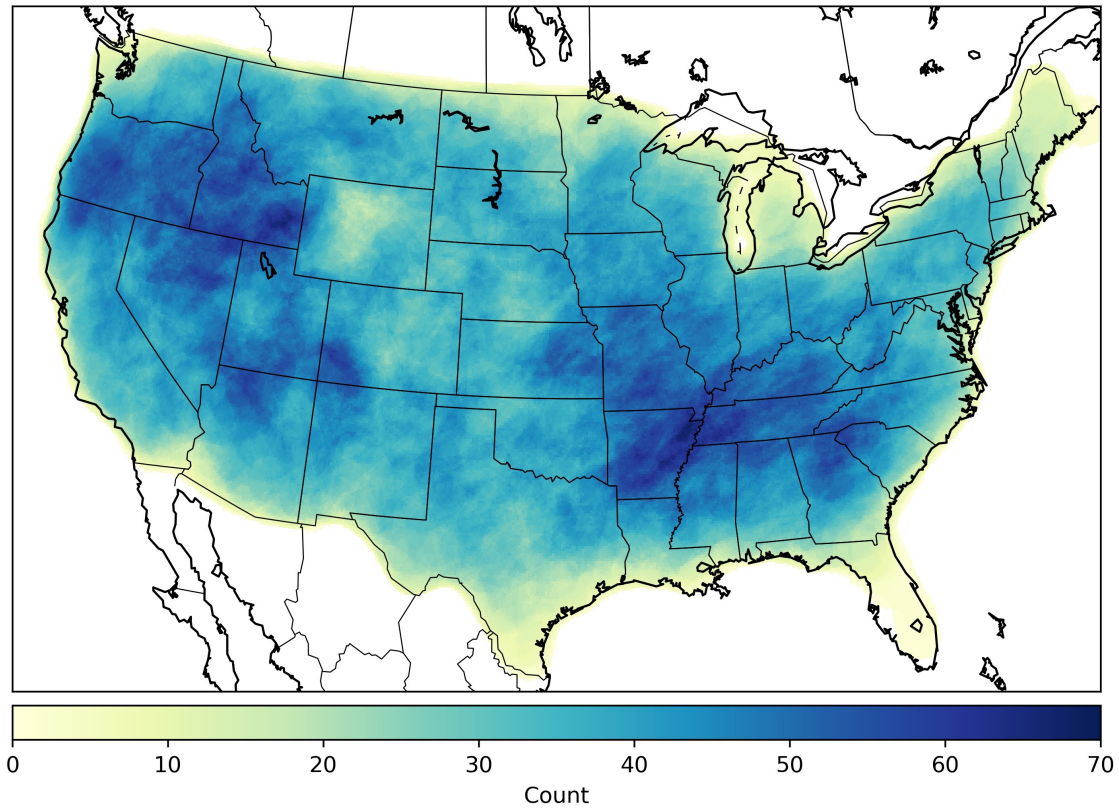


Figure 2.13: Total number of times a grid point is within an extreme period polygon over 1915–2018.

The increased frequency of grid points CONUS-wide being in extreme period polygons is fueled largely by increasing trends in the High Plains, northern Nevada, southern Michigan, and the majority of New England (Figure 2.14). There are two main regions of decreasing trends: southern California and Nevada, and Wyoming. Trends here were calculated using ordinary linear regression as bootstrapping QR trends to assess statistical significance was too computationally expensive when considering a spatial field. Bootstrapping was performed by randomly generating 10 000 new time series, sampling with

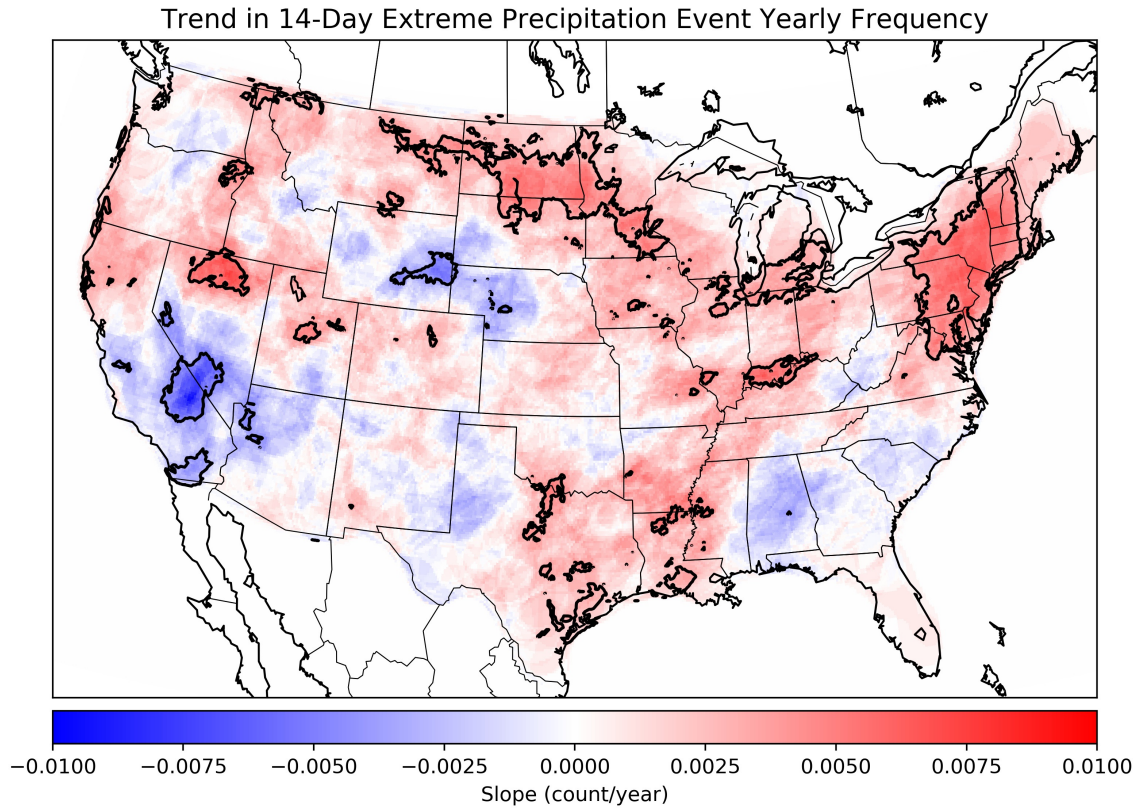


Figure 2.14: Trend (number per year; shading) in yearly frequency of grid points falling inside 14-day extreme period polygons using ordinary linear regression. Black contour delineates statistical significance of the slope at $p < 0.05$ using bootstrapping.

replacement, forming a distribution of slopes using ordinary linear regression, and comparing the point estimate contoured in Fig. 2.14 to the distribution. Roughly 9.8% of the grid points over the CONUS have significant trends (8.8% positive, 1.0% negative).

2.5 Summary and conclusions

The present research is motivated by a dearth of published research examining the characteristics of extreme precipitation with timescales beyond about 7 days. Furthermore, a large number of different definitions of “extreme” exist in the literature. For these reasons, I developed an algorithm for identifying 14-day extreme precipitation events. The algorithm is designed to be easily adaptable within the S2S timescale (e.g., 21-day, 30-day, 60-day). These databases are unique catalogs of extreme precipitation events in that they are identifying S2S extremes as opposed to short-term extremes within an S2S window. Grid points are designated as being extreme if they satisfy two criteria. The first criterion involves accumulation: a grid point must exceed the 99th percentile over the window being considered (e.g., 14 days). The second criterion involves duration: a grid point must record greater than normal daily precipitation on at least half of the number of days in a specified time window (e.g., 14 days). Average daily precipitation is calculated by aggregating all days in a window between 1915 and 2011 in the Livneh dataset and calculating a mean. The combination of these two criteria ensures the database identifies periods throughout the CONUS during which anomalously heavy precipitation occurs for more than a few days. Thus, the catalogued extreme periods likely have more significant, persistent global and synoptic forcings, which may be leveraged to provide forecasts of opportunity for extreme precipitation on the S2S timescale.

Grid points labeled as extreme are grouped into an extreme region using KDE. A 3-dimensional KDE PDF is fit over the CONUS, with 0.1° resolution, where a larger density (normalized by the maximum value) denotes more points labeled as extreme are in close proximity. An extreme event polygon is developed from the density field using the 0.2710 contour level. If such a polygon is greater than 200,000 km² in areal extent, it is identified as an extreme period.

I apply the above criteria to build a database of 14-day extreme precipitation events between 1915 and 2018. In any given year, the typical number of events CONUS-wide is 8.

However, the typical number of events is increasing throughout the period of record (Figs. 2.10 and 2.11), particularly in the Dakotas and throughout much of the Northeast (Fig. 2.14). This increasing trend is consistent with other literature that has found increases in extreme precipitation in the CONUS (e.g., Kunkel et al. 1999; Groisman et al. 2004; Hoerling et al. 2016; Stegall and Kunkel 2019; Kirchmeier-Young and Zhang 2020). However, caution should be heeded since many observational-based precipitation datasets do not homogenize station records, which may affect the trends in the datasets. Nevertheless, since there are well-documented changes in extreme precipitation, future works may explore the possibility of detrending before identifying extreme precipitation events.

I also applied *k*-means clustering to objectively divide the CONUS into regions. Although the masking of data in Livneh and PRISM outside the CONUS presents a challenge when identifying and clustering events, the majority of the CONUS is described by the clusters, and these clusters have fewer boundary effects compared to defining regions *a priori*. Every cluster has events in at least 9 months of the year, and many clusters qualitatively show little to no seasonality since the thresholds themselves follow the climatological cycle. Developing the regions is key not only to identify groups of events that are similarly shaped and positioned geographically but also to isolate events that are exceptionally anomalous in that they are shaped very differently. The next step is to diagnose the large-scale characteristics that force extreme events within each cluster and compare and contrast patterns among regions in a manner similar to Jennrich et al. (2020). This approach to objectively define regions may better isolate the dynamical signal within the extreme events. Defining regions before identifying events ignores the processes that cause the event, whereas my method is more likely to group events that have similar large-scale characteristics. The clusters are also more representative of the atmosphere as many latitude/longitude bounds used to delineate regions were removed, though observationally-based databases are constrained by the CONUS geopolitical boundary. Overall, my approach provides a strong foundation to diagnose the atmospheric drivers across the CONUS, both in the wet and dry seasons, while also allowing us to isolate the physical reasons why unclustered events

were atypical. Despite the continued constraint of the CONUS boundary, the orientation of the clusters follows physical intuition in many cases. For instance, many of the clusters in the Southern Plains into the Ohio River Valley and the Northeast have their major axis in a southwest-northeast direction, consistent with upstream longwave troughing and aligned with the North American storm track.

Examination of such large-scale characteristics should be undertaken using both observational and model sources. The following chapter outlines the changes needed to seamlessly apply my framework across multiple datasets. With these adaptations, additional databases have been developed in reanalysis datasets, such as ERA5 (Hersbach et al. 2020) and climate models, such as those in Coupled Model Intercomparison Project Phase 6 (CMIP6; Eyring et al. 2016). Furthermore, the database developed here can be used to calibrate operational seasonal forecasting models such as the North American Multi-Model Ensemble (Kirtman et al. 2014). Defining databases and diagnosing large-scale characteristics, and how they may change in the future, is important as it will lend insights into the potential predictability of extended-duration extreme precipitation and extend the current boundary in prediction of S2S extreme precipitation and perhaps other hazards.

Chapter 3

S2S Extreme Precipitation Periods: Extending Databases to Reanalysis and Model Datasets

3.1 Introduction

One of the key strengths in the framework presented in Chapter 2 is its ability to easily adapt to various spatiotemporal scales and datasets. Because the algorithm can be applied consistently between observations and models, databases can be developed for virtually any desired region and aspects such as the frequency and trend of S2S extreme precipitation can be thoroughly examined. Furthermore, the large-scale characteristics that precede and force extended-duration precipitation extremes can be quantified and compared between different regions and datasets. These dataset comparisons can be facilitated between two observational datasets, such as ERA5 and the North American Regional Reanalysis (Mesinger et al. 2006), between observations and models, and between several models. Diagnosing the three-dimensional dynamics and thermodynamics, in addition to analyzing the capabilities of models to reproduce these drivers, are important to gain a more comprehensive understanding of S2S extreme precipitation. Analyzing model biases in their representation of extreme precipitation will serve to identify the largest sources of errors and suggest to modelers where to focus efforts. Additionally, quantifying trends in the frequency of these drivers will yield insights into regional changes in extreme precipitation. Therefore, this chapter details the process to extend the previously described database generation algorithm using two example datasets, one reanalysis and one climate model. The

changes outlined here are shown using a 14-day database and would apply to other length extremes (e.g., 30 days).

3.2 Data

I use the ERA5 (Hersbach et al. 2020) reanalysis dataset whereby data had a horizontal grid resolution of 0.5° and hourly precipitation was summed to yield daily totals. Temporally, the ERA5 database was generated using 1950-2019. To demonstrate the algorithm's suitability to a climate model, I chose the historical run, 1850-2014, from the CNRM-CM6-1-HR model (Voldoire et al. 2019; Saint-Martin et al. 2020, hereafter shortened to CNRM) from the CMIP6 suite. Srivastava et al. (2020) showed that the CNRM-CM6-1 model was one of the best performing CMIP6 models in 5-day extreme precipitation over the CONUS in terms of root-mean square-error. Furthermore, Agel and Barlow (2020) found that CNRM's high-resolution counterpart was the best CMIP6 model in reproducing the circulation associated with daily precipitation extremes in the Northeast US. Like ERA5, the CNRM historical simulation was on a 0.5° horizontal grid. The spatial domain was chosen as 20°N , 55°N ; 130°W , 60°W to focus on the CONUS yet ensure that edge artifacts are minimized for any extreme regions affecting the CONUS. When assessing trends among the three databases, the first analysis used the entire period of record for each dataset whereas the second began the trend analysis in 1950.

3.3 Generating new databases

At a high-level, the database generation process can be broken down into the following steps:

1. Calculate accumulation and duration thresholds.

2. Perform KDE on all 14-day periods and calculate long-term 99th normalized density percentile K_{thresh} .
3. Calculate long-term 95th percentile of extreme period polygons area distribution A_{thresh} .
4. Retain all polygons whose area exceed A_{thresh} .
5. Run postprocessing algorithm.

Thresholds for the algorithm's 2 precipitation-related criteria were defined using the long-term statistics of the respective dataset, in which the leading 3 Fourier harmonics were again used to smooth the accumulation criterion. Similarly, K_{thresh} and A_{thresh} were calculated for each dataset and are shown in Table 3.1. The two main adaptations needed to the algorithm come when performing KDE in Step 2: (a) the bandwidth parameter in the KDE algorithm, and (b) the horizontal grid onto which the KDE density is fit.

Table 3.1: Values developed for each dataset during the database generation process.

Dataset	K_{thresh}	A_{thresh} (km ²)
Livneh	0.2710	200 000
ERA5	0.3085	200 000
CNRM	0.3997	275 000

The bandwidth parameter physically controls how far away any given grid point can be from an extreme grid point and still be assigned nonzero weight. As the native grid spacing increases, a larger bandwidth is needed to consistently yield a physically-informed PDF, a task that becomes increasingly difficult as the number of grid points which can be labelled extreme becomes fewer and fewer. Nevertheless, the following linear bandwidth scaling is recommended as a function of the dataset's grid spacing:

$$h(\Delta x) = \begin{cases} 0.02 & \Delta x < 0.5^\circ \\ 0.025 & 0.5^\circ \leq \Delta x < 1^\circ \\ 0.03 & 1^\circ \leq \Delta x < 1.5^\circ \\ 0.035 & 1.5^\circ \leq \Delta x < 2^\circ \\ 0.04 & 2^\circ \leq \Delta x < 2.5^\circ \\ 0.045 & 2.5^\circ \leq \Delta x < 3^\circ \\ 0.05 & 3^\circ \leq \Delta x < 3.5^\circ \\ 0.06 & \Delta x \geq 3.5 \end{cases} \quad (3.1)$$

where Δx is the spacing, in degrees, between latitude and longitude points. Accordingly, both new databases were generated using $h = 0.025$.

However, beginning around $\Delta x \geq 1.5$, physically interpreting the extreme period polygons becomes increasingly difficult. As the size of the grid box grows, the PDF produced by KDE begins to resemble the chosen kernel moreso than extreme precipitation occurrence. Accordingly, future works may opt to adjust the KDE step with two potential ideas briefly presented here. In one approach, a technique like kriging (Cressie 1990) can transform the binary, low-resolution grid of extreme points to a higher resolution grid. Kriging uses covariances to interpolate to the finer grid and factors affecting precipitation such as orography can be accounted for, ensuring the transformation between grids is at least somewhat physically-informed. Kriging is unbiased and minimizes the error variance and is thus similar to optimal interpolation schemes frequently used (e.g., Gichamo and Draper 2022). The new grid of extreme points would no longer be binary and KDE weighted by the point's kriged value may produce more intuitively shaped extreme periods. A second approach drops KDE altogether, instead using a clustering algorithm known as density-based spatial clustering with noise (DBSCAN, Ester et al. 1996). DBSCAN, unlike the clustering methods discussed in Chapter 2, does not require the number of clusters as input, instead

developing clusters of high-density regions (i.e., many data points close to each other) and categorizing any other points as outliers. There are two hyperparameters in the DBSCAN algorithm: (1) the distance between two data points to be considered “close”, and (2) the total number of data points needed to form a cluster. The distance metric could be considered a third hyperparameter as popular implementations have a large number of potential choices, however the Euclidean or Haversine distances are the most relevant for this application. After setting an appropriate distance based on the dataset’s native grid spacing, the extreme period polygon(s) would be the convex hull(s) of the cluster(s) produced by DBSCAN.

The second, simpler change involving KDE involves the horizontal grid for the PDF. I chose to expand the domain grid by 4° , or 8 grid points, in both datasets used herein; in datasets using larger bandwidths, somewhere between 6-12 extra grid points may be needed to ensure extreme polygons are not subject to edge artifacts. The horizontal grid spacing used for these databases continued to be 0.1° , although researchers considering larger domains and/or using datasets with larger grid spacings may consider resolutions as large as 0.5° for the KDE grid. The remainder of the database generation algorithm process outlined above remains virtually the same as described in Chapter 2. For each dataset, K_{thresh} and A_{thresh} were calculated individually and the postprocessing algorithm produced an extreme precipitation period database.

Lastly, the extreme periods in the ERA5 and CNRM databases were regionalized into the Livneh-developed clusters. Although k -means clustering could be utilized to compare preferred regions in each dataset, the focus here is to compare and contrast extreme periods, both in terms of frequency and their large-scale characteristics, that are guaranteed to be in the same region as observationally-derived extreme periods. For each extreme period in a newly-developed database, congruence coefficients are calculated with respect to each average cluster depicted in Fig. 2.9. If no coefficients are greater than 0.25, the extreme period is unclustered; otherwise, the assigned cluster is that with the largest coefficient. This minimum value needed to cluster was determined subjectively by visually inspecting

several instances with coefficients between 0.15 and 0.35 and 0.25 was found to consistently yield the “best” cluster assignments. The remainder of this chapter is dedicated to comparing and contrasting the developed thresholds and databases for each dataset.

3.4 Climatological comparison

3.4.1 Criteria thresholds

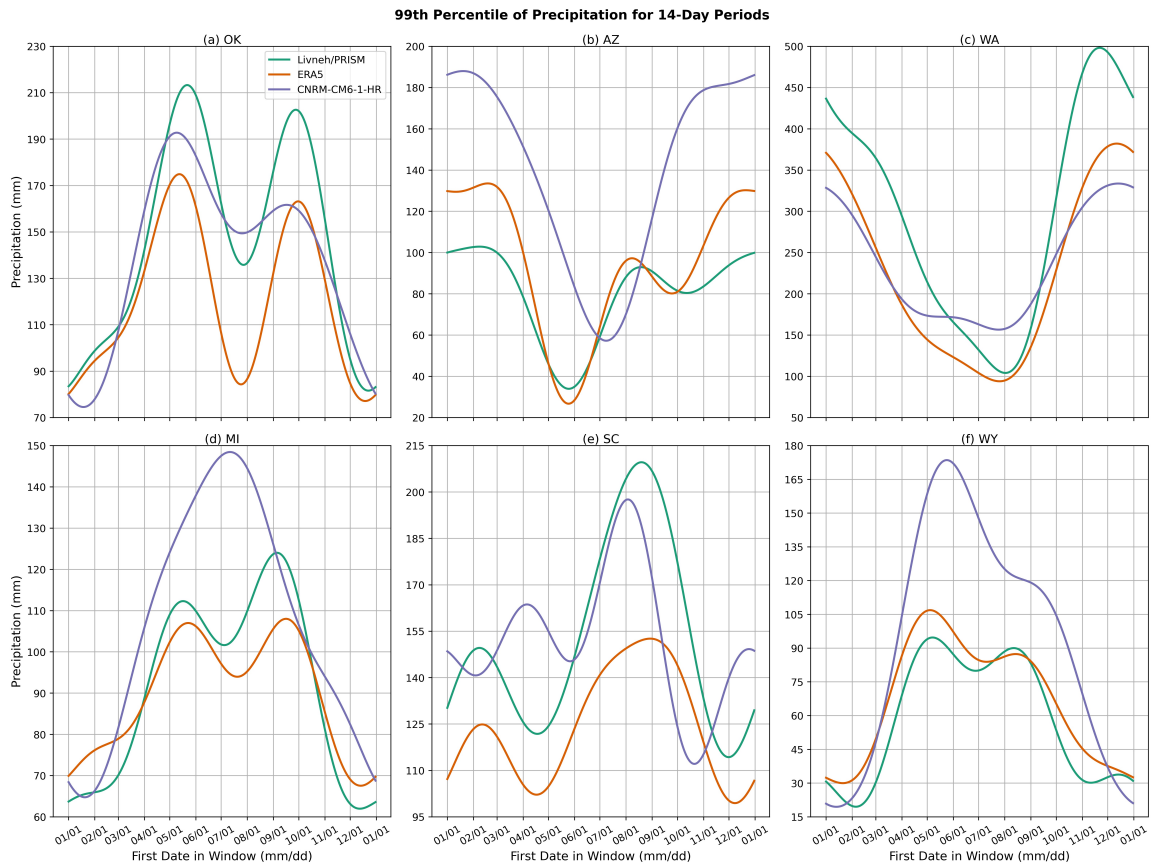


Figure 3.1: Smoothed 14-day 99th percentiles from the (green) Livneh dataset, (orange) ERA5 dataset, and (purple) CNRM-CM6-1-HR historical run using the first three Fourier harmonics for six selected grid points across the CONUS. The plotted date is the start date of each window.

Accumulation thresholds in ERA5 resemble observations (Livneh) at all locations shown in Chapter 2 whereas the CNRM thresholds are less consistent (Fig. 3.1). Note that these

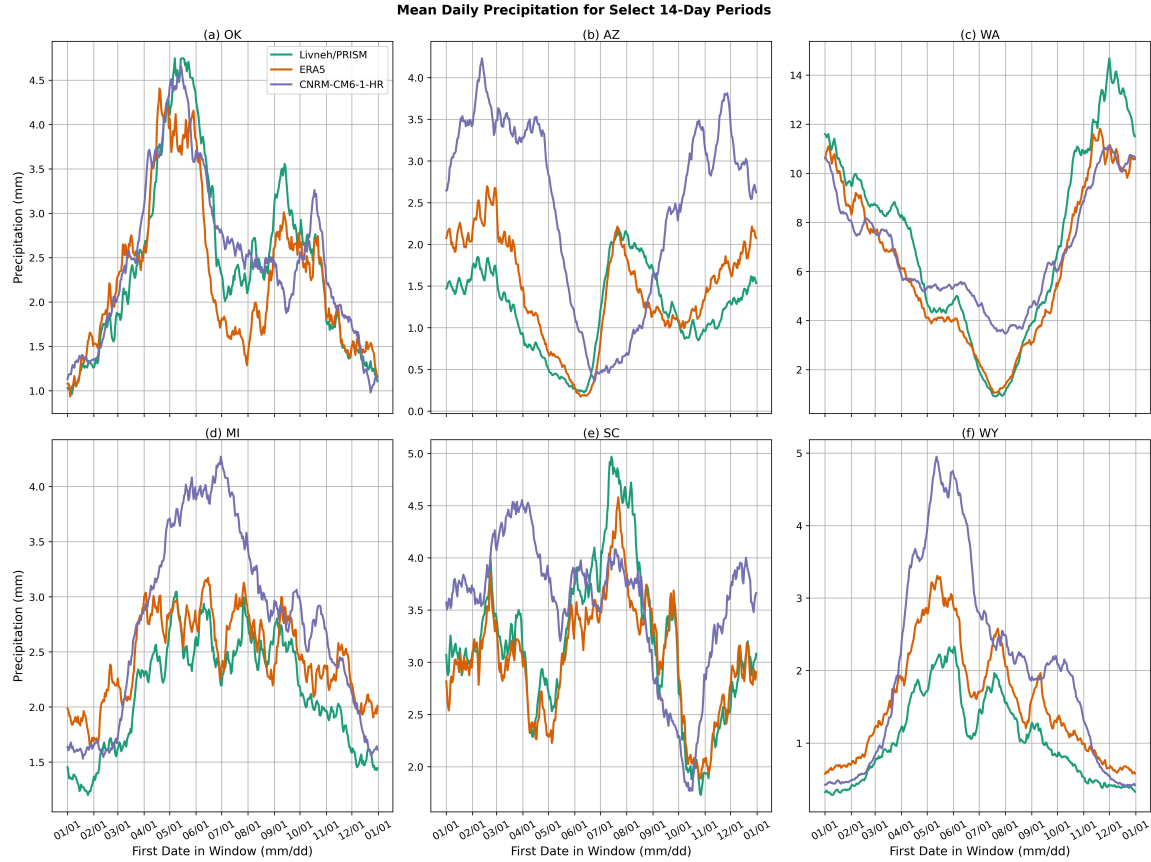


Figure 3.2: Mean daily precipitation from the (green) Livneh dataset, (orange) ERA5 dataset, and (purple) CNRM-CM6-1-HR historical run. The plotted date is the start date of each 14-day window used to calculate the long-term mean.

locations are single grid points within the referenced state. ERA5 thresholds have a systematic dry bias in OK, WA, SC, WY, and MI in boreal summer. In AZ, ERA5 matches Livneh in the summertime before being too wet in the winter. The CNRM climate model is also too wet in AZ in the wintertime, with accumulation thresholds over 80 mm greater than that in Livneh and 55 mm greater compared to ERA5 in much of December and January. CNRM indicates a phase error initially in AZ, being 45 days too late in the timing of the dry season with the minimum threshold biased $\sim 200\%$ high. CNRM thresholds are also biased high across much of the year in MI and WY, and has a phase offset compared to the other two datasets in late winter into early spring in SC. Finally, the CNRM thresholds are highly correlated with Livneh and ERA5 in OK and WY with Pearson correlations exceeding 0.9 in all cases except CNRM and ERA5 in OK ($r = 0.8$). The same overarching

themes emerge when examining the duration thresholds (Fig. 3.2) with errors on the order of 1 mm.

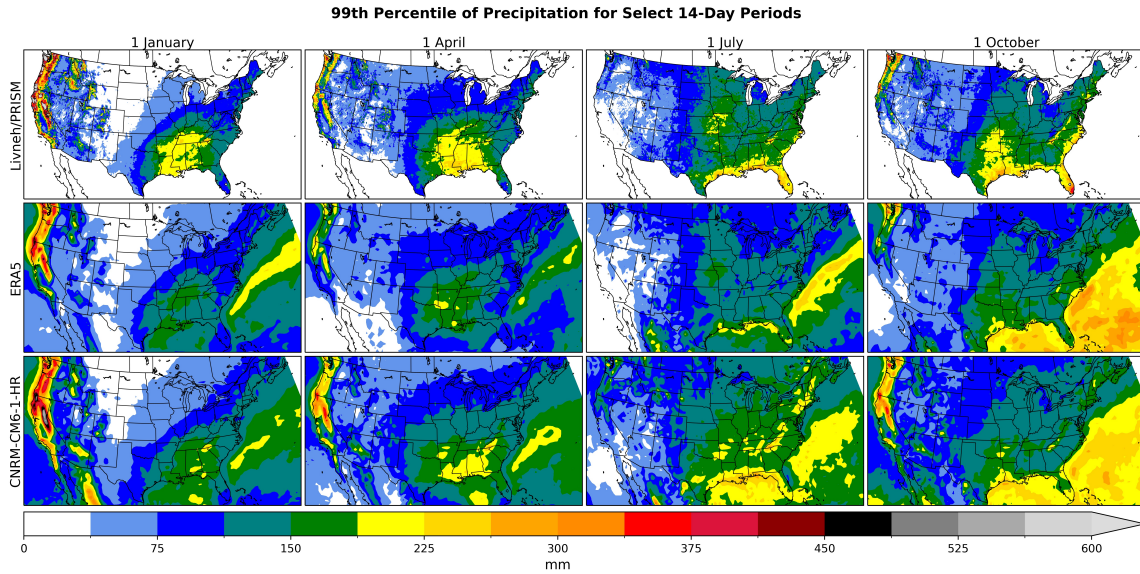


Figure 3.3: The 99th percentile of 14-day precipitation across the CONUS region from the (top row) Livneh dataset, (middle) ERA5 dataset, and (bottom) CNRM-CM6-1-HR historical run for one window in each meteorological season.

Spatially, the 99th percentiles of 2-wk total precipitation across the three datasets are fairly consistent (Fig. 3.3). The CNRM model successfully captures the higher totals in the Sierra Nevada Mountains well in January with totals closer to observations than ERA5 (Fig. 3.3, first column). However, thresholds in the CNRM dataset are too wet by 100 mm or more across these mountains in April and October. The CNRM is also closer to Livneh than ERA5 across the Southeast in April and July, particularly along the Gulf Coast. In the Midwest, however, CNRM is much too wet in the summertime, with biases as large as 50%, as the model struggles with simulating the correct annual climatology (Fig. 3.1d). ERA5 accumulation thresholds are visually well associated with the Livneh thresholds with a general dry bias, an expected result given its coarser grid spacing (e.g., Lopez et al. 2015; Slinskey et al. 2019). Accumulation thresholds between ERA5 and CNRM are also very similar as Pearson correlation values vary from 0.79 to 0.87 depending on the day of year.

3.4.2 Extreme period frequency

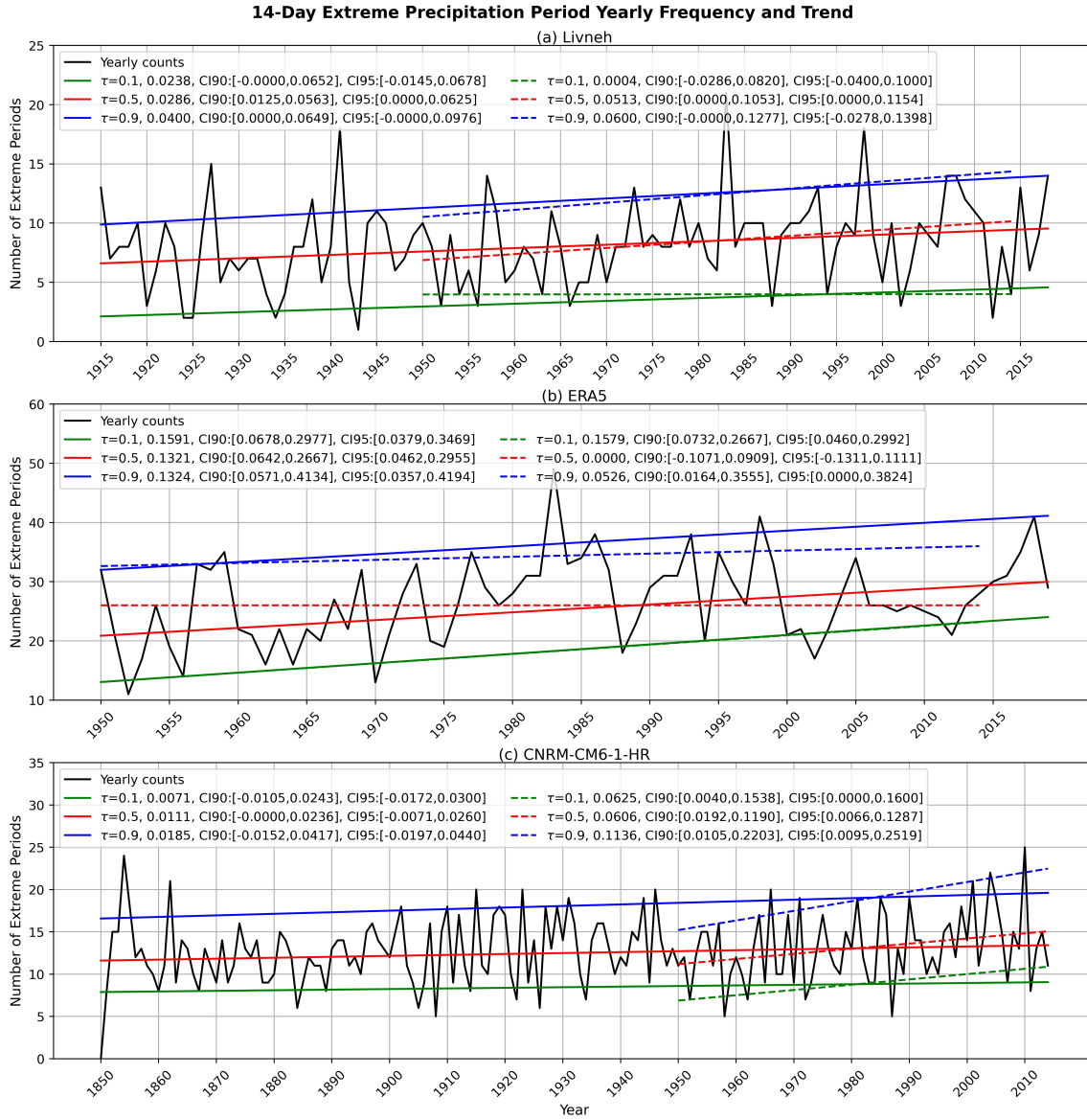


Figure 3.4: (black) Total yearly counts of recorded 14-day extreme precipitation periods domain-wide as a function of time in the (a) Livneh, (b) ERA5, and (c) CNRM-CM6-1-HR historical run 14-day databases. QR models are fit on (green) $\tau=0.1$, (red) $\tau=0.5$ (i.e., median), and (blue) $\tau=0.9$ quantiles of the time series where solid lines were fit on the entire time series and dashed lines were fit using 1950–2014. Slopes are listed as the second number in the legend, and the 90% and 95% confidence intervals on the slope are shown in brackets thereafter. Confidence intervals were calculated using BCa bootstrapping. Slopes and confidence interval bounds were rounded to 4 decimal places.

Domain wide, roughly 25 extreme periods occur per year in the ERA5 database (Fig. 3.4b) whereas the CNRM database has about half as many occurring on a typical year

(Fig. 3.4c). The ERA5 mean frequency is relatively uninformative, though, as there is a statistically significant trend (i.e., rejecting a null hypothesis of no trend) of about 1.3 periods per decade. The upper and lower quantiles are also significantly increasing at roughly the same rate over 1950–2019. A permutation test on the mean frequency between the first and second halves of the ERA5 time series yields $p \approx 0.065$, suggesting the means from the two halves may be drawn from different distributions. Further, a permutation test on the first and last quarters of the ERA5 time series gives $p \approx 0.012$, demonstrating an increasing trend in the frequency of extreme precipitation periods between 1950 and 2019. Conversely, domain-wide trends in CNRM are all statistically indistinguishable from 0 ($p > 0.10$), although the point estimates are all positive and the median confidence interval is very close to 0. A permutation test on the mean frequency applied to the first and last halves of the CNRM historical run gave $p \approx 0.24$, yet the same test on the first and last quarters (thirds) of the time series yield $p \approx 0.02$ ($p \approx 0.11$). Altogether, these results indicate that the distribution of the number of subseasonal extreme precipitation began to change around 1960 and further suggest an anthropogenic link.

When considering only those extreme periods clustered into the Livneh clusters, the mean number of periods per year drops to between 7 and 12 in ERA5 and to 5 in CNRM (Fig. 3.5b,c). These frequencies more closely resemble that seen in the Livneh database. ERA5 continues to depict significant increases in the median and high-end years, although trends in the frequency of low-end years are insignificant ($p > 0.10$) and all three slope magnitudes are an order of magnitude less compared to the full-domain. The magnitudes in the CNRM trends also decrease when considering clustered periods only, although these differences are smaller compared to that seen in ERA5. The change with the smallest magnitude in CNRM when going from the full domain to the CONUS occurs in the median and the trend becomes statistically significant at $0.05 < p < 0.10$. Contrarily, trends in upper and lower quantiles of yearly frequency remain statistically indifferent from 0 ($p > 0.1$). Permutation tests on the ERA5 time series generate p -values of 0.0972 and 0.0252 using the two halves and the first and last quarters, respectively, corroborating the increasing

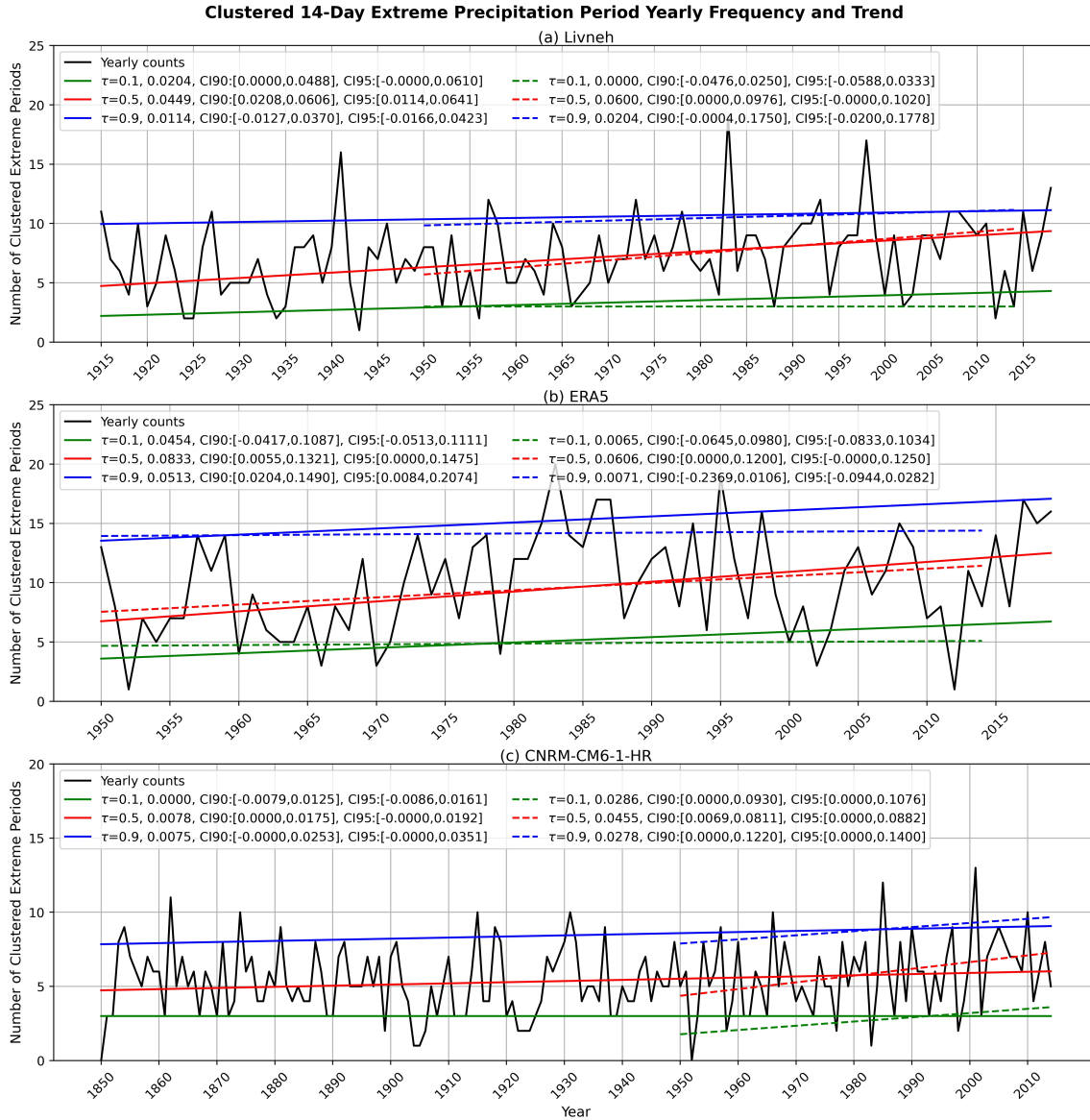


Figure 3.5: As in Fig. 3.4 but only for CONUS-clustered extreme periods.

trends found when using all extreme periods. However, whereas the p -value is unchanged in CNRM when splitting the time series in half, a permutation test using the first and last quarter gives $p \approx 0.14$.

In summary, the Livneh and ERA5 databases both show slight increasing trends in the frequency of subseasonal extreme precipitation across the CONUS during much of the 20th and into the 21st centuries whereas the CNRM historical run depicts little to no trend in these extreme periods until the latter half of the 20th century.

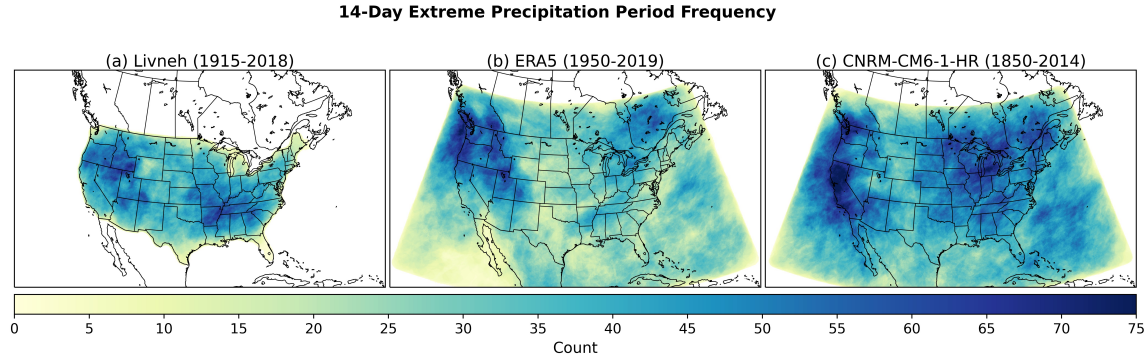


Figure 3.6: Total number of times inside an extreme period polygon from the (a) Livneh (1915–2018), (b) ERA5, and (c) CNRM-CM6-1-HR historical run databases.

The ERA5 and CNRM databases share some spatial similarities in 2-wk extreme period frequency to the observationally-based Livneh database (Fig. 3.6), although there are unique differences from each dataset. The range of frequencies among the different databases are similar despite very different time periods because of the many percentile-derived components (e.g., accumulation threshold, KDE contour, etc.), which allows for a direct comparison among datasets. All three datasets have relative maxima in the Northwest CONUS with the Livneh database maxima located roughly near Oregon and Idaho (Fig. 3.6a) whereas the two globally-gridded datasets have their maxima positioned closer to the Pacific Northwest coast (Fig. 3.6b,c). In ERA5, the highest counts are relatively constant along the Oregon and Washington borders with the Pacific Ocean, whereas the highest counts from CNRM are farther north and east. Moving south along the West Coast, the Livneh and ERA5 datasets maintain relatively large counts into North-Central California prior to decreasing in extreme period frequency. Conversely, CNRM has a second local maxima near the California Bay Area and there are much higher frequencies in southern California compared to Livneh and ERA5.

The second relative maximum in the Livneh database is over Arkansas, Missouri, and western Tennessee (Fig. 3.6a). There is a similar relative maximum in ERA5 in this region (Fig. 3.6b), although there are far fewer extreme periods compared to the Livneh database and its prominence over surrounding grid points is much less pronounced compared to the

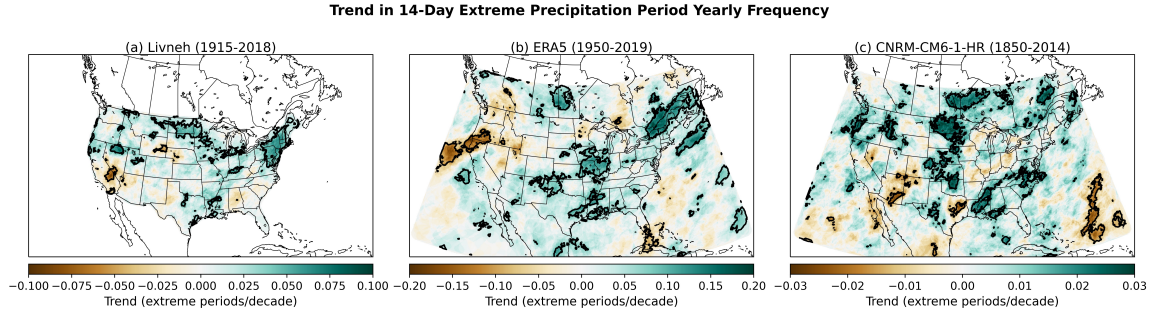


Figure 3.7: (shaded contours) Trend (number per decade) in yearly frequency of grid points falling inside extreme period polygons using ordinary linear regression in the (a) ERA5 and (b) CNRM-CM6-1-HR historical run 14-day databases. Black contour delineates statistical significance of the slope at $p < 0.05$ using bootstrapping with 5000 iterations.

Pacific Northwest. In ERA5, the most prominent maximum east of the Rocky Mountains is in Quebec, Canada with low frequencies across much of the Southeast and U.S. Great Plains. Accordingly, k -means clustering was not able to reproduce the Livneh clusters as the Central and North Plains were shunned in favor of several clusters over the Pacific Northwest (not shown). In contrast, the CNRM database has a clear second maximum east of the Rockies with relatively high and widespread counts across the Midwest and Great Lakes. Although it was not explicitly run, I hypothesize k -means would yield more spatially diverse clusters compared to ERA5.

There are more distinct differences among the three datasets in their spatial trends of 2-wk extreme periods. The main regions of positive trends in the Livneh database are in northern Nevada, the High Plains, regions in southern Michigan and along the Indiana-Kentucky border, and much of New England (Fig. 3.7a). Of these regions, only the High Plains increasing trend is reproduced with CNRM depicting a similar broad region of positive trends (Fig. 3.7c). In ERA5, trends in the Dakotas are small before flipping the sign to the east into Minnesota (Fig. 3.7b). In New England, the other relatively large region of increasing trends, both ERA5 and CNRM show little trend in 2-wk extreme precipitation along the Atlantic Coast, instead favoring the US-Canada border in New York and Vermont (Fig. 3.7b) or areas into Quebec (Fig. 3.7c). Small trends are also seen in northern Nevada and southern Michigan and Indiana. When considering decreasing trends, the three

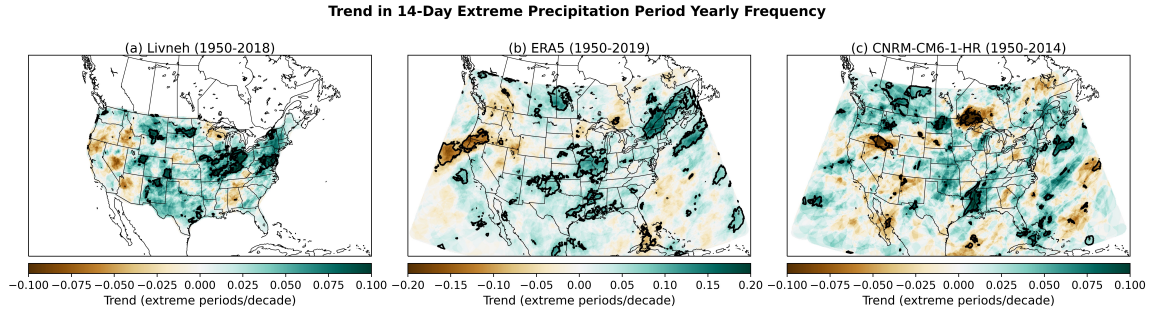


Figure 3.8: As in Fig. 3.7 but performing ordinary linear regression in the Livneh and CNRM databases beginning in 1950.

databases continue to show little agreement. In the Livneh database, decreasing trends are largest in southern Nevada and California with a secondary, significant region in eastern Wyoming (Fig. 3.7a). Neither ERA5 nor CNRM have strong trends in eastern Wyoming or southern California (Fig. 3.7b,c) while CNRM actually has a positive trend in southern Nevada. Similarly, ERA5 has a strong decreasing trend along the coast in northern California and southern Oregon whereas both Livneh and CNRM depict increasing trends in the same region. Pearson correlations between the trend maps for Livneh and ERA5 and CNRM are very small at 0.03 and 0.08, respectively. However, several of these major differences between the Livneh and ERA5 trends disappear when starting the analysis for Livneh in 1950. In particular, the Livneh database now depicts a decreasing trend along the Oregon coast and is also more aligned with ERA5 in Nevada (Fig. 3.8a,b). Using a smaller time range also accentuates the increasing trend throughout much of the Midwest and trends in this region are now in agreement with ERA5. Similarly, there are slightly greater magnitude slopes in upstate New York and Vermont (Fig. 3.8a), offering some corroboration of the strong positive trends in ERA5 seen along the US-Canada border. The correlation between Livneh and ERA5 trends for the 1950-onward period is 0.39, whereas the reduced temporal window caused the correlation with CNRM to rise slightly to 0.11. Across their full period of records, ERA5 and CNRM both have roughly 1.5% of grid points exhibiting significant decreasing trends compared to 7.4% and 5.6% for increasing trends, respectively. When beginning the trend analysis in 1950, the proportion of grid

points with significant positive (negative) trends decreases to 5.5% and 2.7% (0.1% and 1.1%) in Livneh and CNRM, respectively.

3.5 Summary and conclusions

This chapter focused on expanding the framework outlined in Chapter 2 to generate a database of subseasonal extreme precipitation periods. I chose two datasets, ERA5 reanalysis and the CNRM-CM6-1-HR historical run from the CMIP6 suite, to demonstrate the applicability of the algorithm. The most noteworthy change needed was the bandwidth parameter used in KDE and a relatively simple linear scaling as a function of grid spacing was proposed. Although the KDE approach is effective when using datasets with relatively high resolution, it becomes less so as the native grid becomes coarser and two alternatives were proposed to use in place of KDE. Nevertheless, the KDE approach for developing extreme period polygons worked well for ERA5 and CNRM on 14-day periods whereas virtually all other parts of the algorithm were unchanged, a key factor in the development of the framework from the beginning. Extreme periods in both new databases were assigned into Livneh clusters to obtain CONUS-only polygons and adequately compare extreme periods occurring in the same geospatial region.

The developed thresholds between the three datasets were relatively consistent overall with bigger differences between Livneh and CNRM compared to ERA5. Both accumulation and duration thresholds in ERA5 were well associated with Livneh's thresholds (Figs. 3.1 and 3.2) with a tendency to be too dry owing to its coarser resolution. Thresholds in CNRM, however, were less systematic in the errors compared to Livneh, although biases were more often wet than dry, consistent with Srivastava et al. (2020). In AZ and MI, CNRM does not simulate the correct seasonal precipitation cycle as the dry season in the former grid point was late by about 30-45 days whereas accumulations thresholds in MI have one summertime peak rather than intermediary peaks in the spring and fall and a relative minimum in the peak warm season (Fig. 3.1b,d). CNRM also had a phase offset in SC,

particularly in the springtime (Fig. 3.1e). Spatially, accumulation thresholds in both ERA5 and CNRM resemble the overall structure seen in observations well. CNRM simulated precipitation totals particularly well in the Sierra Nevada mountains in the winter (Fig. 3.3) as the 99th percentile was closer to observations than ERA5. However, CNRM overestimates the thresholds in this region in the spring and fall whereas the dry bias in ERA5 is relatively constant. CNRM accurately simulated precipitation in the Southeast US, especially in the summertime, although this may be a result of mischaracterizing the seasonal cycle, especially into the Midwest where thresholds are far wetter compared to Livneh and ERA5 (Fig. 3.3, third column).

There were greater differences among the three databases in the frequency and trends of 2-week extreme precipitation periods. ERA5, developed using the shortest temporal period, had the highest number of extreme periods occurring yearly across North America with a typical year having around 25 documented periods (Fig. 3.4b), although this number dropped to around 10 when considering polygons primarily over the CONUS (Fig. 3.5b). The ERA5 database, in agreement with the Livneh database, depicted significant positive trends in the yearly frequency of subseasonal extreme precipitation periods (Figs. 3.4a,b and 3.5a,b). Conversely, the CNRM historical run did not indicate the presence of any significant trends (Figs. 3.4c and 3.5c).

The spatial frequency of 14-day extreme precipitation periods was also shown to be quite different among the 3 databases. All 3 datasets generally depict a local maximum in the Northwest U.S., although the exact center changes from database to database (Fig. 3.6). In the Livneh database, the maxima was centered near northern Nevada and southern Idaho whereas the ERA5 database peaked to the northwest along Washington coast and CNRM had relatively high counts along much of the central and northern West Coast. East of the Rocky Mountains, the ERA5 appeared more dissimilar compared to the other two databases as it struggled to produce extreme periods across the Plains and parts of the Midwest. The ERA5 and CNRM databases also failed to replicate the observed local trends in extreme period yearly frequency (Fig. 3.7a,b) with virtually no association between the

maps. However, there are some notable areas of overlap, particularly between Livneh and CNRM, in the Northern Plains and in the Arkansas/Mississippi/Missouri region.

Extending this process to CMIP6 models is particularly advantageous because a large number of potential extreme periods can be generated in a myriad of different climates. Agel and Barlow (2020) reported that CMIP6 models can realistically simulate the large-scale circulations forcing extreme precipitation despite having errors in the seasonality and/or distribution of precipitation. Accordingly, databases built using CMIP6 simulations can be leveraged to increase the sample size of extreme periods and gain a more complete understanding of subseasonal extreme precipitation. As such, the following chapter aims to quantify the large-scale characteristics driving subseasonal extreme precipitation in the CONUS and provide a groundwork for developing a predictive model to increase the predictability of these extreme periods.

Chapter 4

Synoptic- and Hemispheric-Scale Characteristics of Subseasonal Extreme Precipitation Periods in the United States

4.1 Introduction

Although the definition for “extreme” precipitation is often inconsistent between studies, distinct large-scale patterns that force the extreme event have emerged. Maddox et al. (1979) was the first study to analyze the mechanisms driving flash flooding in the United States. They found the archetypical pattern to have a mid-level shortwave trough upstream of the location of maximum precipitation. In addition, a quasi-stationary surface boundary and a relatively moist troposphere were also important to maintain and support daily and sub-daily flooding. Similarly, Konrad (2001) showed 500 hPa cyclone centers were commonly observed upstream from extreme precipitation regions when considering 2-day extreme events. On a longer timescale, Flanagan et al. (2018) examined atmospheric patterns associated with pluvial years in the Great Plains, defined as years whose calendar-year total precipitation was 10% greater than the climatological annual total precipitation. During pluvial years, the southern Great Plains was associated with negative height anomalies over the southwestern United States, whereas the northern Great Plains had anomalous troughing centered closer to the northwestern CONUS. In both regions, a key driver for rainfall extremes within pluvial years was anomalously strong moisture fluxes with contributions from both the Atlantic and Pacific Oceans. Zhao et al. (2017) performed cluster analysis on daily extremes and showed a similar anomalous Rossby wave train to be important for daily

precipitation extremes in the northern Plains; wave trains were also seen for regions such as Midwest and across an area resembling my cluster 12 (Fig. 2.9). On a study focused on the Northeast US, Agel et al. (2018) clustered 500 hPa geopotential height anomalies during extreme precipitation days and found troughing upstream of the region present in four of six clusters.

Upstream troughing is the most common mid- and upper-level feature present during extreme precipitation events because it typically promotes quasi-geostrophic forcing for synoptic-scale ascent (Bluestein 1992; Milrad et al. 2014). Numerous studies have shown that these large-scale patterns are not wholly unique to extreme precipitation days (Agel et al. 2015; Agel and Barlow 2020; Schroers and Martin 2022), although the magnitudes of the anomalies are typically deeper on extreme days. Short-duration precipitation extremes occur when there is a combination of synoptic and mesoscale forcings (e.g., fronts Barlow et al. 2019). Anomalous downstream ridging is an especially important factor in forcing favorable environments for extended-duration extreme precipitation. These ridges can form blocking patterns which can aid in the development of, and reinforce, upstream troughs and promote synoptic-scale ascent for longer periods of time. Accordingly, blocks are frequently observed during extreme events (Kautz et al. 2022). Furthermore, high pressure systems can be vital in transporting and maintaining anomalous moisture into the region. For instance, The Bermuda High has been shown to be a key driver for extreme precipitation occurring in the Southeast US through its modulation of the low-level jet and ensuing moisture advection (Keim 1996; Moore et al. 2012, 2015).

In this Chapter, I briefly quantify the synoptic-scale characteristics during subseasonal extreme precipitation days in the Livneh and CNRM databases. Specifically, I examine mid-level geopotential heights, mid-level vertical velocity, low- and upper-level winds, and specific humidity and compare with results shown in Jennrich et al. (2020) and Schroers (2020).

4.2 Data

The ERA5 (Hersbach et al. 2020) reanalysis dataset was used to composite the large-scale fields present during extreme periods in the Livneh database beginning in 1950. Composites were also made for simulated extreme periods using the CNRM CMIP6 historical run database. Each variable was standardized relative to its 1950–2014 climatology; using a moving 30-year climatology resulted in nearly identical composites. Statistical significance at $p < 0.05$, represented as hatching on all figures, was calculated by bootstrapping random 14-day periods with replacement for a total of 5000 iterations. A subset of patterns are shown here for brevity and to compare to the regions defined by (Schroers 2020); the following clusters were chosen as being the closest matches:

- West Coast (WC): cluster 3
- Mountain West (MW): cluster 2
- North Plains (NP): cluster 7
- South Plains (SP): cluster 9
- Great Lakes (GL): cluster 11
- Southeast (SE): cluster 14
- Northeast (NE): cluster 15

4.3 Results

Each cluster is characterized by an anomalous trough-ridge dipole, although the orientation of the dipole is not constant. The WC region has the deepest trough among all clusters positioned near the California-Oregon border (Fig. 4.1a). The trough is flanked by ridging in the Aleutian Islands to the northwest and a second region stretching from Hawaii though

Baja California and into the central CONUS. There are also broad regions of troughing off eastern Asia and across the western Atlantic. Overall, the WC composite is consistent with Jennrich et al. (2020). Extreme periods in the CNRM are driven, in the midlatitudes, by a very similar pattern as depicted by ERA5 (Fig. 4.2a), although the ridge to the south of the WC trough is less expansive and slightly weaker. In the MW, an anomalous wavenumber 4 Rossby wave train is seen across the hemisphere with significance in nearly every node (Fig. 4.1b). Broad troughing is in place across nearly the entire western third of the CONUS with strong Alaskan ridging upstream and moderate ridging across the Southeast US. The CNRM model reproduces part of the wave train, particularly the nodes closest to the MW (Fig. 4.2b), although the influence of trend may be having the biggest impact on this region compared. There is also some agreement with Schroers (2020), especially over the CONUS, however their MW composites did not depict the other nodes of the wave train.

Following the work of Flanagan et al. (2018) and recommendations from Jennrich et al. (2020), Schroers (2020) split the Plains into separate northern and southern regions along the Kansas-Nebraska border. The results here are in broad agreement with Flanagan et al. (2018) in that there are unique drivers for the two regions. In the NP, the composite depicts a wave train moreso than seen in the SP. Both regions have anomalous troughs over the Rocky Mountains with the NP trough centered over Utah whereas SP periods have troughs that dig further south (Fig. 4.1c,d). In both regions, anomalous ridging extends from northern Canada into the Southeast US. The NP composite here matches the findings of Schroers (2020) despite their NP region having a relatively small sample size with the biggest difference being the Atlantic trough here being positioned slightly further north. The PL composite in Jennrich et al. (2020) is most similar to the SP composite. CNRM also has similar patterns in both regions (Fig. 4.2c,d), although the SP trough is positioned further south and the downstream ridge is stronger than seen in reanalysis.

The GL and SE regions are both driven by trough-ridge dipoles that are positioned in a northwest-southeast orientation which, like the Great Plains, are different primarily in their

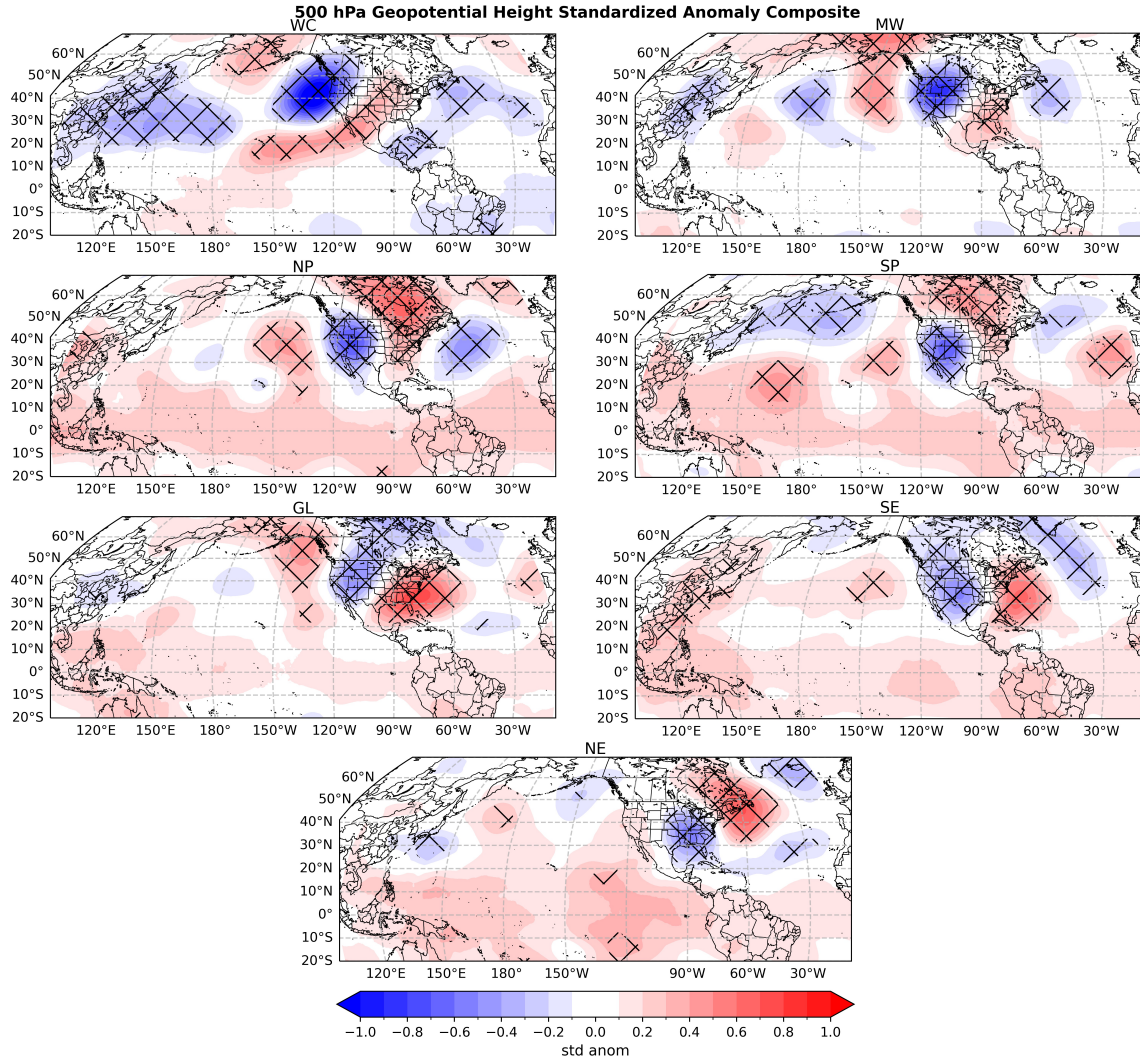


Figure 4.1: (shading) 500 hPa standardized geopotential height anomaly composite from subseasonal extreme precipitation days in the Livneh database across various regions of the CONUS. Hatching denotes statistical significance at $p < 0.05$ after bootstrapping.

exact positioning (Fig. 4.1e,f). In the GL, the main centers of action feature an anomalous trough in the Rocky Mountains into the NP and Canada sandwiched by ridging in southern Alaska and across much of the SE and western Atlantic (Fig. 4.1e). Compared to Jennrich et al. (2020), the trough seen here is marginally further southwest whereas the two anomalous ridges are nearly identical. The biggest difference in composites between ERA5 and CNRM are seen in this region as the upstream trough extends too far northwest in CNRM and the Alaskan ridging is completely missed (Fig. 4.2e). For SE extreme periods, the anomalous trough pushes further southeast and, accordingly, the downstream

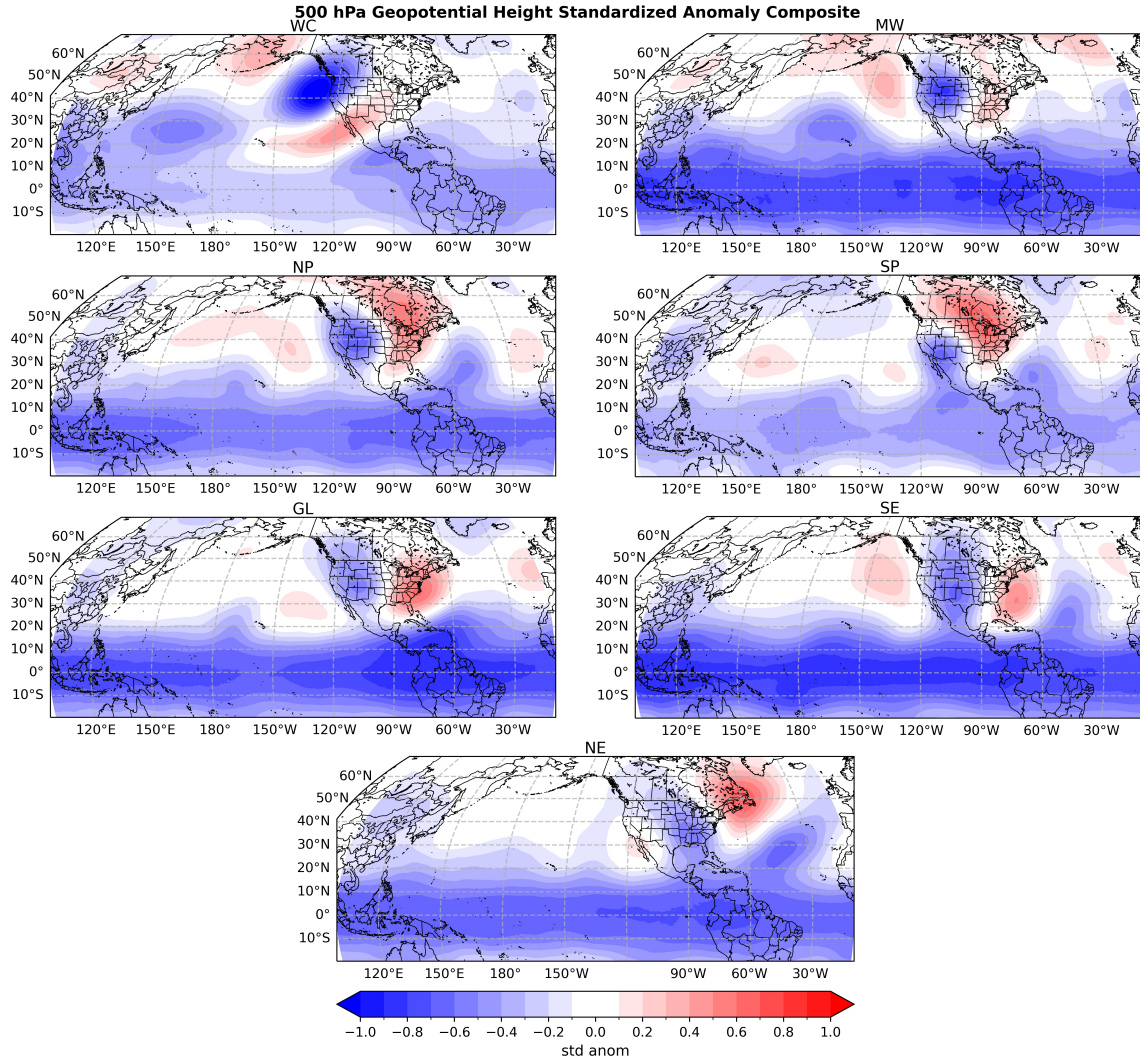


Figure 4.2: As in Fig. 4.1 but using CNRM simulated periods.

ridge is centered off the East Coast. There is also an anomalous ridge in the Central Pacific with the center being about 7° further north than that in Jennrich et al. (2020). The ridge is even further north in simulated periods by CNRM, and the ridge in the Western Atlantic is stronger than seen in observed extreme periods (Fig. 4.2f). Finally, the NE region features a southwest-northeast oriented trough-ridge dipole and is well depicted in both observations and simulations (Figs. 4.1g, 4.2g). Although this main dipole is consistent with past work, the remote sources are different using my NE cluster. There is anomalous troughing near the Bering Strait in Schroers (2020) during NE periods whereas Figure 4.3g shows

this trough positioned further east. Similarly, troughing is seen further north here with significant anomalies centered over Iceland. The CNRM model simulates much of the dipole correctly, although the ridge extends northeast into Greenland (Fig. 4.2g) instead of across the Hudson Bay as in ERA5.

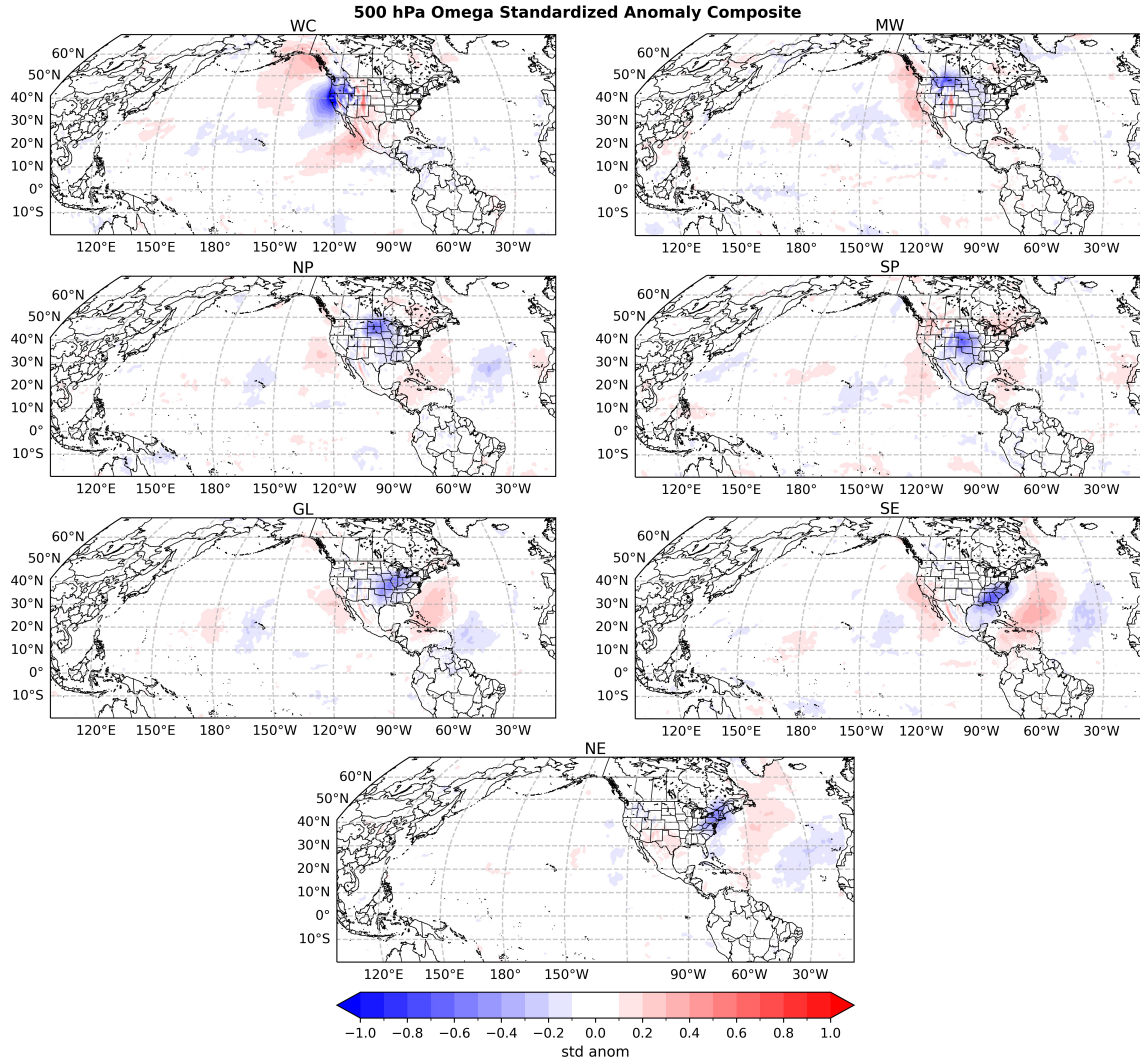


Figure 4.3: As in Fig. 4.1 but for 500 hPa vertical velocity.

Across all regions, the trough-ridge dipole favors synoptic-scale forcing for ascent in a quasi-geostrophic framework and, accordingly, anomalous ascent is seen in each region during extreme periods (Figs. 4.3 and 4.4). Ascent is strongest in the WC (Fig. 4.3a) with large anomalies along nearly the entire coast. There are also relatively strong signals for ascent in the GL, SE, and NE regions. The impact of the anomalous ridges is seen across the

domain as there are significant regions of descent seen across all regions. In the WC, MW, and Plains clusters, descent is strongest in the upstream ridge, whereas downward motions are stronger downstream of SE and NE regions. There are also significant remote signals in several regions. The SP and GL, and to some extent the SE, regions have anomalous sinking motion near the dateline in the subtropical Pacific, and the MW has a similar feature positioned further west. The NE region has anomalous descent on the Equator in the West Central Pacific (Fig. 4.3g). This is co-located with negative outgoing longwave radiation anomalies north of the Solomon Islands and positive anomalies over Indonesia, consistent with phase 4 of the MJO (not shown).

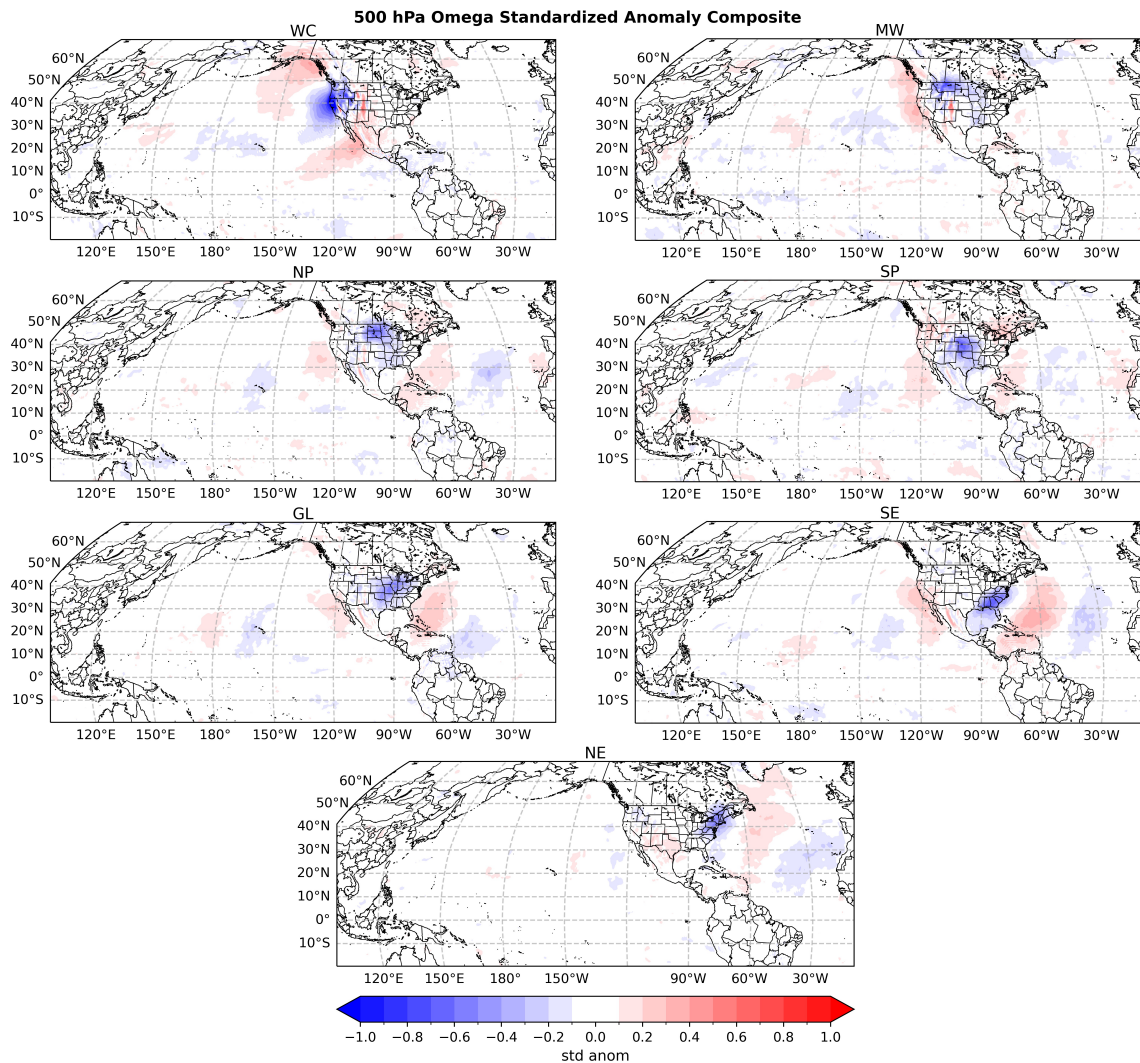


Figure 4.4: As in Fig. 4.3 but using CNRM simulated periods.

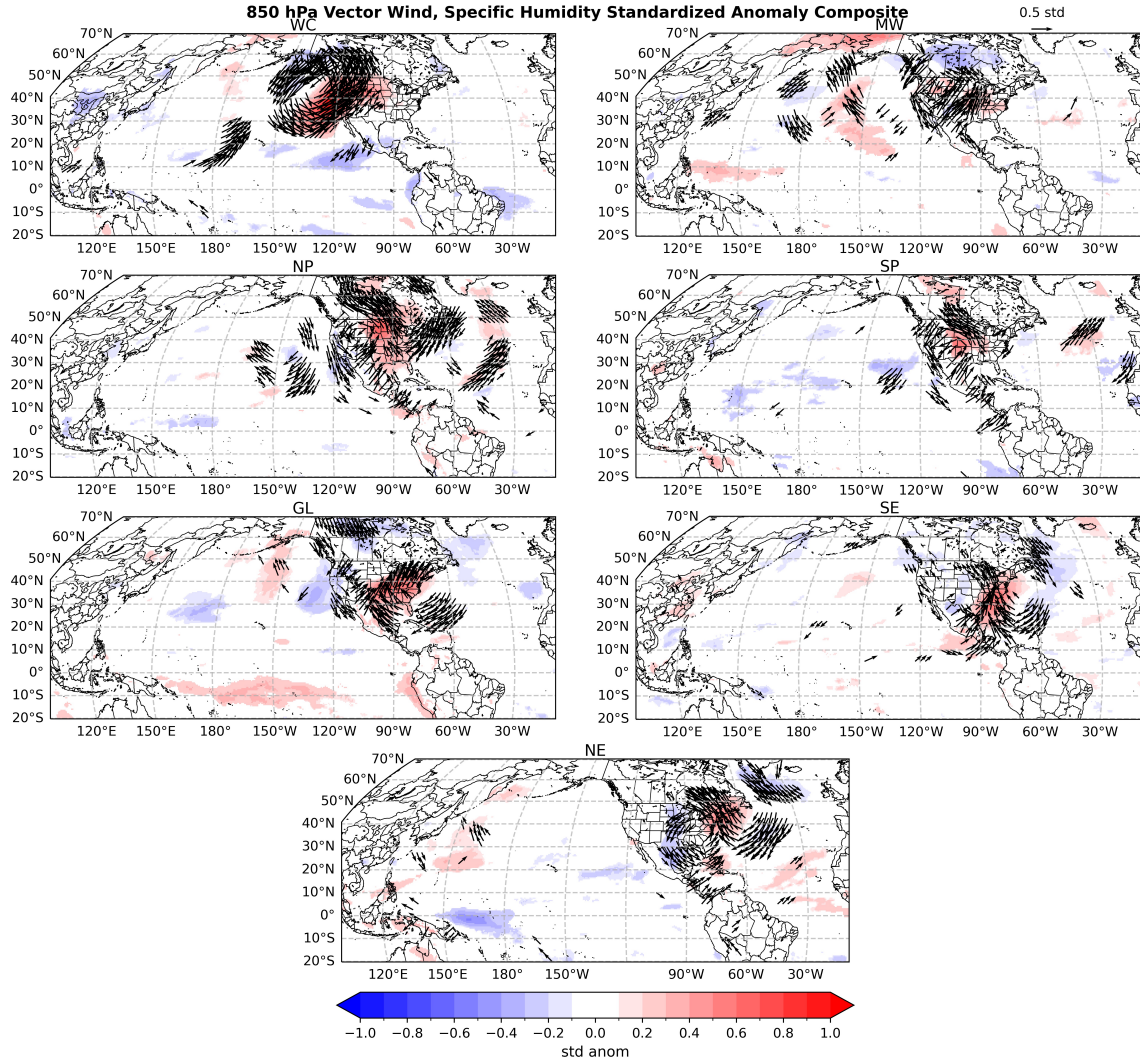


Figure 4.5: Composite of (shading) 850 hPa specific humidity and (vector) wind standardized anomalies. Only significant regions of specific humidity and vectors are shown.

Moisture advection is important to initiate and maintain an extended-duration precipitation extreme. At lower levels, specific humidity anomalies are stronger and centered on all regions (Fig. 4.5). On the WC, onshore flow is exceptionally strong and nearly entire western third of the CONUS has above normal specific humidity. The NP, GL, and SE all have moisture sources originating from the Gulf of Mexico and converge with anomalous southwesterly flow upstream. The NE region also has some moisture advection contribution from the Gulf of Mexico, the strongest anomalies originate in further east in the Atlantic Ocean and flow to the northwest onshore (Fig. 4.5g). Similar patterns are seen near the

top of the troposphere (250 hPa, Fig. 4.6), although the centers of anomalous specific humidity are slightly further downstream compared to that seen at 850 hPa. The anomalous vectors in all regions suggest an enhanced jet favoring synoptic scale ascent through convergence aloft. Overall, specific humidity composites, especially at 850 hPa, are similar to those developed by Schroers (2020) and although Jennrich et al. (2020) only examined the zonal component of the wind, the results here are consistent with enhanced momentum deposition into each region and helping to provide a lifting mechanism.

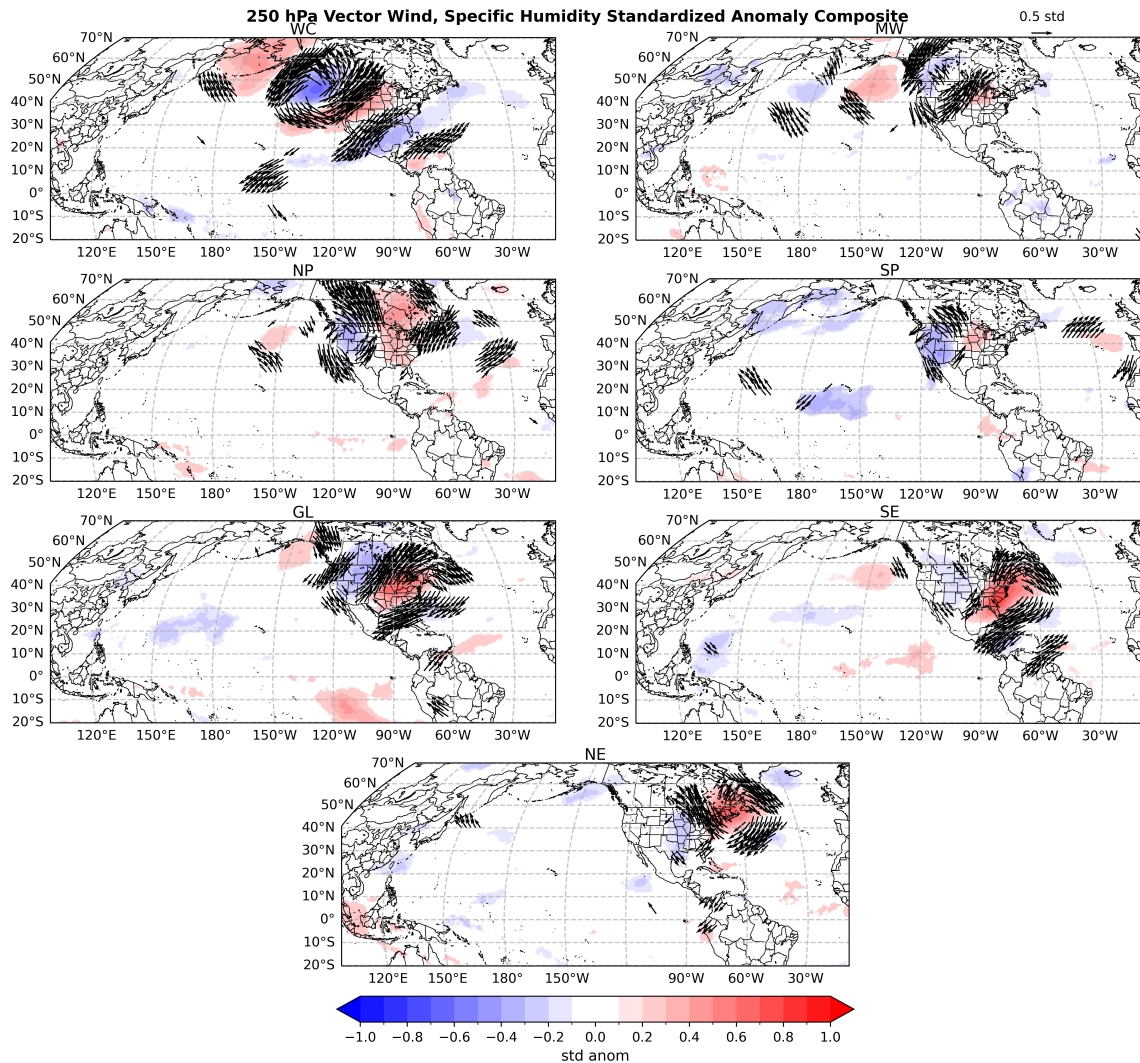


Figure 4.6: As in Fig. 4.5 but for 250 hPa.

4.4 Summary and conclusions

Short duration flooding events have been extensively studied in terms of their impacts, trends, and forcing mechanisms. However, there is a relative lack in literature exploring extended-duration precipitation extremes and how their large-scale patterns compare to single-day events. Composites show that subseasonal extreme precipitation days are driven by similar features to single-day events with anomalous trough-ridge dipoles seen across all regions of the CONUS (Figs. 4.1, 4.2). In observations, there was a clear Rossby wave train signal across the Pacific Ocean across the CONUS in the MW, NP, GL, and SE regions (Figs. 4.1). The SP composite also had a wave train pattern along with anomalous troughing near Alaska. The CNRM also correctly simulated the majority of the mid-latitude patterns across the seven clusters, although trend had a clear and obvious impact on the tropics (Fig. 4.2). The patterns seen also aligned well with some of the only prior work on synoptic features from subseasonal precipitation extremes (Jennrich et al. 2020; Schroers and Martin 2022).

Strong and spatially expansive moisture anomalies, at both high- and low-levels, were also present among regions (Figs. 4.5 and 4.6). Throughout subseasonal extreme precipitation days, a common theme is a strong moisture signal and the dynamical forcing for ascent controls the extent, duration, and magnitude of the extreme (e.g., Barlow et al. 2019). The wind plays a crucial role not only in moisture advection, but also in contributing to ascent through convergence aloft. Accordingly, anomalous ascent is indeed present across all of the regions on extreme period days (Figs. 4.3 and 4.4).

There still remain many avenues of future work. Firstly, there are different temporal evolutions among the extreme periods. Categorizing periods within each cluster into sub-groups may provide insights into the different ways in which the extreme periods unfold and yield more clear large-scale signals. Furthermore, grouping extreme periods by their precipitation time series would make precursor analysis easier since lagged composites, for example, could be constructed relative to the heaviest points within the extreme period.

Some of the climate mode analysis conducted by Jennrich et al. (2020) may accordingly been affected. Finally, ERA5, at the time of writing, was recently extended back to 1940, offering the potential for 84 additional observed extreme periods to be included (at least 1 in every cluster, at least 3 in 80% of clusters). I now look to use the patterns identified here and explore the potential predictability of subseasonal extreme precipitation using two machine learning methods.

Chapter 5

Predicting U.S. Subseasonal Extreme Precipitation Periods

Part I: Random Forests

5.1 Introduction

Accurately predicting extremes and communicating the appropriate risks at any timescale is an immensely difficult problem. Forecasting precipitation extremes are particularly challenging owing to complex interactions between planetary, synoptic, and mesoscale drivers and surface hydrologic responses which determine the spatial location, scale, magnitude, and duration of the event. Indeed, Herman and Schumacher (2018) described extreme precipitation forecasting as “often considered among the most challenging problems in contemporary numerical weather prediction”. Forecasts of extreme precipitation have generally improved over time (Sukovich et al. 2014) through increased understanding of the dynamics of extreme precipitation and improvements in numerical modeling (e.g., Clark et al. 2009; Ikeda et al. 2013; Iyer et al. 2016), among other factors.

Statistical postprocessing of numerical weather prediction forecasts is a popular method to increase the predictive skill of precipitation forecasts. Bias correction techniques, such as quantile mapping, which transforms model output from its climatological cumulative distribution function to the corresponding quantile in observations, have been shown to improve precipitation forecasts at all timescales (e.g., Ines and Hansen 2006; Cannon et al. 2015; Shah et al. 2017; Kolachian and Saghaian 2019). One downside, though, to traditional quantile mapping is its tendency to artificially inflate long-term trends (Cannon

et al. 2015). In a recent study, Hamill et al. (2023) proposed an updated quantile mapping scheme for the Global Ensemble Forecast System version 12 (Guan et al. 2022). The scheme substantially improved forecast (Brier score) reliability across the entire CONUS with improvements in 6-hourly probability of precipitation forecasts out to a lead time of 240-h (Stovern et al. 2023). Unfortunately, however, this skill increase did not translate to the same degree for heavy precipitation, which the authors attribute to biases arising from landfalling TCs, MCSs, and other convectively-driven events, especially in the warm season.

Machine learning (ML) techniques are often applied to postprocess extreme precipitation forecasts because they can effectively exploit nonlinear relationships between complex inputs. ML popularity continues to grow as faster computations are brought about by improvements in hardware and software. These methods can also efficiently process large amounts of data and aid in high-impact decision-making scenarios (McGovern et al. 2017). Random forests are one commonly employed ML algorithm and they have been shown to increase forecast reliability for extreme precipitation (Gagne et al. 2014; Herman and Schumacher 2018; Loken et al. 2019; Hill and Schumacher 2021). Whan and Schmeits (2018) employed quantile regression forests, which model the target variable's (e.g., likelihood of extreme precipitation) conditional distribution as opposed to a deterministic output. Their forest outperformed logistic regression (Wilks 2011) and a zero-adjusted gamma distribution bias correction scheme (Rigby and Stasinopoulous 2005) as evidenced by larger Brier skill scores at nearly all lead-times up to +48-h.

To date, improvements in short-term precipitation forecasts have not translated at the same rate on the S2S timescale. Although skill varies by region and season, raw precipitation forecasts typically have little to no skill relative to climatology in the midlatitudes beyond week 2 (de Andrade et al. 2019; Pan et al. 2019; Becker et al. 2020; Zhang et al. 2021; de Andrade et al. 2021; McAfee et al. 2023). de Andrade et al. (2019) showed significant correlations between observed precipitation and models in the subseasonal-to-seasonal (S2S) database (Vitart et al. 2017) over the CONUS into week 2, although correlations were

insignificant by week 3. Further, they indicated a tendency for models not coupled to the ocean to perform worse than models with ocean-atmosphere coupling. Similarly, Zhang et al. (2021) showed that anomaly correlation coefficients decreased below 0.5 for most of CONUS beginning at week 2 in the North American Multi-Model Ensemble (NMME), signaling that forecasts offer little improvement above climatology. Becker et al. (2020) also found increases in NMME model skill for temperature forecasts since 2011, yet precipitation skill did not experience similar increases. Further, Pan et al. (2019) showed that models in the S2S database offered little to deterministic or probabilistic skill in the wintertime along the U.S. West Coast. Accordingly, the Climate Prediction Center's (CPC) Week 3-4 precipitation outlook remains experimental despite the corresponding temperature forecast being operational since May 2017 (CPC 2023).

Many studies adopted Bayesian approaches to improve forecast reliability of S2S forecasts, which have the added benefit of quantifying uncertainty more than quantile mapping alone. Li et al. (2021) found added skill and more reliable ensemble-mean forecasts in China from the European Centre for Medium-Range Weather Forecasts (ECMWF) subseasonal model after fitting a joint probability model on hindcast data. Huang et al. (2022) also found improved ECMWF precipitation forecasts for catchments in China by calibrating raw forecasts with a combined Bernoulli-Gamma-Gaussian model. Other studies included large-scale teleconnections such as the El Niño Southern Oscillation (ENSO) and Madden-Julian Oscillation (MJO) as these are two major contributors towards subseasonal predictability (Krishnamurphy et al. 2021; Sun et al. 2022). Specq and Batté (2020) developed a scheme that used forecast precipitation, ENSO, and the MJO and obtained significant skill increases in the tropical Pacific up to week 4, most notably when forecasting probabilities for binary events (e.g., weekly precipitation exceeding the upper quintile). Likewise, Li et al. (2022) exhibited skillful and reliable mean pentad precipitation forecasts in China by including outgoing longwave radiation and 850 and 200 hPa zonal wind and geopotential height.

The vast majority of S2S forecast products, however, remain limited to tercile-like approaches (i.e., below normal, normal, above normal), providing decision-makers little information on the potential for high-impact events on such scales. The closest forecast product available for practitioners to make decisions on subseasonal extreme precipitation is the 8-14 hazards outlook published by the CPC. However it, too, remains experimental at the time of writing and many end-users are not aware of this product. Accordingly, current capabilities in predicting extreme precipitation events at the S2S timescale are low, leading to inadequate conveyance of risk to stakeholders. This work presents, to the best of my knowledge, a first step towards improving the predictability of subseasonal extreme precipitation by using random forests fit onto my novel database. Therefore, I pose two central research questions:

1. How well can a random forest classify days as being extreme or not extreme and is there a spatial dependence?
2. Which atmospheric variables are most important in classifying extreme versus non-extreme days, and where is their importance maximized?

Section 5.2 describes random forests and the data used to train and validate the models. The skill of each model is quantified Section 5.3.1 while the importance of each predictor is discussed in Section 5.3.2. A summary and conclusions follow.

5.2 Data and methods

5.2.1 Data

The observationally-based database developed in Chapter 2 was used to fit the random forests. Specifically, a binary time series between 1950 and 2018 was created whereby a given day was denoted with a “1” if it was during an extreme period and a “0” otherwise. Random forests were developed for 3 regions across the CONUS: the Central West Coast

(CWC, cluster 3), the Central Plains (CP, cluster 9), and the Ohio River Valley (ORV, cluster 12). The time series' were split into 1950–2000 to train the random forests whereas 2001–2018 were reserved for testing data.

To predict subseasonal extreme precipitation days, I used 7 variables (see Table 5.3 from the ERA5 reanalysis dataset (Hersbach et al. 2020). The mass-weighted mean was used for geopotential height and wind components to incorporate 3-dimensional information into the model; using multiple isobaric levels as predictors dramatically increased the dimensionality and led to appreciably longer training times. All predictors were regridded from a 0.5° horizontal grid to a 1.5° grid using a neighborhood average of all adjacent grid points centered on grid points that were shared by both grids. This regridding was employed to match the grid spacing of the models in the S2S model database (Vitart et al. 2017). These variables were chosen to give the random forest variables previously identified important variables for subseasonal extreme precipitation (e.g., Chapter 4 here, Jennrich et al. 2020; Schroers and Martin 2022). The domain for all 3 models was 20° - 80° N; 160° - 310° E, allowing the forests to potentially learn local and remote forcing sources.

Lastly, two data scales were tested in each cluster's model, one using standardized anomalies and one using detrended anomalies. Standardized anomalies were made by removing the daily mean and dividing by the daily standard deviation relative to the 1981-2010 climatology; detrended anomalies were calculated by removing the least-squares regression line between 1950 and 2020.

5.2.2 Decision trees and random forests

The fundamental component of a random forest is a decision tree, a flow chart that iteratively poses questions (typically binary, but not always) to make predictions for both classification and regression tasks. There are two types of nodes in a decision tree: decision nodes and leaf nodes. Decision nodes spawn two new nodes, and a numerical threshold from a single input predictor (e.g., temperature $> 20^\circ$ C) determines which child an

incoming example should traverse to. Although all predictors used in this context are numerical, note that decision trees can also utilize categorical data. Leaf nodes represent the end of a branch whereby a deterministic prediction is made. The Python module *scikit-learn* (Pedregosa et al. 2011), which was employed for this work, uses the Classification and Regression Tree algorithm to train decision trees. The algorithm recursively splits the training set into two subsets using a single predictor k and threshold t_k whereby the optimal threshold minimizes the following cost function:

$$J(k, t_k) = \frac{m_{\text{left}}}{m} G_{\text{left}} + \frac{m_{\text{right}}}{m} G_{\text{right}} \quad (5.1)$$

where G_{left} is the Gini impurity of the left subset and m_{left} is the number of samples in the left subset; analogous interpretations hold for “right” subscripts. The Gini impurity is defined as:

$$G_i = 1 - \sum_{k=1}^n p_{i,k}^2 \quad (5.2)$$

where $p_{i,k}$ is the ratio of class k examples among the training instances in the i^{th} node. The Gini impurity measures the frequency in which a randomly chosen sample would be incorrectly labelled if it was assigned a random class label according to the class distribution.

The algorithm continues to make splits until it cannot find a split that will decrease the Gini impurity, although a number of hyperparameters such as the maximum depth may also affect the tree. A leaf node consisting of samples from a single class is known as a “pure” leaf and has a Gini impurity of 0. Predictions from a decision tree are typically deterministic for both classification and regression tasks, although the class proportion within a leaf node can be utilized to make probabilistic classifications.

There are a number of advantages to using decision trees. Firstly, they are intuitive and simple to interpret. Feature selection is also inherently built-in to decision trees since irrelevant features will not produce meaningful splits and thus not get picked. However, decision trees are severely prone to overfitting and are very unstable. Accordingly, random forests are an ensemble of decision trees which retain most of the advantages of decision trees and

typically result in a large decrease in variance (Breiman 2001), provided the individual trees are uncorrelated. Two techniques commonly employed to produce uncorrelated trees are tree bagging and feature bagging. Tree bagging fits each tree in the forest on bootstrapped samples from the full training set. Similarly, feature bagging considers only a random subset of the predictors when looking for optimal split points. Overall, individual trees within the forest tend to have higher bias than a decision tree fit on the original training data but bootstrapping both the sample and predictor space generally results in an ensemble with similar bias and a much smaller variance.

The random forests trained herein consisted of 100 trees and considered the square root of the total number of predictors at each split, which has been empirically shown to be an effective balance between the generalization ability of individual trees and an ensuing increase in intra-forest correlations (Breiman 2001). Furthermore, each model weighted samples inversely proportional to label frequency, forcing the model to have increased emphasis on the extreme days. By definition, the number of “yes” days is much smaller than the number of “no” days and weighting the two classes equally would result in models that are highly biased towards the “no” days. For each cluster, the random forest was fit by optimizing two hyperparameters: (1) the minimum weighted fraction for a node to be a leaf (0.01, 0.02, and 0.05) and (2) the minimum impurity decrease to continue splitting (0, 0.005, and 0.01); the model that had the highest average critical success index (CSI, defined below; Schaefer 1990) score after 10-fold cross validation was chosen as the final model. It should be noted that these hyperparameter choices do not represent the entire space of options that could be tested, and other choices may result in better models. The final parameters for each model are listed in Table 5.1.

Table 5.1: Optimized hyperparameter configurations for each of the fit random forests.

Cluster	Leaf fraction	Minimum impurity decrease
Standardized Anomalies		
CWC	0.01	0
CP	0.02	0.01
ORV	0.01	0.005
Detrended Anomalies		
CWC	0.01	0
CP	0.02	0
ORV	0.01	0

Table 5.2: 2x2 contingency table used to assess skill in binary forecasts.

		Observed	
		Yes	No
Forecast	Yes	<i>a</i>	<i>b</i>
	No	<i>c</i>	<i>d</i>

5.2.3 Model validation

Skill in each of the random forests was quantified through two scores derived from a 2x2 contingency table: the CSI and the symmetric extreme dependency score (SEDS, Hogan et al. 2009), defined as:

$$\text{CSI} = \frac{a}{a + b + c} \quad (5.3)$$

$$\text{SEDS} = \frac{\log[(a + b)(a + c)/n^2]}{\log(a/n)} \quad (5.4)$$

where the ranges of each score are [0,1] and [-1,1] for CSI and SEDS, respectively.

I also calculated “relaxed” CSI and SEDS values by including nearby clusters as potential sources for hits. For instance, nearly the entire state of Oregon is within the polygons for both clusters 1 and 3. Accordingly, for each cluster a set of nearby clusters were used

to search for extreme periods that overlapped with, or were relatively close to, the cluster. Specifically, the nearby clusters were defined as:

- CWC: clusters 1, 2, and 4;
- CP: clusters 7, 8, 10, and 11;
- ORV: clusters 11, 13, 14, and 15.

where the unclustered periods in the database were additionally considered for all clusters. Next, I found the overlap between the original predictions with the labels from nearby clusters and artificially added the days subjectively based on the extreme period polygon's overlap with the respective cluster summary polygon (Fig. 2.9). In other words, I found the intersection between predicted extreme days for a given cluster with the labels from the nearby clusters. See Appendix B for a list of all dates and if they were classified as “hit” days. The adjusted contingency tables were developed as following:

1. Let a_{orig} , b_{orig} , and c_{orig} be defined as the original number of hits, false alarms (FAs), and misses, respectively.
2. Classify overlap days as “hit” days or FA days based on the extreme period polygon's geographic proximity and orientation relative to the main cluster. Let a_{adj} be the total number of new “hit” days.
3. Calculate the number of maximum misses, $14n$, where n is the number of independent extreme periods contributing to the added hits. The total number of new “miss” days is then $c_{\text{adj}} = 14n - a_{\text{adj}}$.
4. The new total number of hits and FAs was $a_{\text{orig}} + a_{\text{adj}}$ and $b_{\text{orig}} - a_{\text{adj}}$, respectively, whereas the new number of misses was $c_{\text{orig}} + c_{\text{adj}}$. The number of null days, element d from the contingency table, is left unchanged.

To illustrate this adjustment, let us walk through the CWC standardized anomaly case.

Table 5.3: Variables used to train the cluster-specific random forests.

Variable Acronym	Variable Description	Isobaric Level(s) (hPa)	Lags
SLP	Sea level pressure	n/a	0
PWAT	Precipitable water	n/a	0
Z	Geopotential height	Avg. in [850, 300]	0, 7, 14
U	Zonal component of wind vector	Avg. in [850, 300]	0
V	Meridional component of wind vector	Avg. in [850, 300]	0

5.3 Results

5.3.1 Model skill

Overall, models for each of the regions have limited skill in differentiating subseasonal extreme days from non-extreme days. In this framework, developing models capable of getting hits comes at a cost of a high FA ratio. In nearly all cases, the number of FAs far outnumber the number of misses (Tables 5.4-5.5). When using standardized anomalies, all 3 models correctly predicted nearly all extreme days within the training set with the CWC forest having the greatest number of misses at 7. The CWC and ORV models have similar performances in both the training and testing sets registering testing CSI scores of about 0.1. The ORV forest has the largest number of hits of the 3 models in the testing set, correctly identifying 38 days, yet also missed 137 days, 2.8 and 5.5 times as many as the CP and CWC forests, respectively. As a result, the CWC has a superior SEDS compared to the ORV forest. The CP model has the lowest CSI and SEDS of the three clusters in both the training and testing datasets as a result of having fewer hits and more FAs. It follows dynamically that the CP region has the lowest skill as the random forests are relying on large-scale synoptic patterns to make forecasts and events in this region likely have a larger mesoscale influence compared to the CWC and ORV regions (e.g., Stevenson

and Schumacher 2013; Bunker 2020). However, the comparatively large number of misses in the ORV suggests that a key driver is missing from the predictor set. Furthermore, all 3 models have a large number of FAs, illustrating the need for additional predictors to differentiate extreme and non-extreme days.

Table 5.4: Performance for cluster-specific random forests within the training datasets. CSI and SEDS have been rounded to 3 significant figures.

Region	Hits	FAs	Misses	CSI	SEDS
Standardized Anomalies					
CWC	356	625	7	0.360	0.739
CP	268	690	3	0.279	0.697
ORV	320	419	3	0.431	0.792
Detrended Anomalies					
CWC	357	735	6	0.325	0.713
CP	268	201	3	0.568	0.865
ORV	320	281	3	0.530	0.843

Trend in the predictors may also play a role in the random forests' capability to make predictions on unseen data. The impacts of detrending are most pronounced in the CP, with an increase in the training CSI score by nearly 0.3 and slightly better generalization in the testing dataset (Tables 5.4 and 5.5). Skill in this region rises due primarily to decreasing the number of FAs from 305 to 67, although this comes at the cost of 2 fewer hits in the testing dataset. Nevertheless, the SEDS in the CP increases by nearly 0.2 in the testing dataset when using detrended anomalies, suggesting that trend plays a key role in the predictability of subseasonal extreme precipitation. The ORV forest has a larger CSI and SEDS in the training set compared to its standardized anomaly counterpart, yet performs worse in the testing set with a decrease of 0.02 seen in both CSI and SEDS. Conversely, the CWC model has 3 more hits when using detrended data coming at a cost of about 60 more FAs. Nevertheless, the CWC continues to be the most skillful cluster with an SEDS of 0.465.

Table 5.5: As in Table 5.4 but over the testing datasets.

Region	Hits	FAs	Misses	CSI	SEDS
Standardized Anomalies					
CWC	31	254	25	0.100	0.475
CP	9	305	49	0.025	0.178
ORV	38	211	137	0.098	0.338
Detrended Anomalies					
CWC	34	311	22	0.093	0.465
CP	7	67	51	0.056	0.346
ORV	21	106	154	0.075	0.317

Employing the relaxed verification scheme saw improvements in CSI of around 0.1 on average whereas the SEDS increased by 0.15-0.35, depending on the data scale and cluster (Table 5.6). The CWC remains the cluster with the fewest misses and thus has SEDS values > 0.6 for both standardized anomalies and detrended anomalies. Yet, the detrended anomaly CWC model produces between 3 and 5 times as many FAs compared to the other two clusters. The CP has the largest increases in SEDS of all the forests going from the strict to the relaxed scheme and, surprisingly, becomes more skillful than the ORV. However, the ORV is more skillful than the CP from a CSI perspective for both input data scales, displaying the sensitivity to chosen skill score. Furthermore, end-users tend to be much more sensitive to misses compared to FAs (see Chapter 7) and may prefer the verification statistics in the CP over the ORV.

5.3.2 Variable importance

In addition to making forecasts for subseasonal extreme precipitation, this work aims to employ explainability methods to understand why the model is making its predictions. A key advantage of random forests is their relative ease in explaining their predictions. Using the Gini impurity, a given predictor is considered to be more important if it can effectively

Table 5.6: As in Table 5.5 but for a relaxed scheme. The change in scores relative to the strict verification scheme are added as the final two columns.

Region	Hits	FAs	Misses	CSI	SEDS	Δ CSI	Δ SEDS
Standardized Anomalies							
CWC	79	206	61	0.228	0.648	0.128	0.173
CP	61	253	123	0.140	0.525	0.115	0.346
ORV	83	166	204	0.183	0.527	0.085	0.188
Detrended Anomalies							
CWC	81	264	59	0.200	0.616	0.108	0.151
CP	21	53	107	0.116	0.567	0.060	0.221
ORV	46	81	171	0.154	0.499	0.080	0.182

split samples higher in each decision tree. Consequently, every grid point in the domain has an impurity importance score for each predictor. Using this metric, lag-0 geopotential height and zonal wind are consistently two of the most important variables in predicting subseasonal extreme precipitation days (Figure 5.1). Although SLP was relatively unimportant for the CP and ORV clusters, it was the next best predictor in the CWC. Similarly, the meridional wind has much more importance in the ORV compared to the other two clusters owing to its dependence on strong southerly flow to supply moisture from the Gulf of Mexico. Accordingly, precipitable water is also important in the ORV with about 10-15% of the total importance. Conversely, precipitable water, along with 2-week lagged geopotential heights, is unimportant in differentiating subseasonal extreme days from other days, suggesting dynamical forcing is more important than the thermodynamics in the CWC. The 1-week lagged geopotential height fields have non-negligible importances in all clusters, implying there are useful precursor signals to subseasonal extreme precipitation events.

Although Figure 5.1 displays domain-summed importances, we can also see where in space the predictors are utilized in the trees. Upstream anomalous troughing is important in driving subseasonal extreme precipitation in the CWC and CP with a 0-day lag (Fig. 5.2). There are also isolated grid points with non-zero importance from the CP forest in the

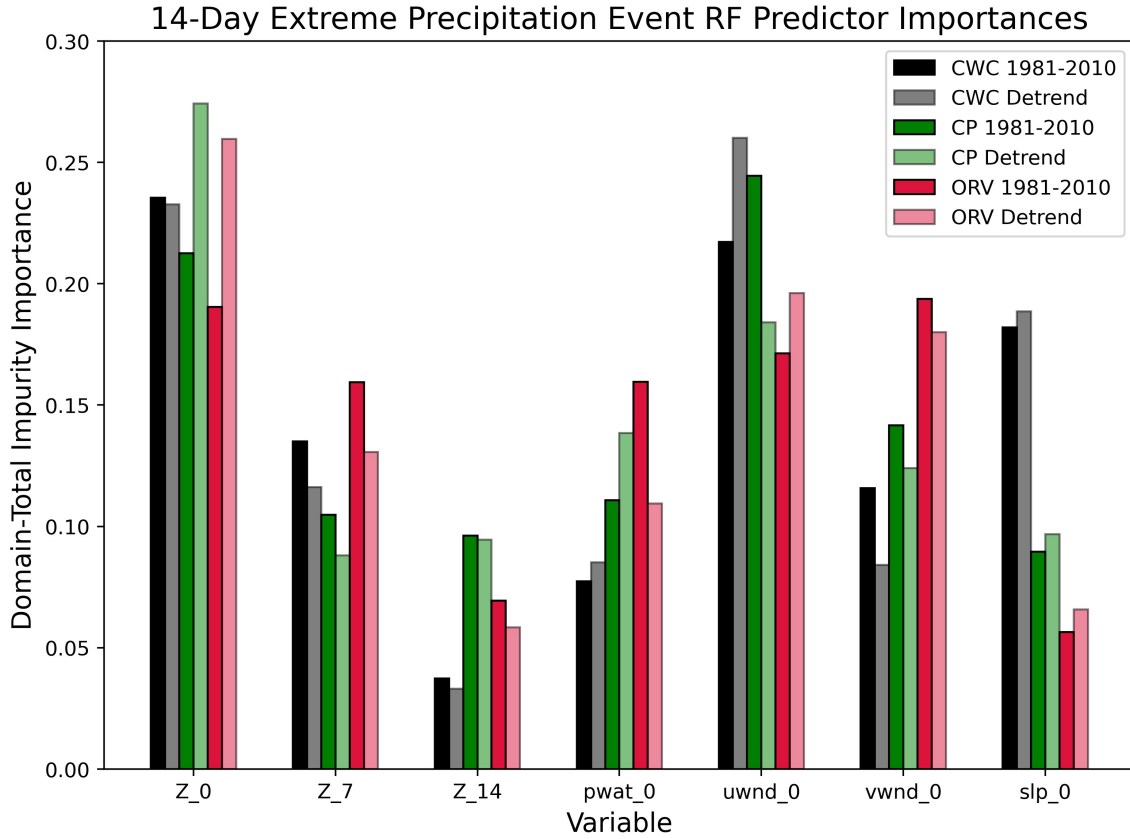


Figure 5.1: Domain-total Gini impurity importances for the 7 predictors for the (black) CWC, (green) CP, and (red) ORV. Darker (lighter) shades are for the standardized (de-trended) models.

subtropical Pacific located under an anomalous ridge. Subseasonal extreme precipitation days in the ORV are more influenced by anomalous ridging downstream in the Gulf and Caribbean than the broad, positively tilted trough located in the Rocky Mountains. Comparing composites over hits and FAs may reveal sources of model error. For example, in CP hits, the region lies in an inflection point co-located with a strong zonal gradient in geopotential heights. Comparatively, this zonal gradient is much weaker in FA days and the upstream trough is fully closed off. Furthermore, FA days in the CP have much weaker anomalous ridging across Canada and the anomalous trough is weaker and shifted farther north compared to hit days. Similarly, the extent of the anomalous trough is farther equatorward in ORV hits compared to that for false alarms. In all three regions, the trough-ridge dipole looks quite similar at lag-7 as in lag-0 (Fig. 5.3) for hit days, suggesting there are

useful precursor signals to at least some of the extreme periods in my database. The signal is most coherent in the CWC, and FA days also feature a relatively deep anomalous trough off the West Coast. There is also a trough setting up in the ORV at lag-7, extending from the high latitudes into the northern CONUS, however importances are greatest in the Bermuda High region and Southwest U.S., again isolating the crucial role of the barotropic high in the Atlantic.

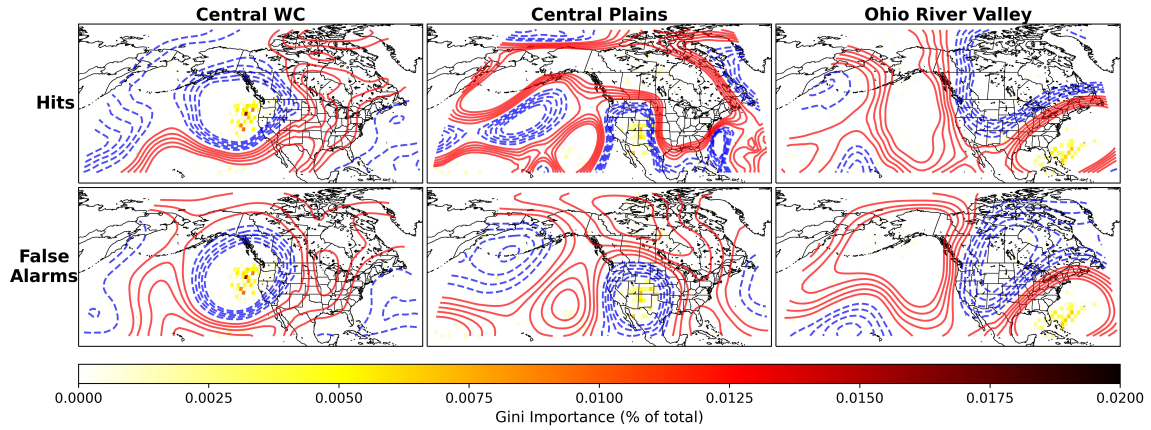


Figure 5.2: (contours) Lag 0 500 hPa geopotential height standardized anomalies with contours beginning at ± 0.1 in increments of 0.1. (shading) Gini importance with lighter shades of red less important and darker shades representing higher importance. The 3 clusters are given in the columns while the top (bottom) row is a composite over model hits (false alarms).

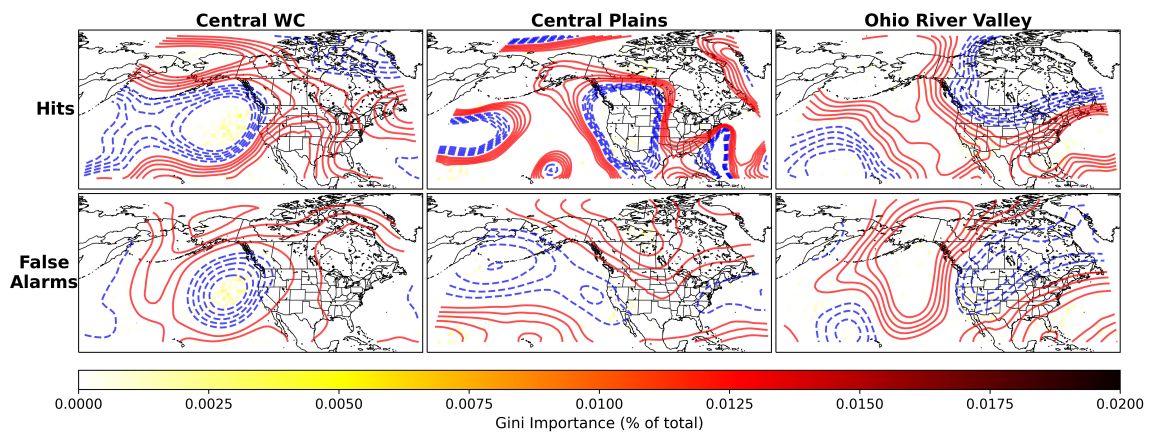


Figure 5.3: As in Figure 5.2, but for a 7-day lag.

The zonal wind component is another variable that is important in all regions (Figure 5.4). In both the CWC and ORV regions, greater importance lies to the south of the region,

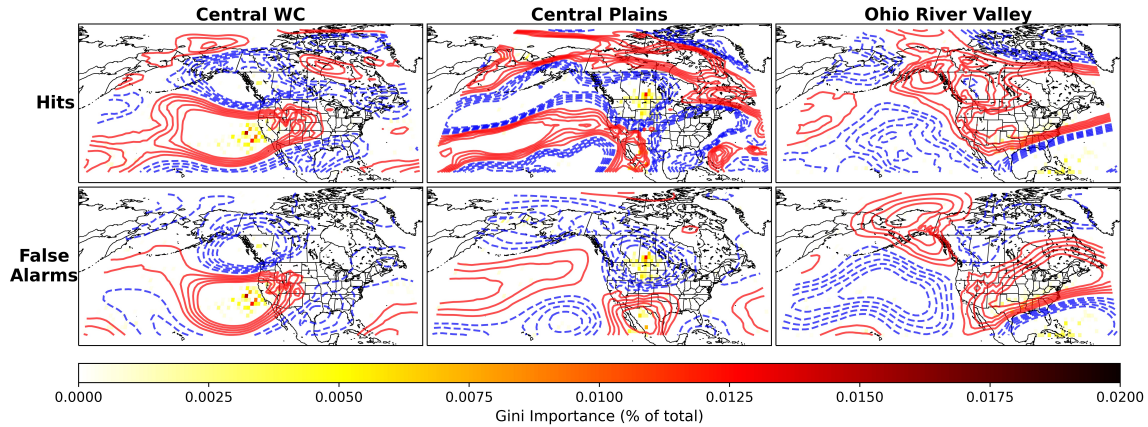


Figure 5.4: As in Figure 5.2, but for the zonal wind component.

although distance from the region is greater in the ORV case. The centroid of the CWC cluster is around the inflection point between anomalous easterlies and westerlies with enhanced westerly flow into California, which is crucial to driving precipitation, likely a signal of landfalling atmospheric rivers. The CP and ORV regions both have higher importance when the zonal wind is weaker. In the CP, precipitation days are enhanced when the jet is stronger in northern Mexico and weaker in the High Plains. However, despite low importances, a key differentiation between hits and FAs in the CP is seen across much of the Appalachian Mountains with hits having anomalous easterlies whereas FAs have near climatological flow. Lastly, the ORV again has higher importance near Cuba as the jet is weakened and perhaps farther poleward. For the meridional component, there is strong southerly flow into all three regions with larger importances in the CP and ORV regions (Fig. 5.5). In all three regions, the positive anomalies in both wind components act to supply copious amounts of moisture, driving broad regions of anomalously high PWATs (Fig. 5.6).

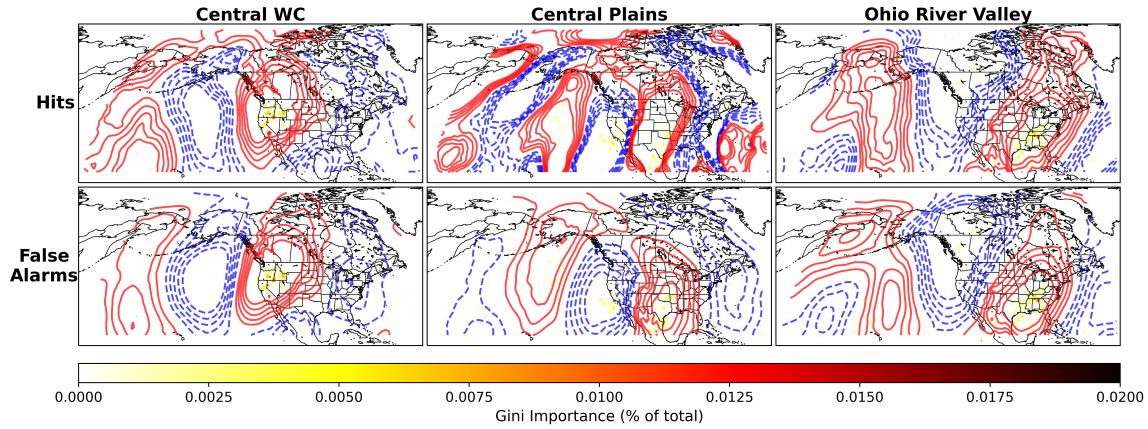


Figure 5.5: As in Figure 5.2, but for the meridional wind component.

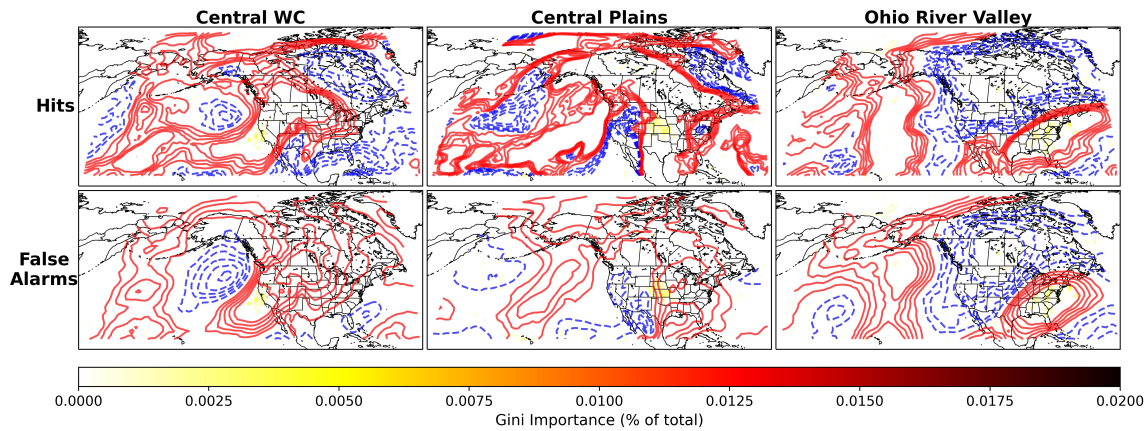


Figure 5.6: As in Figure 5.2, but for precipitable water.

5.4 Summary and conclusions

There currently exist limited forecast products for decision-makers to rely on for extended-duration precipitation extremes. To begin to address this problem, this Chapter examined the utility of random forests to predict subseasonal extreme precipitation days in three regions of the CONUS: (1) the Central West Coast, (2) Central Plains, and (3) the Ohio River Valley. The forests overall showed modest skill in predicting subseasonal extreme precipitation days. The most common error was the preponderance for FAs. The CWC had the best skill of the three regions with a CSI and SEDS of about 0.1 and 0.7, respectively (Table 5.5, across the two input data types). The CWC forest had a probability of detection of 55-60%, however it struggled to differentiate non-subseasonal extreme days, totaling >250

FAs across the testing dataset (2001–2018). The CP also had a relatively large number of FAs using standardized anomalies (305), this was reduced to only 67 after detrending. However, CP model had very few hits regardless of the input. The ORV model stood out from the other two particularly when considering misses, having at least 4 times as many as the CWC and CP models. I hypothesize adding vertical velocity as a predictor would be of particular importance for the ORV model, and the following chapter adds this to the predictor set.

Each of the models in this Chapter was developed on one specific cluster within the Livneh database. Although each cluster is distinct, there is non-negligible overlap among many adjacent clusters. Accordingly, some of the FAs may arise in parts of the country that are in multiple clusters (e.g., Oregon in clusters 1 and 3) and the prediction was, from an atmospheric perspective, correct; indeed, the labels were slightly too strict. Among all FA days from each of the models, I found instances where the given extreme period had overlap with, or was within very close proximity of, the summary polygon of the cluster. These dates are listed in Appendix B. This relaxation greatly improved skill in all regions, with SEDS values > 0.5 in all six forests. However, a caveat to the relaxation scheme is that I only considered predicted extreme days from the given forest when searching for overlap with nearby clusters. There may be extreme periods that would be “close enough” in the nearby clusters, thus impacting the number of misses and null days. A second hypothesis for the number of FAs is due to the setup of the model. The random forests were trained and validated on *daily* data. Because subseasonal extreme periods consist of several embedded daily heavy to extreme precipitation days, the forests likely learned synoptic setups that were favorable for single-day flooding events and accordingly overforecasted extreme days in the testing set. This hypothesis is currently under investigation.

The contingency tables presented above were all constructed using a 50% probability threshold to delineate between extreme and non-extreme days. Although intuitive, such a threshold is not guaranteed to be the optimal decision boundary. In discussions with stakeholders (Chapter 7), minimizing misses is a far bigger priority, and if FA days are

indeed connected to impacts, these forests may provide some utility. Thus, the following chapter incorporates the F -score which can be used to differentially weight FAs and misses.

One advantage of random forests is their ability to be easily interpreted. A metric known as Gini importance, related to the Gini impurity used to fit the forest, gives us clues into the, for lack of better term, importance of each predictor. Geopotential heights and the zonal component of the wind were the most valuable for the forests to learn for all three models. In the CWC and CP, the forests found a nearby trough to be the biggest contributor, co-located with jet streaks and thus favorable for synoptic-scale ascent in a quasi-geostrophic framework. The ORV forest put more weight on the downstream ridge centered in the western Atlantic. Anomalously strong southerly flow throughout the tropospheric-column was modestly important in all three regions and was associated with moisture fluxes into each region. The results shown here are in general agreement with past studies on subseasonal extreme precipitation (e.g., Jennrich et al. 2020; Schroers and Martin 2022). However, the noisiness of Gini importance isolates a weakness of this random forest design: the forests lack capability to employ spatial correlations. Therefore, it is hypothesized that more complex methods that leverage the covariability that is present may lead to stronger, more valuable forecast products. Accordingly, the next chapter explores a more complicated model: a deep neural network.

Chapter 6

Predicting U.S. Subseasonal Extreme Precipitation Periods

Part II: Convolutional Neural Networks

6.1 Introduction

Artificial neural networks (ANNs) are a subset of the broad ML envelope that have burgeoned in recent years, especially in the atmospheric sciences (Chase et al. 2022). The rise in popularity of such techniques led to the creation of a new American Meteorological Society journal titled *Artificial Intelligence for the Earth Systems*, debuting in January 2022. ANNs are inspired by the biological neurons in the human brain and have been used to tackle a wide variety of problems including image classification and speech recognition.

One of the most popular ANNs initially were multi-layer perceptrons (MLP, Fig. 6.1). These architectures share weights across all possible combinations of predictors across two adjacent layers, allowing the networks to learn and model complex functions. These types of layers are also called fully-connected layers and have been used in many applications in the atmospheric sciences. Barnes et al. (2019) developed a simple MLP to predict the year given global surface temperatures and used the model to examine climate variability and change. Satellite products also use similar networks to derive precipitation, such as the Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks suite (Nguyen et al. 2018). On the S2S scale, ANNs have been successfully applied to predict geopotential height anomalies (Mayer and Barnes 2021) and sea-level anomalies (Lee et al. 2021), as well as correct summertime temperature forecasts (Straaten et al. 2023). Fan et al. (2023) also showed significant decreases in precipitation root-mean

square-error at weeks 3 and 4 using fully-connected layers. However, in a fully-connected layer, data are treated as scalars in vector format and, thus, does not optimally exploit the spatial covariability that exists.

Convolutional neural networks (CNNs), on the other hand, are fed images and have accordingly been widely adopted in geosciences. CNNs are particularly popular when correcting precipitation forecasts and have been shown in a variety of studies to reduce mean errors and make more skillful forecasts across a myriad of lead-times (Hu et al. 2021; Li et al. 2022; Sha et al. 2022; Osborne et al. 2022). On the S2S scale, Scheuerer et al. (2020) employed CNNs and obtained improved subseasonal precipitation forecasts over California across weeks 2-4. Other successful applications of CNNs include forecasting the spatial coverage of convection (Lagerquist et al. 2021), severe hail (Gagne et al. 2019), and front identification (Lagerquist et al. 2019). In this chapter, I assess the viability of a CNN to predict subseasonal extreme precipitation in the CONUS using the “true” (i.e., reanalysis) large-scale fields. Furthermore, I do not include precipitation as a predictor so that the CNN learns the relationship between synoptic-scale variables and extreme precipitation as the former have better predictability for longer leads (Li et al. 2017; Cavalcanti et al. 2021).

First, I discuss the various CNN architectures tested in section 6.2. Sections 6.3 and 6.4 present bulk statistics and graphical examples of three extreme precipitation periods from the Livneh database and one real-time scenario. An overall summary and conclusions are given in section 6.5.

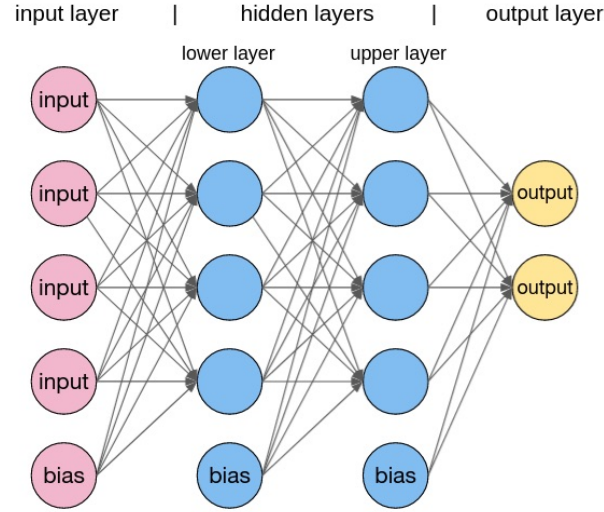


Figure 6.1: Diagram of a multi-layer perceptron. Taken from <https://medium.com/code/introduction-to-how-an-multilayer-perceptron-works-but-without-complicated-math-a423979897ac>.

6.2 Data and Methods

6.2.1 Data

I use the Livneh database developed in Chapter 2 as the truth for the CNN to learn subseasonal extreme precipitation. I specifically use the *preprocessed* database because the CNN will be trained and tested using sliding 14-day windows in the same manner as cataloging the periods in Chapter 2 (see 2.3). The database contains the coordinates to the extreme polygons which were rasterized onto the grid described below.

To predict these subseasonal extreme precipitation periods, I used a combination of variables (see Table 6.1). I used atmospheric fields from ERA5 (Hersbach et al. 2020) on a 1.5° horizontal grid to match the spacing of the models in the S2S model database (Vitart et al. 2017). The CNN was trained on the years of 1950-1997 whereas I reserved blocks of 1998-2008 and 2009-2018 for the validation and testing sets, respectively. The domain was 15° - 61.5° N; 217.5° - 312° E (Fig. 6.2), giving 32 latitude and 64 longitude points. All predictors were standardized by removing the 1950-2018 daily mean and dividing by the standard deviation over the same period. Climatological daily means (and standard

Table 6.1: Variables used to train our U-Nets. n/a indicates a 2-dimensional variable.

Variable acronym	Variable description	Isobaric surface (hPa)
Z	Geopotential height	850, 500, 300
q	Specific humidity	850, 500, 300
T	Air temperature	850, 500, 300
u	Zonal component of wind vector	850, 500, 300
v	Meridional component of wind vector	850, 500, 300
ω	Vertical Velocity	500
SLP	Sea level pressure	n/a
doy	Day of year	n/n
CONUS-mask	Binary mask; 1 for CONUS, 0 otherwise	n/a

deviations) were computed by smoothing the long-term mean on each calendar day with a centered 15-day running mean. The daily standardized anomalies were also averaged across each 2-week period (e.g., 1-14 Jan, 2-15 Jan, etc.). The calendar day of year of the beginning of the window was also used as a predictor. Finally, to emphasize the model's focus on the CONUS, I added a binary mask whereby a grid point's value was 1 if it had an extreme period in the database and 0 otherwise. I hypothesize that 2-week mean fields can be leveraged as predictors for a deep CNN because the synoptic-scale forcing will be captured and a complex network can learn many nonlinear relationships.

I also evaluate our CNN's forecasts when using real-time dynamical subseasonal forecasts. Specifically, I used the National Center for Environmental Prediction (NCEP; Saha et al. 2014) and ECMWF (Vitart et al. 2008) models from the S2S model database. The NCEP model, the Climate Forecast System (CFSv2), is initialized 4 times daily with each initialization having 4 ensemble members, thus real-time runs in the S2S database for the CFSv2 consist of the 4 ensemble members from 0000Z, and each of the 4 members from the prior 3 initializations (i.e., 1800Z, 1200Z, and 0600Z). The CFSv2 hindcast period is

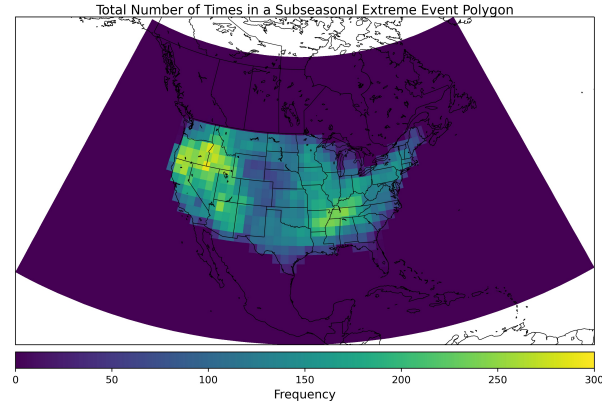


Figure 6.2: Frequency of subseasonal extreme precipitation from 1950-2018.

1999–2010. Using the hindcast climatology, daily long-term means and standard deviations were calculated for each calendar day and lead time and then smoothed with a 30-day triangular kernel as in Pegion et al. (2019). For the ECMWF model, version CY47R3 was the operational version in January 2023. Each time a real-time forecast is made from ECMWF, a hindcast climatology of the most recent 20-years is generated, and I used these hindcasts to standardize the real-time forecast. ECMWF real-time forecasts have 51 ensemble members in total. For the current study, I show forecasts only from the 29 December 2022 initialization; however, similar forecasts are made from earlier initializations and 14-day windows.

Lastly, we use reports from the National Center for Environmental Information Storm Reports Database (<https://www.ncdc.noaa.gov/stormevents/>) to quantify the impacts arising from these extreme precipitation periods. Here, I divide relevant report types into two categories:

1. Liquid reports: flash flood, flood, and heavy rain;
2. Frozen reports: heavy snow, blizzard, winter weather, and winter storm

Frozen reports like sleet and ice storm as the case studies presented herein were primarily snowfall-driven.

6.2.2 Convolutional neural network

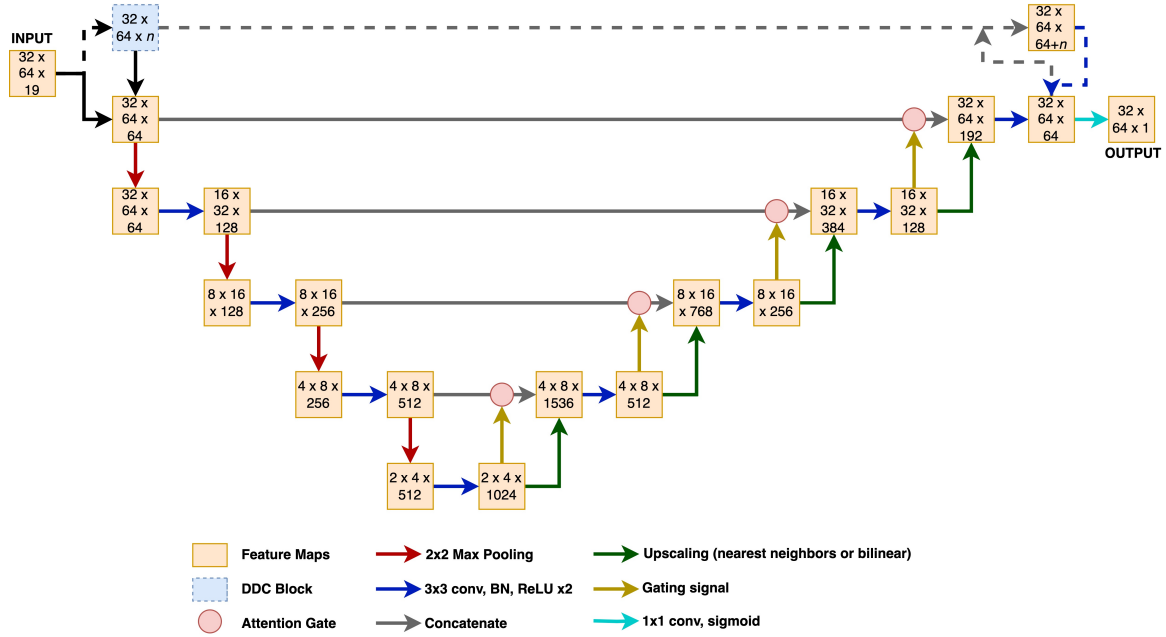


Figure 6.3: U-Net1 architectures used throughout the study. Dashed lines indicate optional steps depending on the experiment. Gating signals are only used for models that use attention gates. Numbers inside each feature map block indicate the number of latitudes, longitudes, and predictors/feature maps; in the decoder, the number of feature maps constitutes a sum of the maps coming from the encoder and the next-lowest decoder level.

I trained a U-Net1 style architecture (Ronneberger et al. 2015) to predict the probability a given 2-wk period experienced a subseasonal extreme precipitation event. The U-Net1 is a symmetric encoder-decoder style architecture (Fig. 6.3) that falls in the class of semantic segmentation, whereby each pixel is assigned its own output class. Hence, the model will receive a series of atmospheric and other fields as predictors and forecast the probability of a 14-day extreme precipitation event in a grid cell. The encoder is the left side of the “U” and, although its output is coarse, it is a numerical representation of the most important combinations of the input. The input fields are passed through a convolutional layer initially and the output, known as feature maps, is subject to a second convolution. In a convolutional layer, weights in a $N \times M$ (here, $N = M = 3$) matrix are slid across each grid point and updated repeatedly in such a way to identify patterns in and amongst all

input fields that are present during extreme precipitation periods. Specifically, convolutional weights are updated during backpropagation, an algorithm that uses the chain rule to propagate the network's prediction errors backwards from the output. Stacking two 3x3 convolution layers allows a given pixel to be affected by a 5x5 region, termed its receptive field. Two 3x3 kernels equates to 18 total trainable parameters whereas a single 5x5 kernel yields 25 parameters, thus making the encoder blocks of 2 3x3 convolutions more efficient despite utilizing the same receptive field. For more details and visualizations of these mechanisms, see Lagerquist et al. (2020).

After each convolution, the feature maps undergo 2 further operations. First, batch normalization (BN; Ioffe and Szegedy 2015) is applied to standardize feature maps across the batch. Standardization ensures feature maps that may have very different variances do not have unequal contributions towards the output, and was also shown to speed up training. Because statistics computed on batches opposed to the entire set may be biased, BN uses an exponential moving average during training to update moving means and standard deviations to be used in the validation and testing sets. Secondly, feature maps are passed through the rectified linear unit (ReLU) activation function, $f(x) = \max(0, x)$, a piecewise linear function which retains all positive activations and setting negative activations to 0, allowing the model to learn nonlinear functions whilst speeding convergence during training. A single block within the encoder consists of 2 iterations of 3x3 convolution, BN, ReLU, and concludes with max pooling. In max pooling, a user-set (here, 2x2) kernel slides across the feature map and only the maximum value within the kernel is kept for the output feature map. The kernel slides across every other grid point (moving 2 both meridionally and zonally), keeping only the maximum activations, and making the grid spacing larger by a factor of 2.

The encoder repeatedly chains these blocks to find increasingly higher-order representations of patterns that, after being trained, are conducive for subseasonal extreme precipitation. As the grid spacing becomes coarser with each max pooling operation, the number of feature maps increases. As done in Ronneberger et al. (2015), I applied 5 convolutional

blocks with the number of feature maps increasing in powers of 2: 64, 128, 256, 512, 1024, although the fifth and final block does not experience max pooling (see Fig. 6.3). Instead, the maps are passed to the decoder, whose task is to “reverse” the high-level representation produced by the encoder and recover the original input resolution. A single block within the decoder first upscales the feature maps by a factor of 2 and then, similar to the encoder, applies 2 iterations of 3x3 convolutions. I ran all experiments with 2 different upscaling methods: (1) nearest neighbors and (2) bilinear interpolation. Transpose convolutions are another popular method; however, predictions using this technique exhibited checkerboard artifacts and thus are less trustworthy in an applied setting (Odena et al. 2016).

The viability of the U-Net style architecture arises from concatenating encoder and decoder feature maps, known as skip connections. Without these skip connections within the model, the process would be analogous to compressing and decompressing a photo: the image would generally remain visible but blurry. Merging maps from both the encoder and decoder minimize information loss and combine the strengths of each segment of the architecture. The final convolutional layer takes a 1x1 convolution across the 64 feature maps from the final decoder block, mixing information from the maps and then applies the sigmoid activation function:

$$S(x) = \frac{e^x}{e^x + 1} \in [0, 1] \quad (6.1)$$

which yielded probabilities on the same grid as input latitude-longitude grid.

In addition to experimenting with 2 different upscaling techniques, I applied 2 existing and 1 new type of blocks to improve the aforementioned base U-Net: residual blocks (He et al. 2016), attention gates (Oktay et al. 2018), and my developed differential depthwise convolution (DDC) module. Residual blocks are similar to the base double convolution block, but it adds the input from the block to the output before continuing in the network. Letting the convolutional block be denoted as F , the output of a normal block is simply $F(x)$. In contrast, the output from a residual block is $F(x) + x$. Residual blocks have

been shown to greatly enhance the ability of deep neural networks to learn, especially early layers, as the residual connection reduces the effects of vanishing gradients during backpropagation (He et al. 2016). I also tried adding attention gates (Oktay et al. 2018) to the decoder. In this context, the attention gate learns coefficients in the interval $[0,1]$ and are used to scale feature maps coming from the encoder, forcing the model to have an increased focus on more relevant parts of the image.

The DDC module was designed to be appended to the beginning of a U-Net. This module leverages depthwise convolutions; in a normal convolution, the same 3×3 kernel is applied to all 19 predictors whereas a depthwise convolution has different kernel weights for each predictor. I used 10 filters for each of the atmospheric predictors whereas each of the other predictors were encoded with a single filter. I added 3 final permutations to the experiment revolving around appending the a DDC module to a base U-Net. The first test outputs a concatenation of the individual predictors' feature maps to the U-Net; the second test adds a normal 3×3 convolution to the concatenated feature maps. The final convolution serves two purposes: (a) it mixes information across the individual feature maps, looking for interactions amongst the predictors, and (b) it reduces the dimensionality from 172 feature maps to 64 to then be passed to a base U-Net. The third and final test using the DDC module was to add another concatenation before the final convolutional layer. By adding another skip connection, I effectively made a U-Net with 6 blocks, rather than 5. I hypothesize that learning patterns in the individual predictors and subsequently including interactions is superior to performing both tasks simultaneously, particularly for geophysical applications.

The U-Nets were trained using Focal loss (Lin et al. 2017). By definition, extreme precipitation is a rare event, thus the total number of instances a given pixel within the CONUS was a 1 compared to a 0 was 1% or less (Fig. 6.2). Similarly, the extreme regions represent relatively small parts of the input image, further accentuating the imbalanced nature of our problem. A major consequence of this substantial class imbalance was that a traditional loss function like binary cross entropy accurately delineating the correct geographical areas

in testing predictions but with very small probabilities. Focal loss enhances binary cross entropy by adding two parameters: α and γ . Focal loss first calculates binary cross entropy and then raises it to the power of γ ($\gamma \geq 0$), which downweighs the loss from confident predictions and places a much higher relative loss amount on high-error predictions (e.g., predicting 0.1 for a label of 1). The loss is also multiplied by α when pixels are in extreme periods and $1 - \alpha$ otherwise, which acts to reduce the influence of the class imbalance. For this study, I use $\alpha = 0.9$ and $\gamma = 2$. The base U-Net had better forecasts on non-extreme periods for larger values of γ but extreme period forecasts worsened overall (not shown), so I chose to use the γ value recommended by Lin et al. (2017). The authors generally recommended smaller values for α . However, I felt larger values were better suited for this highly-imbalanced problem, and the results remain relatively consistent for small changes in α .

The model was trained on an NVIDIA V100 GPU using shuffled batches of 32 and the Adam optimizer (Kingma and Ba 2015) using TensorFlow (Abadi et al. 2015). During training, the area under the precision-recall curve was calculated over the validation set and training stopped when the score did not improve by 0.0001 over 10 epochs. From Table 5.2, precision, also called success ratio, and recall, typically coined probability of detection in meteorology, are defined as:

$$\text{Precision} = \frac{a}{a + b} \quad (6.2)$$

$$\text{Recall} = \frac{a}{a + c} \quad (6.3)$$

Precision and recall balance false alarms and misses, whereas the area under this curve (using increasing thresholds to binarize forecasts) is a better indicator of forecast quality in imbalanced settings compared to receiver operating characteristic curves (Davis and Goadrich 2006).

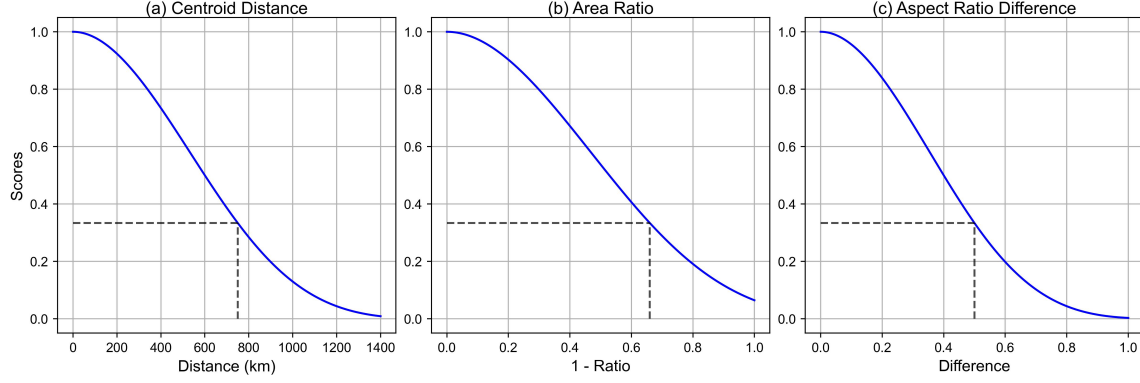


Figure 6.4: Similarity functions for (a) centroid distance, (b) area ratio, and (c) aspect ratio difference between forecast-truth object pairs. Grey dashed lines indicate the e -folding distance.

6.2.3 Object-oriented verification

To assess the quality of the predictions, I used an object-oriented framework (e.g., Davis et al. 2019; Johnson et al. 2020). Object-oriented verification treats the variable of interest as a polygon (e.g., daily precipitation exceeding x mm) and compares properties such as spatial location, size, and orientation of a forecast polygon to the truth. This approach is advantageous for two reasons. Firstly, the database generation process directly provides polygons. Secondly, forecasts valid on the subseasonal timescale should not be expected to have the same spatiotemporal precision as short-term forecasts. Indeed, end-users have built-in expectations of spatial and temporal shifts in long-range forecasts (VanBuskirk et al. 2023), suggesting a forecast which has a phase shift by 100 km, for example, is still a useful forecast and may be considered a “hit” forecast.

I used a modified version of the framework developed by Johnson et al. (2020) to assign truth-forecast object pairs. The algorithm uses 3 metrics: (1) distance between object centroids, (2) ratio of the two polygons’ areas, and (3) the difference in aspect ratios. When comparing any two polygons, individual scores in the range $[0,1]$ are calculated for each metric with larger scores indicating more similar polygons. The scores are calculated using a quasi-Gaussian piecewise function with a user-defined e -folding distance (Gaspari and

Cohn 1999, see Fig. 6.4), allowing for thresholds to be adjusted to the specific application. The final similarity score is calculated as the product of the 3 individual scores:

$$S = f_c * f_a * f_r \quad (6.4)$$

where S is the total similarity, f_c is the similarity score for centroid distances, f_a is the score for the area ratio, and f_r is the aspect ratio difference score. Objects are paired if $S \geq 0.2$. Based on subjective determinations and the values chosen by (Johnson et al. 2020), I set the e -folding lengths to be set at 750 km, 0.66, and 0.50 for f_c , f_a , and f_r , respectively. We hosted a workshop in April 2023 to gather stakeholder feedback and fine-tune these lengths and adjustments to the scheme are currently under investigation. I then constructed 2x2 contingency tables (Wilks 2011) based on forecast-truth pairs. In addition to metrics defined in the previous chapter, I also computed the F-score, defined as:

$$F_\beta = \frac{(1 + \beta^2)a}{(1 + \beta^2)a + \beta^2c + b} \quad (6.5)$$

where β controls the relative weight on precision and recall such that recall is β times more important than precision. Here, we use $\beta = 0.5$ and $\beta = 2$ to assess skill from the perspective of end-users who may not place equal weight on false alarms and misses.

In addition to these skill scores, I also assess how well the probability field align with the rasterized extreme event polygons using the congruence coefficient. The congruence coefficient, g , is defined as:

$$g(x, y) = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} \quad (6.6)$$

where x is a vector of forecast probabilities for a given 2-wk period and y is the corresponding labels vector. Note that $\epsilon = 10^{-6}$ was added to both vectors to avoid division by zero. The congruence coefficient is nearly identical to the correlation coefficient, but calculated with respect to the origin and not with respect to the centroid position (i.e., the

mean). This property makes it a good candidate for identifying the closeness of the match between pattern vectors.

6.3 Results

6.3.1 Summary statistics

I first examine congruence coefficients between the labels and forecasts to get a sense of forecast quality over the CONUS on both extreme (Fig. 6.5; purple) and non-extreme (i.e., “other”, green in Fig.6.5) 2-week periods. Across all combinations, different models appear stronger than others depending on the considered aspect. In the base model, the smoother fields produced by bilinear interpolation likely explain the slight downshift in both extreme and other 2-week periods. Although the lower tails are comparative, the nearest neighbors the right tail forecasts than bilinear interpolation. Using residual convolutions reduces the difference between the two upscaling methods, with the bilinear model now having poor forecasts more often than its nearest neighbors counterpart. The interquartile range (IQR) stays constant or increases on both extreme and non-extreme periods when adding residual connections. Adding attention gates decreases forecast IQRs compared to UNets without these gates and the forecasts during non-extreme windows are considerably better. The labels matrix during these relatively quiet periods is all zeros; thus, the forecasts from the models with attention gates tended to have relatively constant probabilities throughout the CONUS. During extreme periods, the median congruence for forecasts using nearest neighbors was significantly lower than the base-near U-Net. Conversely, the median increase of 0.08 when using bilinear interpolation is significantly greater than the base-bilin U-Net at $p < 0.05$. Using both residual connections and attention gates markedly decreases the spread of congruences using nearest neighbors upscaling, however the extreme event forecast distribution has nearly the lowest median and right tail. Conversely, the same network using bilinear interpolation has a median congruence of roughly 0.58, the highest among

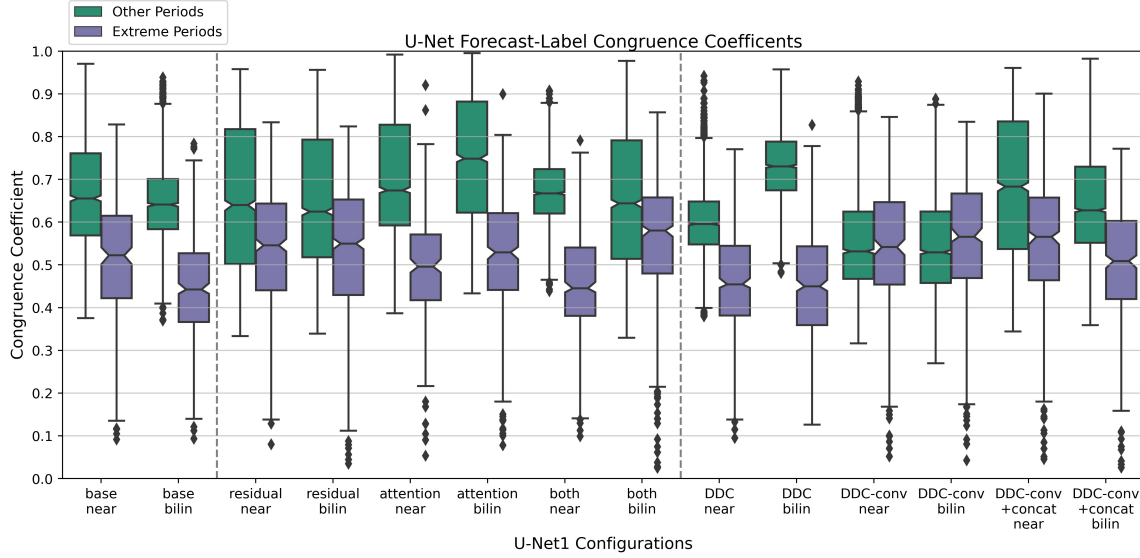


Figure 6.5: Congruence coefficient boxplots for the different U-Net1 configurations for 2-week periods with an extreme somewhere in the CONUS (purple) and non-extreme periods (green). Nearest neighbor interpolation was used for models with “near” whereas models with “bilin” use bilinear interpolation. Notches represent a 95% confidence interval on the median calculated using bootstrapping with 5000 iterations.

models that do not use our DDC module. Overall, attention gates add value to our U-Nets by suppressing activations on non-extreme periods whereas residual connections do not appreciably impact our U-Nets, both in terms of forecast-label similarity and training time (not shown).

The addition of our DDC module onto a base U-Net makes the differences between nearest neighbors and bilinear interpolation become less apparent, and these networks all have among the lowest IQRs in both extreme and non-extreme 2-wk periods. The DDC-bilin U-Net has one of the best forecast distributions during non-extreme periods as it is one of only 2 U-Nets with a median congruence >0.7 . Predictions during extreme periods, though, tend to have lower congruence coefficients than the non-DDC networks. Adding an extra convolution at the end of the DDC module shifted the extreme period distribution towards larger congruence coefficients with both DDC-conv models’ 25th percentile being equal to or greater than the median congruence DDC models. However, the DDC-conv models have smaller congruences on other periods as the distributions both shifted down

and have larger IQRs. Forecasts on these periods improve when including another skip connection for both upscaling types with the median value increasing by roughly 0.1. On extreme periods, the model with nearest neighbors upscaling has an insignificant change in the median compared to the U-Net without the final skip connection at ~ 0.56 whereas its right tail has the largest congruence among all models. Conversely, its bilinear upscaling counterpart does not see any added predictive benefit on extreme periods. Overall, adding our DDC module with a convolution across both the space and predictor dimensions provides some of the best forecasts on periods with subseasonal extreme precipitation whilst concatenating the DDC module's feature maps with the base U-Net output is needed to have consistently larger congruence coefficients on non-extreme 2-week periods. Furthermore, the models with the largest congruences also tend to have some of the smallest congruences, too, suggesting these models may be more precise in their spatial and/or temporal locations. As such, we employ object-oriented verification to shed further light on the performance of these neural networks.

Model objects are developed using the 50% and 60% probability isopleths from a given forecast, to demonstrate how a potential user could select their own risk level and pick the “best” model according to specific needs. Most models overforecast the frequency of subseasonal extreme precipitation periods using a 50% isopleth (Fig. 6.6) at a ratio of between 1.4 and 2.1 to 1. Although most of the models have CSIs between 0.2 and 0.3, the types of errors in the various configurations are nonuniform. On one end, the residual-bilin U-Net's precision is 0.48 (Fig. 6.6, purple plus), significantly larger than any other U-Net, coming at the cost of fewer hits and ultimately the lowest POD. The two F-scores also depict this dichotomy: the residual-bilin has the lowest F_2 score (Fig. 6.7a) and the highest $F_{0.5}$ score (Fig. 6.7b), though both scores are not significantly lower than all other models. Conversely, the base-bilin (Fig. 6.6, purple circle) and attention-near (Fig. 6.6, green “x”) models have PODs > 0.6 as they are more aggressive to label a pattern as conducive for extreme precipitation. These two models forecast extreme periods at nearly 4 times the rate as observations and have about 1.75x more false alarms than the next highest model.

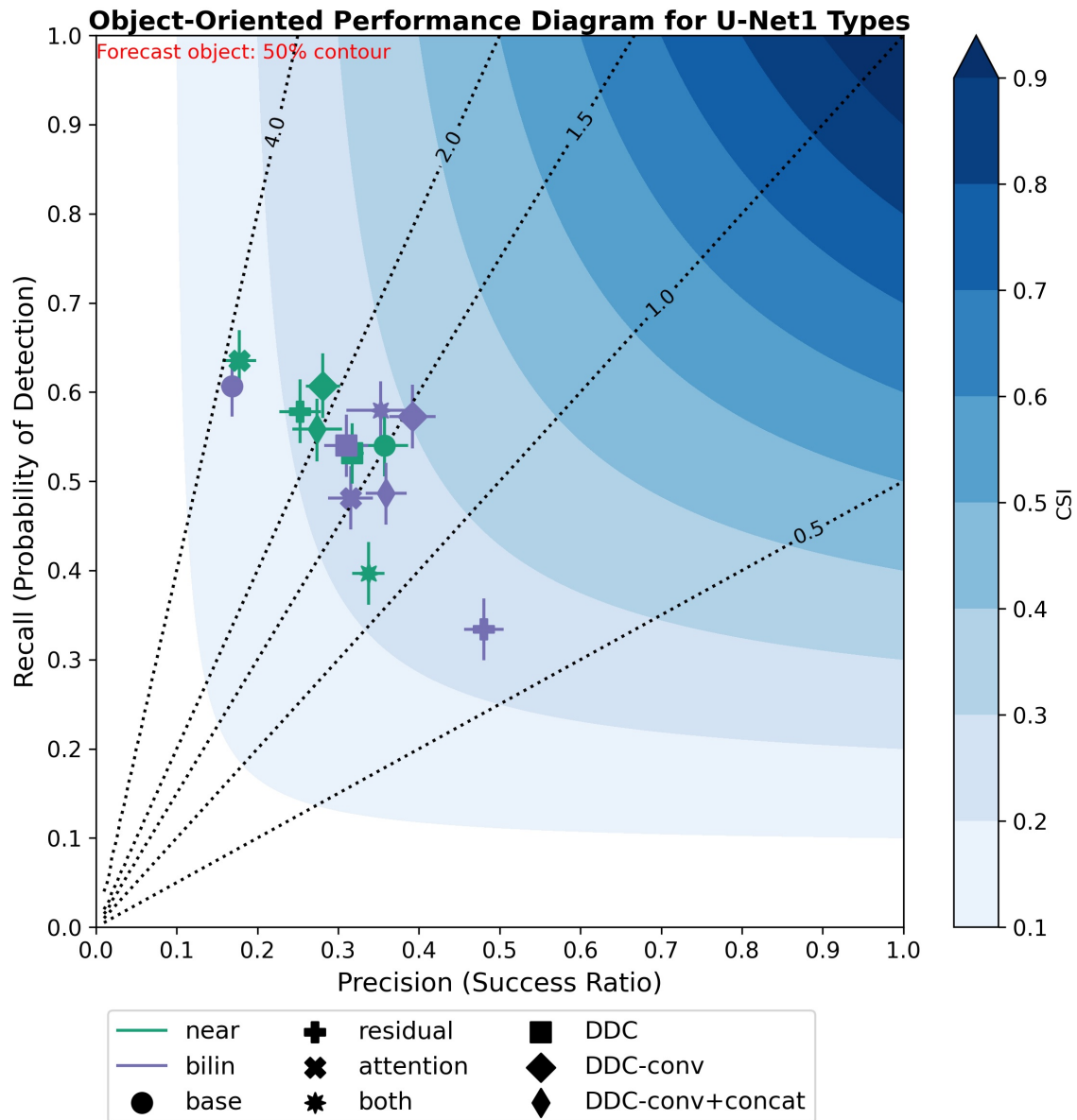


Figure 6.6: Object-oriented performance diagram for various U-Net configurations (see legend) in the testing dataset. Forecast objects were created using the 50% probability contour. Error bars represent a 90% confidence interval and were calculated by bootstrapping with replacement using 5000 iterations.

The high false alarm ratio forces poor $F_{0.5}$ scores, lower than any other U-Net by nearly 0.1 (Fig. 6.7b). Accordingly, forecasts may be identifying periods with precipitation totals in the 90th or 95th percentile (see below). The DDC-conv-bilin model boasts the largest CSI and F_2 scores and second-highest $F_{0.5}$ score (Fig. 6.7b, purple diamond) as it had among the fewest number of misses coupled with the second-highest number of hits. However,

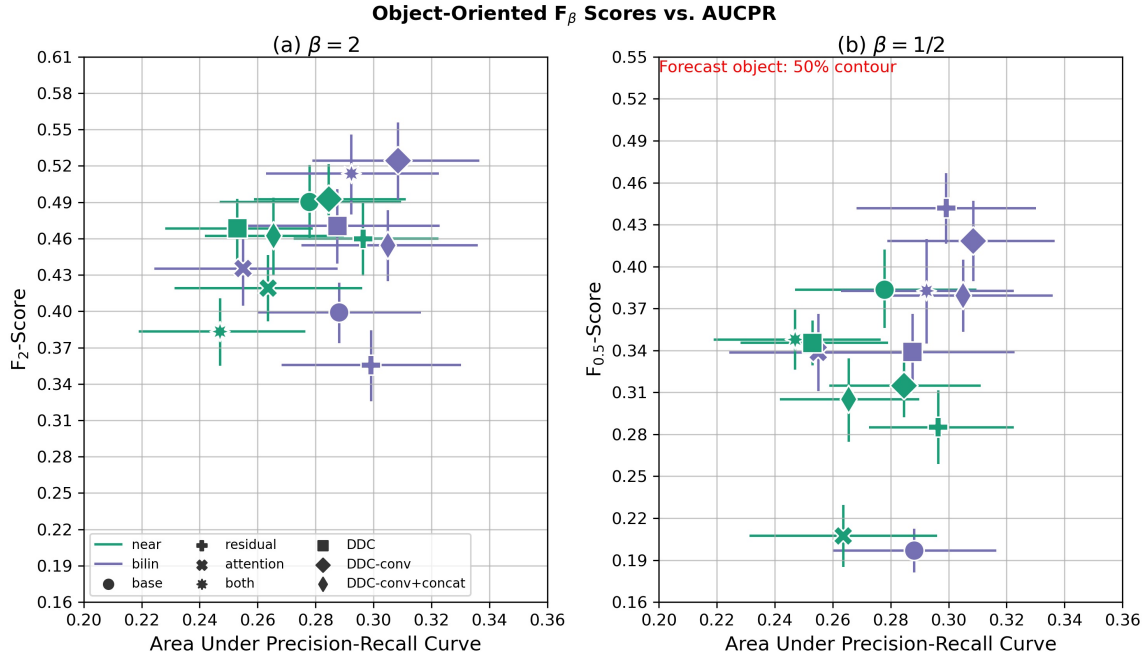


Figure 6.7: (a) F_2 and (b) $F_{0.5}$ scores vs the area under a precision-recall curve for different U-Net configurations. Forecast objects were created using the 50% probability contour. Error bars represent a 90% confidence interval and were calculated by bootstrapping with replacement using 5000 iterations.

these scores are statistically indistinguishable from several other models, suggesting many of these models are making similar forecasts when aggregated across a decade.

Using the 60% probability isopleth to make forecast objects leads to the majority of models having biases roughly between 0.75 and 1.1 (Fig. 6.8). Most of the models continue to yield CSIs in the 0.2-0.3 regime with scores dipping slightly compared to using 50%. The DDC-conv-bilin U-Net still has the largest CSI with a precision of 0.52 and POD of nearly 0.4 (Fig. 6.8, purple diamond). The two aggressive models noted earlier continue to overforecast but yield larger CSIs owing to a nearly constant POD and increased precision. These two models now have the largest F_2 scores (Fig. 6.9a) and no longer have significantly smaller $F_{0.5}$ scores compared to other models (Fig. 6.9b). Contrarily, the conservative residual-bilin model, along with the attention-bilin model, underforecasts extreme precipitation periods by more than half the observed rate (Fig. 6.8, purple plus and cross). The residual-bilin model has the largest precision point estimate of nearly 0.6, however the

DDC-conv-bilin has similar precision with significantly larger POD. As such, the DDC-conv-bilin U-Net has a significantly larger $F_{0.5}$ score than the residual-bilin model as the latter score decreased by $\tilde{0}.02$ compared to an increase of nearly 0.08 (Fig. 6.9b, purple diamond).

In general, U-Nets using bilinear interpolation to upscale yield larger areas under the precision-recall curve in the testing dataset despite bilinear interpolation's tendency to over-smooth compared to nearest neighbors. Yet, the bootstrapped confidence intervals are relatively large, and no model is significantly different than the others. Furthermore, correlations between the areas and CSI and $F_{0.5}$ are 0.4 and 0.55, respectively, suggesting that it may not have been the most appropriate metric to use to stop training. Ideally, object-oriented metrics would be calculated during training whereby the final model was that with largest object-oriented CSI over the validation set. However, at the time of writing, these computations were too expensive to calculate during training and more apparent differences in the architectures may have become apparent if such a scheme was implemented.

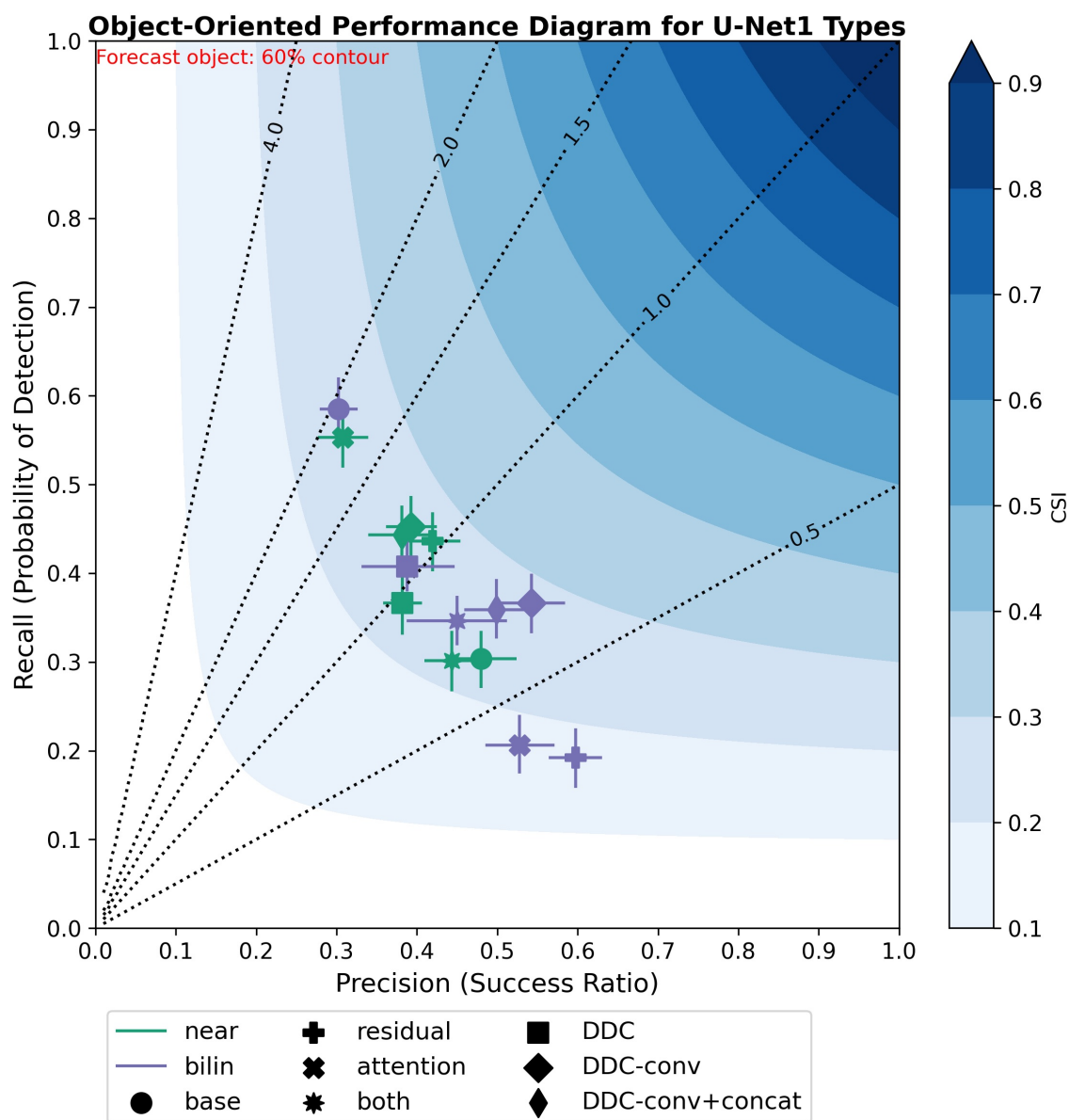


Figure 6.8: As in Fig. 6.6 but using the 60% contour.

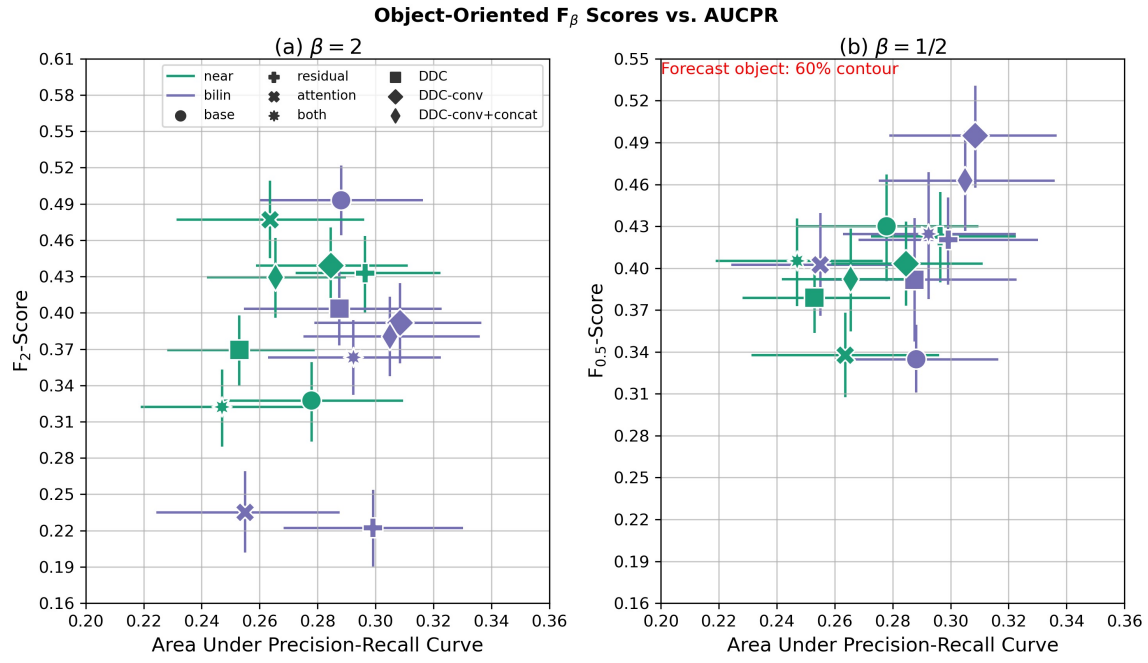


Figure 6.9: As in Fig. 6.7 but using the 60% contour.

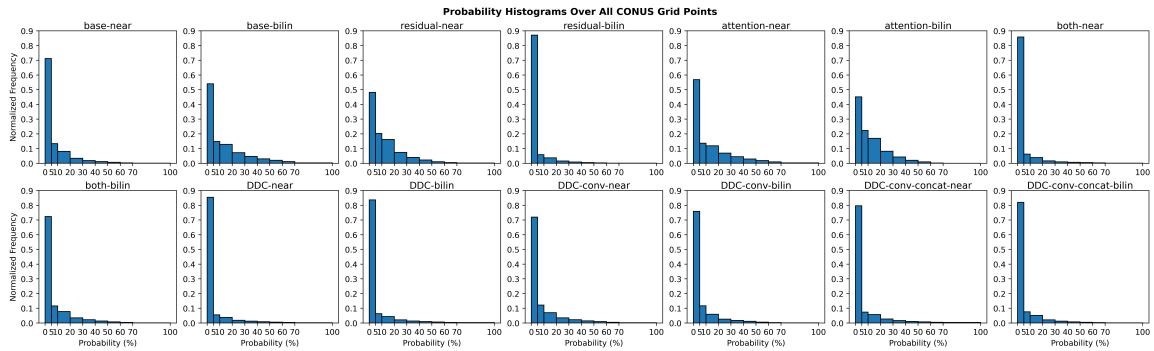


Figure 6.10: Histogram of probabilities across the CONUS in the testing set for each U-Net.

6.4 Case studies

I next present four case studies to illustrate strengths and weaknesses of our trained U-Nets. Specifically, we select 3 models to examine: (1) base-bilin, an aggressive model; (2) DDC-conv-bilin, a balanced model; and (3) residual-bilin, a conservative model. These case studies represent scenarios with both extreme regions and near misses, occurring in different seasons, and differing forcing mechanisms. Because we are assessing the U-Nets' ability using known patterns, all case studies use 2-week composite patterns over the same 2-week period unless otherwise noted.

6.4.1 May 2013 pluvial

The second half of May 2013 featured 2 anomalous upper-level troughs with one centered in the Central Plains around 20 May while a second trough formed in the Pacific Northwest in part owing to a strong Aleutian Low. Between 19 May and 1 June, most of the eastern half of Montana, the western half of North Dakota, and parts of northwestern South Dakota experienced 6-10 inches of rain (Fig. 6.11a), climatologically rare for this time of year as a spatially extensive swath of 2-week precipitation met or exceeded the 95th percentile for this 2-week time span (Fig. 6.11b). All 3 U-Nets predict probabilities exceeding 60% for subseasonal extreme precipitation in eastern Montana and the Dakotas (Fig. 6.11d-f). Highly anomalous precipitation totals also fell across the Midwestern Plains. Throughout the 2-week period, there was consistent strong southerly moisture advection coupled with forcing for ascent from both longwave troughs and embedded shortwave troughs (Fig. 6.12) which coupled to support multiple waves of thunderstorms across the region. Although total precipitation in parts of Iowa, Missouri, and Illinois exceeded their 99th percentiles, not all grid points met the Dickinson et al. (2021) duration criterion, resulting in a polygon that missed the area criterion by only $\sim 23000 \text{ km}^2$. All three U-Nets generalize well to highly anomalous precipitation, both in the High Plains and throughout the Midwest, and the predictions are well associated with the resulting impacts (Fig. 6.11c) despite training

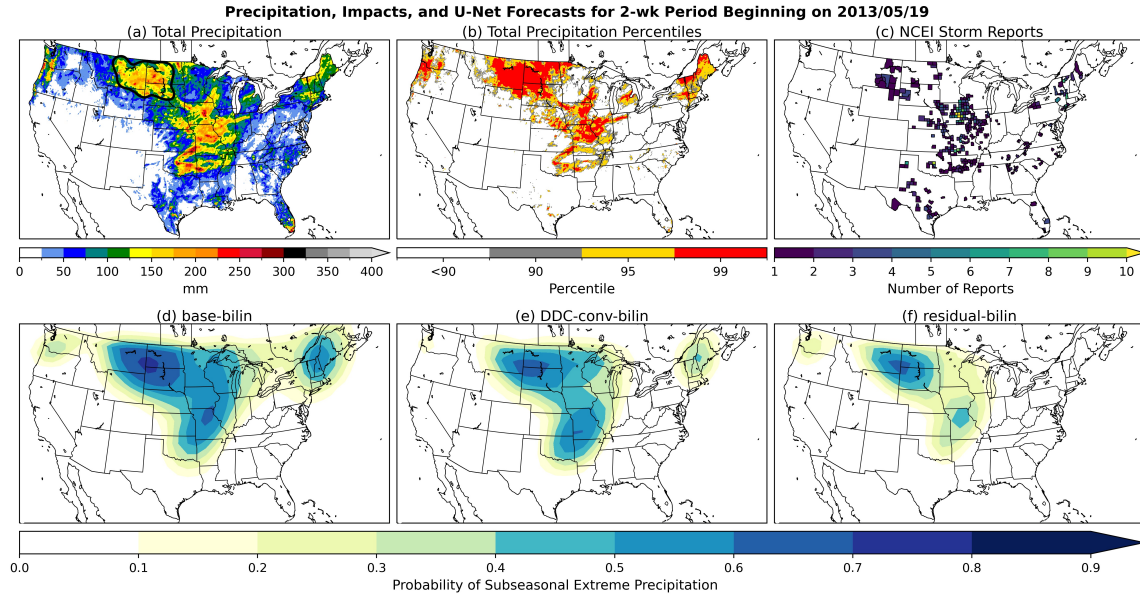


Figure 6.11: (a) Total precipitation (mm; from PRISM) and extreme period polygon (black outline), (b) total precipitation percentile, (c) liquid NCEI storm reports, and subseasonal extreme precipitation predictions from the (d) base-bilin, (e) DDC-conv-bilin, and (f) residual-bilin U-Nets for the 19 May - 1 June 2013 period.

primarily with large-scale atmospheric fields. The base-bilin and ddc-conv-bilin models also predict probabilities up to 30% in New England where heavy rain fell, but only met the 95th-percentile. The more conservative residual-bilin model does not have any appreciable risk in New England and is thus assigned the highest congruence. Despite this, end-users who are less sensitive to false alarms may prefer the other two forecasts with lower congruences, suggesting that association/skill and usability are not interchangeable and increases in one may not necessarily yield increases in the other (VanBuskirk et al. 2023).

6.4.2 February 2018

In February 2018, anomalously high precipitation occurred across much of the CONUS as both heavy rain and snow resulted in near-record totals across parts of Montana and as well as from the Southern Plains into the Great Lakes and mid-Atlantic. In the first half of February, strong and persistent Alaskan ridging was present, fostering northwesterly flow

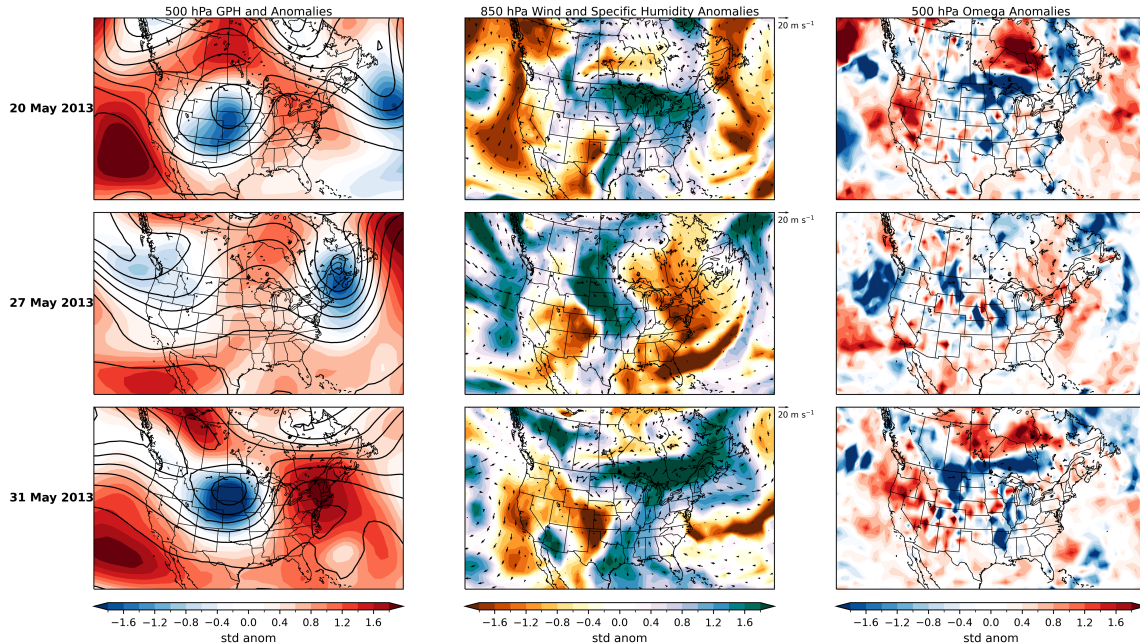


Figure 6.12: (left) 500 hPa Geopotential height (contours) and standardized anomalies (shading), (middle) 850 hPa vector wind and (shading) specific humidity standardized anomalies, and (right) 500 hPa vertical velocity standardized anomalies on (top row) 20 May, (middle) 27 May, and (bottom) 31 May 2013.

into the northern Rockies and, by mid February, helping to force a strong positively tilted upper-level trough from central California north into Montana (Fig. 6.14). The Alaskan High also transported anomalously high amounts of moisture from the central Pacific into the Pacific Northwest region, especially early in the period.

Multiple waves of heavy snow and blizzard conditions affected Montana between 3-16 February and snowfall totals exceeded 3 feet across much of the state; many of the counties in Montana had frozen storm reports that often had time lengths of more than a day (Fig. 6.13c). Although liquid-equivalent precipitation totals were modest (Fig. 6.13a), these totals exceeded the long-term 99th percentile across a large portion of Montana (Fig. 6.13b). However, even the most aggressive model forecasts a peak probability magnitude of around 25% (Fig. 6.13d).

On the other side of North America, anomalously high pressure off the East Coast favored enhanced moisture advection from the Gulf of Mexico into the southeastern CONUS. Longwave troughing was present throughout the eastern two-thirds of the CONUS with

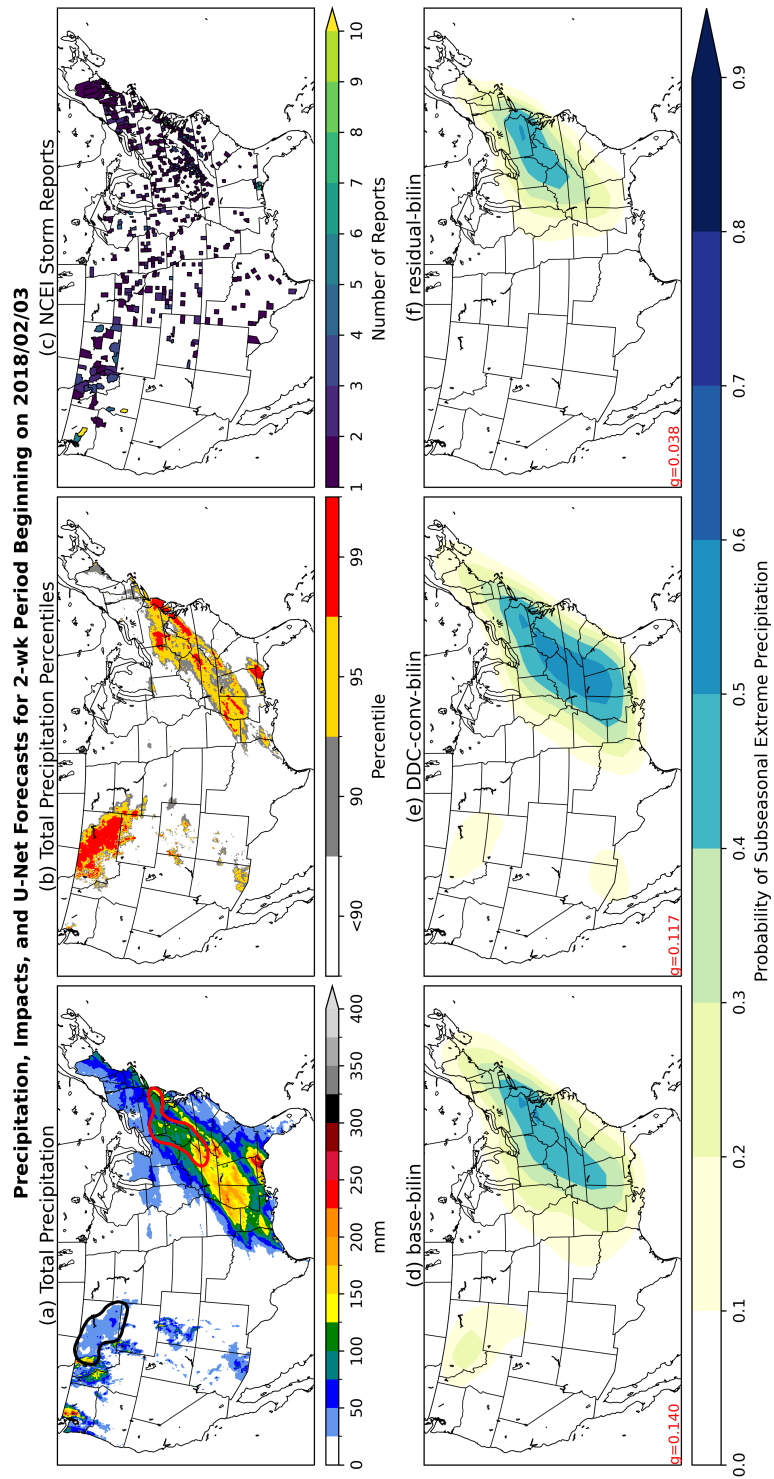


Figure 6.13: As in Fig. 6.11 but for the 3-16 February 2018 period. The red polygon in (a) is an extreme period which starts 2 days later (see text). Both liquid and frozen storm reports are included in (c).

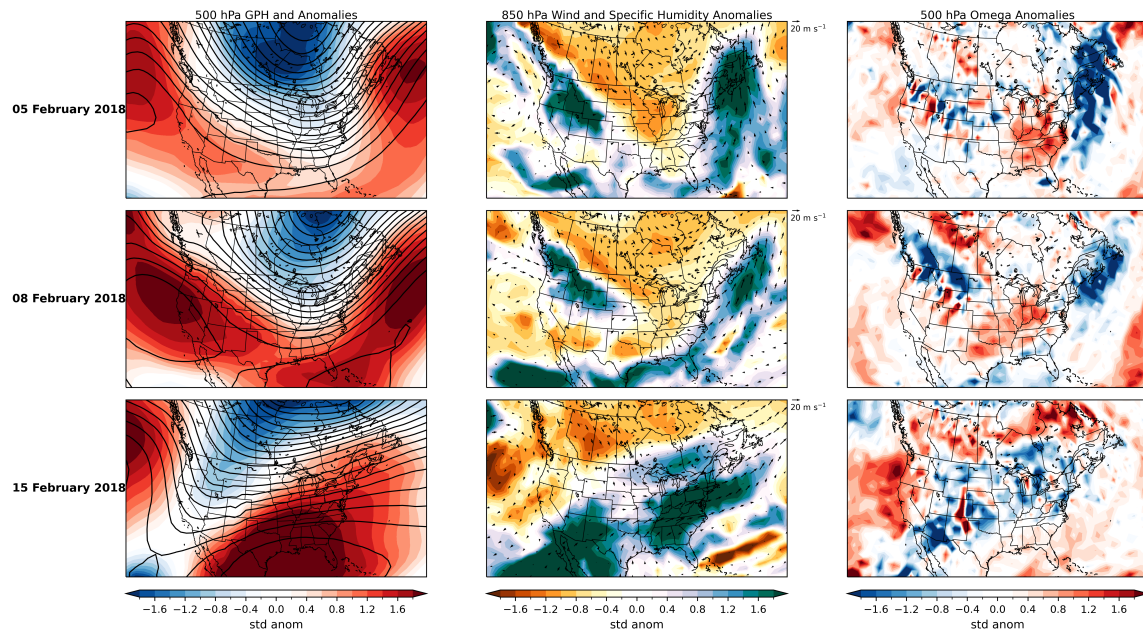


Figure 6.14: As in Fig. 6.12 but for (top) 5 February, (middle) 8 February, and (bottom) 15 February 2018.

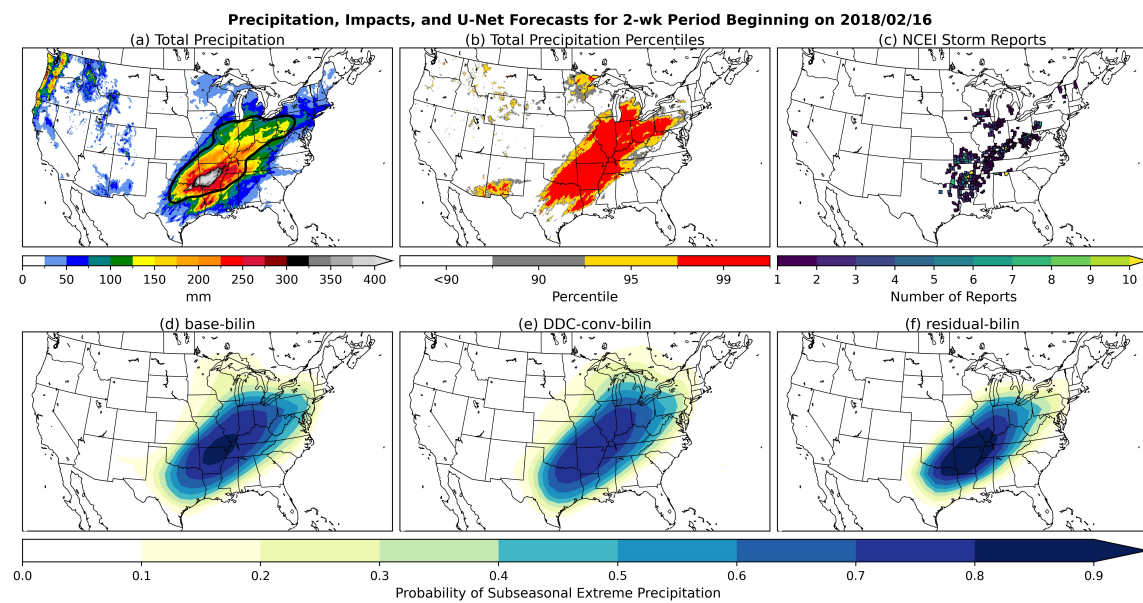


Figure 6.15: As in Fig. 6.11 but for the 16 February - 1 March 2018 period.

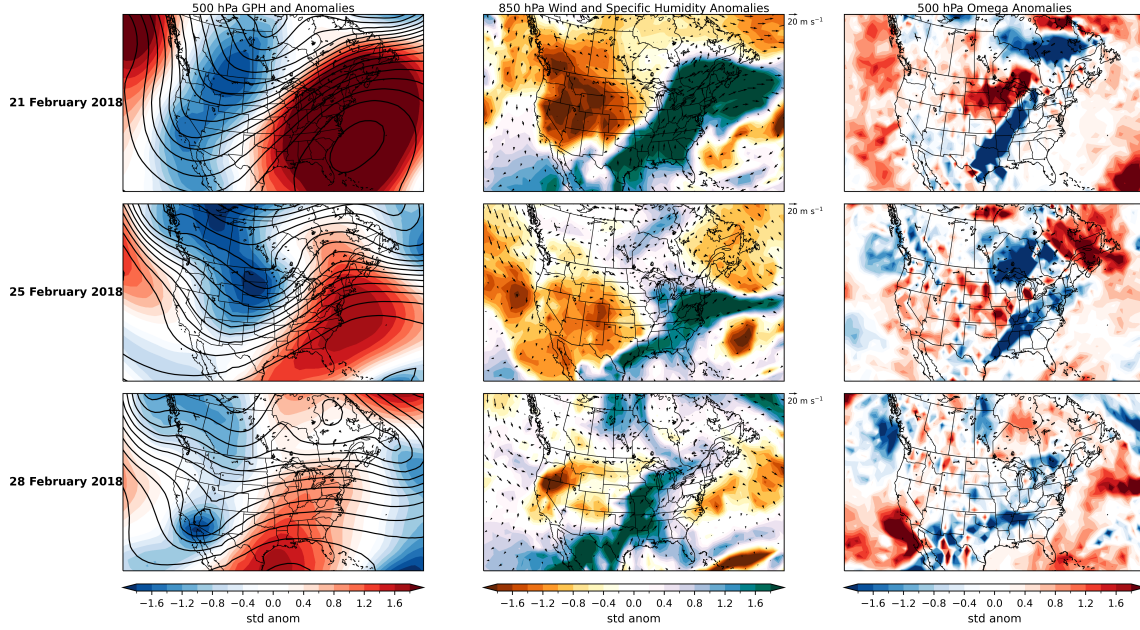


Figure 6.16: As in Fig. 6.12 but for (top) 21 February, (middle) 25 February, and (bottom) 28 February 2018.

multiple jet streaks positioned from Kentucky into New England supported synoptic forcing for ascent (Fig. 6.16). Visual comparison of all 3 U-Net predictions indicate they are well associated with anomalously heavy precipitation bringing impacts as probabilities between 30-50% span from northern Mississippi and Alabama into West Virginia and Pennsylvania (Fig. 6.13d-f). Despite these forecasts being categorized as false alarms, they would likely bring value to many users, particularly when combined with other forecast products such as antecedent conditions. Shifting the start date by only 2 days yields an extreme event throughout the mid-to-south Appalachian Mountains (red polygon in Fig. 6.13a). The 3-16 February forecast object developed using the 50% probability isopleth from the DDC-conv-bilin U-Net (Figs. 6.13e) would pair (i.e., $S > 0.2$) with the 5-18 February extreme event. The predictors for each 2-week window share 12 days and thus will be highly correlated. Therefore, our U-Nets generalize well to the large-scale drivers of subseasonal extreme precipitation and, like the May 2013 example, can bring value despite small spatiotemporal errors.

The second half of February 2018 saw no relief from extreme precipitation across much of the eastern two-third of the CONUS. This subseasonal period was characterized by highly amplified flow as ridging strengthened near Alaska and off the US East Coast. One longwave trough dug south into California on 20 February before stalling for several days (Fig. 6.16). A second trough deepened along the West Coast, forming a closed low in southern California into Arizona by 28 February and, along with the prior quasi-stationary wave, forcing strong synoptic forcing ascent. The low-level jet, with meridional wind standardized anomalies consistently above 1, was also anomalously strong, supplying ample moisture from the Gulf of Mexico. The combination of lift and moisture resulted in precipitation totals exceeding 6 inches from northeastern Texas through the Ozarks and into the Midwest and Great Lakes, exceptionally anomalous for late February (Figs. 6.15a-b). Predictions from all 3 U-Nets have strong associations with the labels (i.e., extreme period polygon) with probabilities greater than 70% across the region and clearly are well-associated with heavy rain and flooding reports (Figs. 6.15c-f).

6.4.3 September 2008: Hurricanes Gustav and Ike

Tropical cyclones (TCs) can bring prolific rainfall totals both during landfall and throughout extratropical transition. In September 2008, Hurricanes Gustav and Ike made landfall in Louisiana and Texas, respectively, bringing heavy rain among many other hazards. The extreme event polygon (Fig. 6.17a) depicts the path Gustav followed after landfall and during extratropical transition as it interacted with a surface baroclinic zone draped from the central Plains into the Great Lakes (Fig. 6.18). Several inches of rain fell from Louisiana and as far north as southern Michigan (Fig. 6.17b). By 12-14 September, Hurricane Ike was moving towards the Gulf states, driving daily rainfall totals exceeding 1.5 inches as far north as southern Michigan. Strong surface high pressure systems were in place in the High Plains as both TCs impacted the CONUS. The combination of vigorous moisture advection from the TCs with the stationary baroclinic zone provided the ingredients for repeated

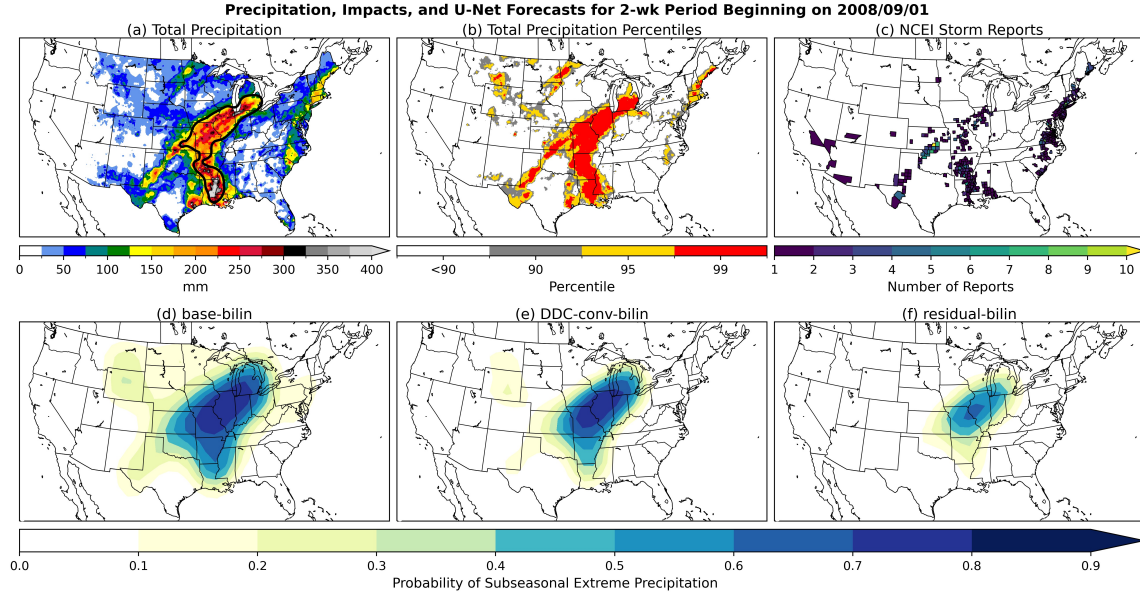


Figure 6.17: As in Fig. 6.11 but for the 1-14 September 2008 period. Note that total precipitation in (a) is from the Livneh dataset.

waves of convection. The CNNs depict increased risk of subseasonal extreme precipitation centered over the mid-Missouri River Valley, spatially consistent with the region most impacted by extratropical forcings (Figs. 6.17d-f).

The vast majority of cases in the Dickinson et al. (2021) database are larger-scale extratropical cyclones and, accordingly, the networks all favor regions further north over Louisiana and Arkansas. Nevertheless, the base-bilin and DDC-conv-bilin models both hint at extreme precipitation as far south as the Gulf of Mexico (Figs. 6.17e,f). Although TCs can bring exceptional rainfall totals, they are relatively short-lived after landfall and thus are not traditionally considered subseasonal in nature. As such, the present iteration of the CNNs may be biased low when predicting extreme precipitation when TCs are making landfall and separate, TC-specific models should be accordingly developed.

6.4.4 Realtime forecast: January 2023

Lastly, we examine a scenario using forecasts from two subseasonal models. One potential advantage of our approach in using standardized anomalies allows for a model to be

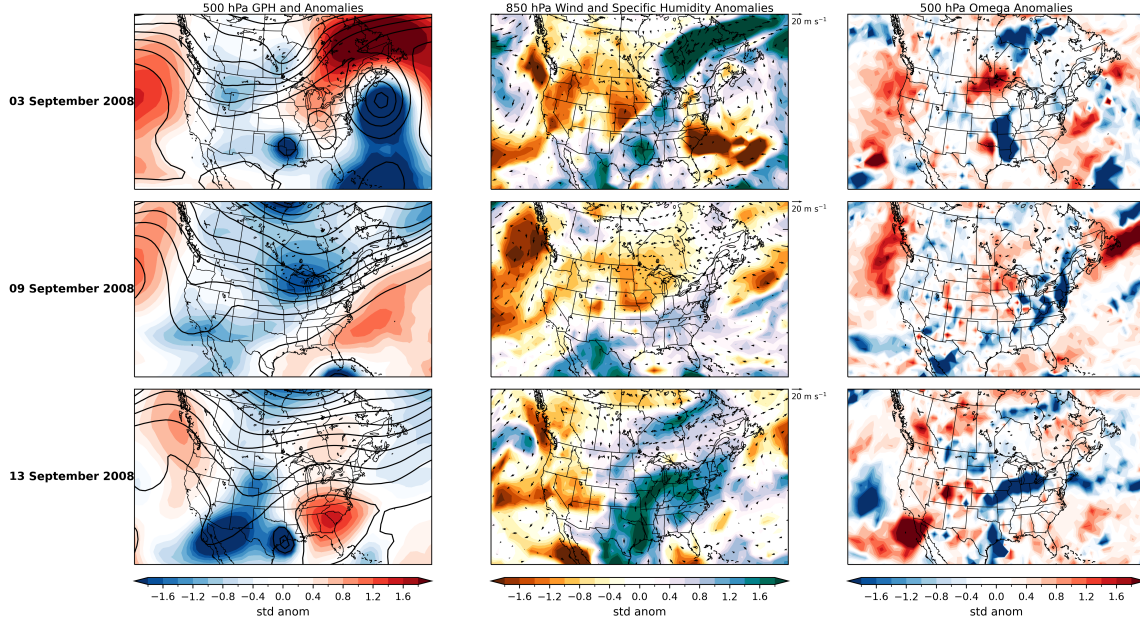


Figure 6.18: As in Fig. 6.12 but for (top) 3 September, (middle) 9 September, and (bottom) 13 September.

standardized with its own climatology and used as input to my U-Nets. In January 2023, a series of landfalling atmospheric rivers affected California and other parts of the Interior West bringing massive rain and snowfall totals to the region (Fig. 6.19a,b). Our U-Nets forecast 30-50% probabilities for subseasonal extreme precipitation across interior California, Nevada, and western Utah (Fig. 6.20). Although the majority of media attention was on California, the powerful ARs contributed to highly anomalous snowfall totals from Nevada to southern Minnesota (Figs. 6.19a,b). The base-bilin CNN and, to some extent the DDC-conv-bilin CNN, suggest this inland push of moisture and forecast a broad 10% probability region in the North-Central Plains for both the ECMWF and NCEP dynamical models. Additionally, the ECMWF dynamical model also suggests slightly increased odds for extreme precipitation in the Northeast US (Figs. 6.20a,b). The NCEP model also has around 10% probabilities in southern New England, however it places more risk in the Southeast (Figs. 6.20d-f). Swaths of enhanced precipitation fell across parts of Georgia and along a SW-NE oriented axis from northwestern Louisiana and central Arkansas into southwestern Pennsylvania, associated with flooding reports early in the window (Fig. 6.19c) and convective

Precipitation and Impacts for 2-wk Period Beginning on 2023/01/01

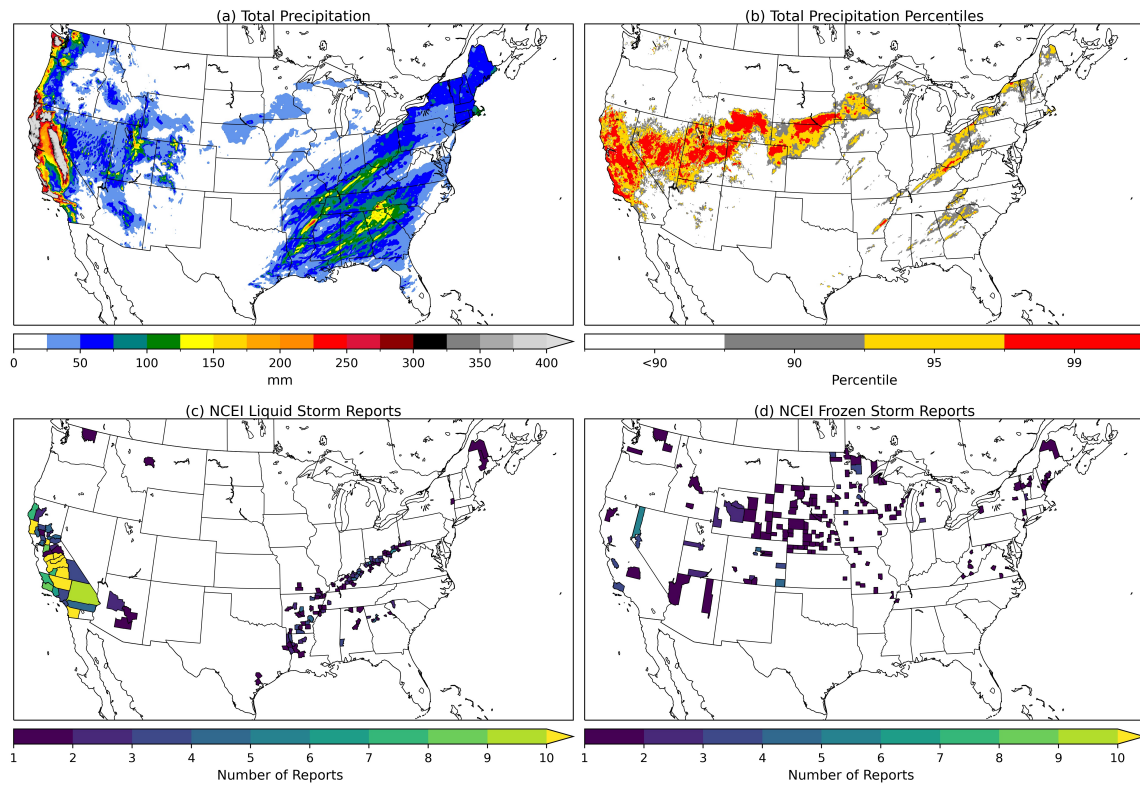


Figure 6.19: (a) Provisional PRISM total precipitation, (b) total precipitation percentile, and NCEI storm reports for (c) liquid and (d) frozen report types for the 1-14 January 2023 period.

impacts (e.g., thunderstorm wind) on 12 January (not shown). Overall, these results are a promising start towards building an operational subseasonal extreme precipitation forecast system.

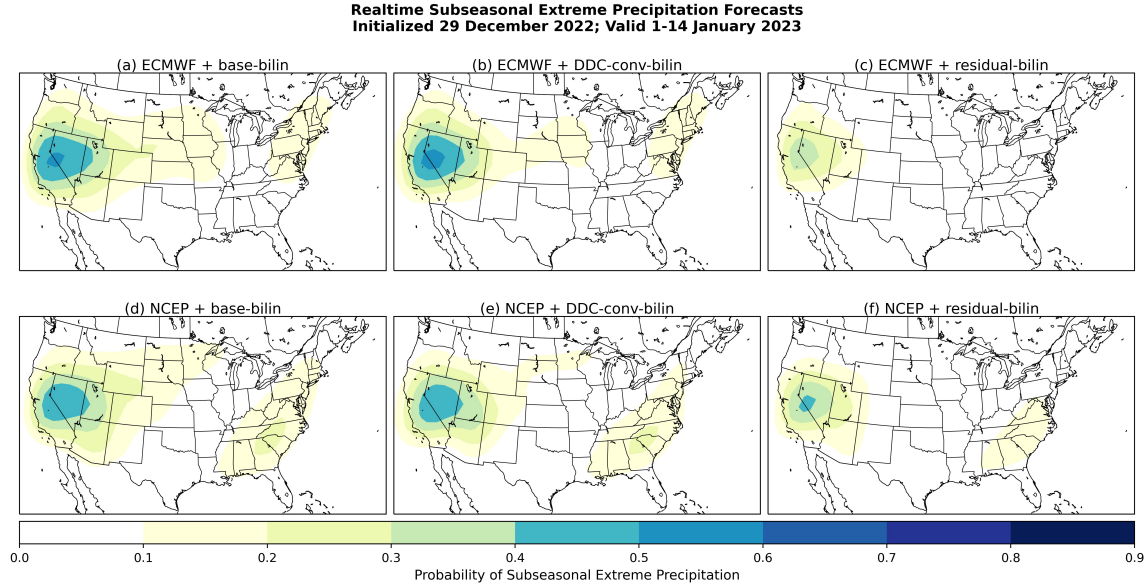


Figure 6.20: Realtime forecasts initialized on 29 December 2022 from (a-c) ECMWF and (d-f) NCEP subseasonal models valid for 1-14 January 2023. The dynamical forecasts were input into the (a,c) base-bilin, (b,d) DDC-conv-bilin, and (c,e) residual-bilin U-Nets.

6.5 Summary and conclusions

In this Chapter, I examined the utility of a CNN architecture coined U-Nets, developed to take in a series maps as input and output predictions with the same input resolution, in predicting locations of subseasonal extreme precipitation. I started with the base U-Net proposed by Ronneberger et al. (2015) and applied 3 types of modifications to test a variety of architectures: residual connections (He et al. 2016), attention gates (Oktay et al. 2018), and a block I developed utilizing depthwise convolutions. I showed that the DDC block with an extra convolution yielded some of the best forecast associations with the extreme period polygons in my database whereas attention gates were particularly beneficial in suppressing activations on non-extreme days. However, areas under the precision-recall curve were not well correlated with any of our contingency table scores and using a different metric to determine when training is terminated may lead to more apparent differences between model designs.

I also employed an object-oriented scheme to quantify the skill of the trained U-Nets and found generally consistent forecast quality across all model architectures (Figs. 6.6-6.9). However, the types of errors were not systematic; indeed, models like base-bilin attention near were more aggressive, predicting larger and more spatially widespread probabilities, whereas the residual-bilin model had a lower POD but significantly larger SR (Fig. 6.7). These biases were reflected spatially in all 4 case studies we presented despite different forcing mechanisms across the cases (Section 4). In each example, the predictions from the U-Nets were well associated with both the long-term precipitation percentile ranking and NCEI storm reports (Figs. 6.11-6.19), illustrating the models ability to generalize to subseasonal extreme precipitation. Furthermore, many of the false alarms produced by the U-Nets were in regions that narrowly missed my extreme period definition, illustrating that metrics arising from a 2x2 contingency table may not map directly to stakeholder usability, especially at extended lead-times.

Our trained U-Nets also showed some capability to make accurate predictions using forecasts from subseasonal models (Fig. 6.20) despite being trained on a different dataset. This is one key advantage to our approach of training on standardized anomalies, since any input data can be standardized with respect to its own climatology. Although there may be some forecast improvement arising purely from the U-Nets, bias correction is still needed to remedy forecast drift and other systematic errors. Li et al. (2017) found that the CFSv2 simulated the circulation patterns during East Asian cold snaps and showed that forecasts for variables such as geopotential height and sea-level pressure were skillful at much longer leads than precipitation. Accordingly, bias correction on large-scale fields like those used herein will likely be an easier task than correcting the precipitation forecast directly. ML bias correction architectures have been shown to change the sign of model errors, and thus a combined bias correction-forecast architecture is the source of future work. In this way, the forecast component can be trained on observational and/or reanalysis datasets, giving a much longer period of record and more positive examples for the network to learn from. A second advantage is speed: predictions from a trained U-Net are made quickly enough

for operation adoption, on the order of seconds for a given forecast window and ensemble member and the order of minutes or less across an entire ensemble, depending on the number of members. Moving forward, future studies should include explainability techniques to understand why the U-Nets are making predictions. These techniques will build trust in the model, making it easier to identify sources of model error and potentially offer sophisticated end-users additional information to use in their decision-making processes.

Chapter 7

Broader Impacts: Incorporating Stakeholder Feedback

7.1 Introduction

S2S extreme precipitation, as described in the introduction, has the ability to cause significant destruction across a vast number of sectors. Economic losses can be felt through numerous impacts such as destruction of crops and degrading the quality (or completely destroying) health and transportation infrastructure. End-users (i.e., decision-makers, stakeholders) accordingly yearn for usable forecast products at extended ranges for extreme precipitation.

Statistical models typically can be designed for virtually any forecast product one could imagine and is the key starting point for any research-to-operations study. However, prior studies have shown that the co-production of knowledge between researchers (those who create a product) and end-users (those who use a product) result in products that are far more usable (Lemos et al. 2012). These products also tend to be viewed as more trustworthy (Meadow et al. 2015) and are easier to integrate into their existing workflows (Lemos et al. 2012). Therefore, the PRES²iP team hosted 3 stakeholder workshops to receive input from decision-makers who would be using our products in pursuance of making them as beneficial as possible. The first workshop took place in-person in July 2018 and aimed to establish a baseline for what stakeholders considered as “extreme precipitation” as well as what they did in current practice to manage extreme precipitation situations. The remainder of this Chapter focuses on the two other workshops. Workshop 2 was virtual in October 2021 and was concerned with how stakeholders interpret educational products made by the science team and how they dealt with forecast uncertainty. The second workshop linked

prior work done by the team (i.e., Chapter 2) while taking a brief look at forecast scenarios. Workshop 3 built on the ending of Workshop 2 in April 2023 and gauged the usefulness of a variety of forecast product types and further elucidated information on uncertainty. The “Methods” section presents the setup of each of the workshops before discussing the results immediately afterward. A brief set of conclusions follow.

7.2 Methods

7.2.1 Workshop 2

Workshop 2 took place virtually on 4–6 October 2021 and was comprised of 16 total stakeholders from backgrounds such as water management, emergency management, and tribal professionals. The team also invited two National Weather Service meteorologists from local forecast offices and one climatologist from the CPC. Leading up to the workshop, I created two fact sheets for the practitioners summarizing the research presented in Chapter 2. These fact sheets were focused on the creation of the database (Fig. 7.1) and another which focused on the climatology of extreme precipitation periods (Fig. 7.2).

The workshop also featured an interactive, role-playing exercise whereby participants were told to role-play as someone with their current position but in St. Louis, Missouri. The activity used a 14-day extreme precipitation period between 20 April and 3 May 2011 and forecast products were made using hindcasts from the U.K. Met Office S2S forecast system (MacLachlan et al. 2015). Extreme precipitation was defined as the long-term 95th percentile.

7.2.2 Workshop 3

Our third and final workshop took place in-person in Norman, Oklahoma 12–15 April 2023. Nine total practitioners of similar backgrounds as Workshop 2 were able to attend this workshop with four of the nine having attended at least one of the two prior workshops.



DEFINING A 14-DAY EXTREME EVENT

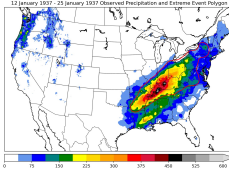
Between 1980 and 2019, there were 14 billion-dollar extreme precipitation-related disasters over periods longer than 14 days. But how do we **define** extreme over 14+ days?

- We analyze events in "**windows**"; 14-day blocks beginning on each calendar day.
- Thresholds for accumulation **and** duration are developed for **each window**.
- Spatial points meeting criteria grouped together to form **event polygon**.

Over an event window, a point must experience total precipitation greater than the 99th percentile. We are also building databases with other percentiles, such as the 95th. The point must also have at least 7 days of above normal **daily** precipitation. An algorithm to group points meeting both criteria results in an **extreme event polygon** (see right). Polygons are kept if area exceeds 200,000 km².

Database v1.0 spans 1915 – 2018 and events were grouped into clusters based on geographic location.

<http://pres2ip.com>

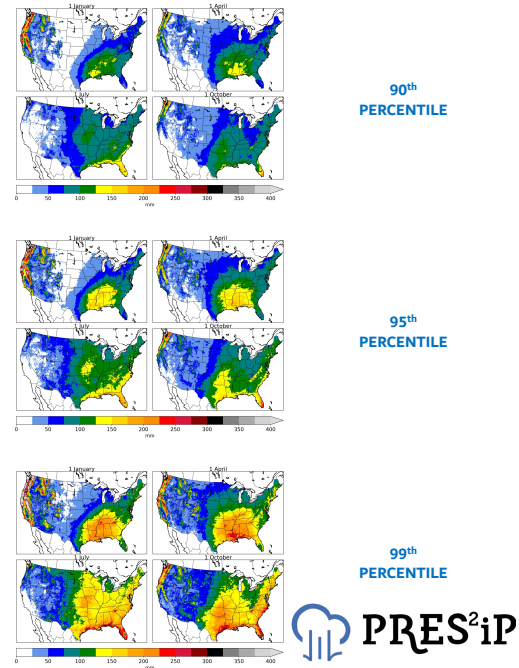


12 January 1917 - 25 January 1917 Observed Precipitation and Extreme Event Polygon

To see a list of events with corresponding images (like above) along with quick stats, see <http://pres2ip.com/extreme-event-tables>.



(a) Front



(b) Back

Figure 7.1: Fact sheet on defining a subseasonal extreme period.

This workshop focused on gathering feedback on pros and cons of different forecast products and to better understand how end-users would best interpret forecasts in an object-oriented sense; the discussion herein focuses solely on the latter.

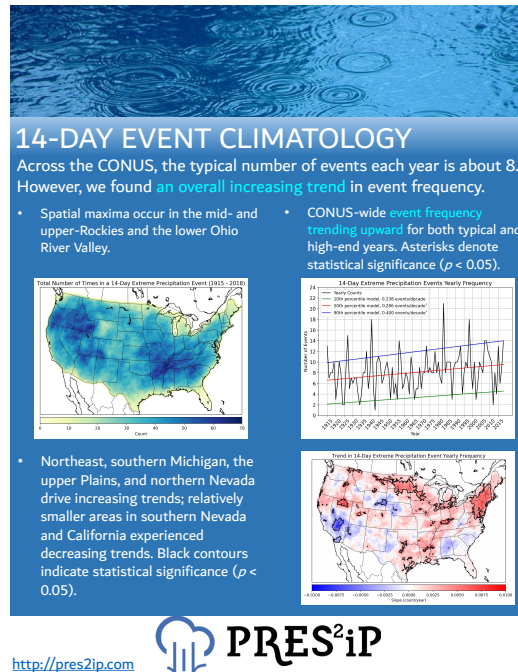


Figure 7.2: Fact sheet on the 14-day extreme period climatology.

7.3 Results

7.3.1 Fact sheets

Prior to the arrival of the participants, the PRES²iP team sent a pre-survey to gauge initial thoughts on the fact sheets. Eighty percent of those who responded said that they found these 2 fact sheets useful. Most participants stated that they liked the layout and organization of the fact sheets. Several stakeholders found the trends in the extreme precipitation periods to be very useful as it explicitly showed where changes were occurring and how intense the changes were. The participants were less enthusiastic, however, about the lack of context in the definition fact sheet. They desired factors such as monetary loss, infrastructure damage, and presidential disaster declarations to be included in the extreme period definition.

Additionally, some key concepts were unhelpful or confusing to the stakeholders. The concept of a sliding 14 day window could be difficult to convey to the public, some participants stated. Others could not understand why it was not a requirement for there to be

precipitation for every day in the 14 day window. Overall, this stemmed from a lack of continuity of background of the stakeholders. Their varying backgrounds caused them to have varying levels of understanding of scientific jargon and, thus, varying needs from the fact sheets. To address the wide variety of end-user needs, I moved revised versions of the fact sheets to our PRES²iP team website (<http://pres2ip.com/fact-sheetz/>). The webpage is split into two tabs with varying amounts of technical details to be more widely accessible to a wider target audience. Lastly, following the suggestions from the stakeholders, I replaced “extreme event” with “extreme period” and continue to use extreme period even in scientific settings since it better communicates the temporal scale of these extremes.

7.3.2 Forecast uncertainty

The final major session of Workshop 2 was centered on learning how practitioners would use and interpret deterministic versus probabilistic subseasonal forecasts. The activity began with the deterministic forecast and stakeholders were asked the question, “If you were given this graphic with a lead time of 7 days, would this deterministic product cause you to take any action?”. In response, 8 stakeholders said they would take actions such as notifying policymakers, messaging partners, and further monitoring the forecasts. Several of those who said they would not take action stated that they would continue to monitor the forecasts but they did not consider that to be an action taken in response to the forecast. Stakeholders said that additional context, such as precursor conditions and seasonal timing of the extreme precipitation period, would need to be considered when deciding how to further act.

Then stakeholders were next given a corresponding probabilistic forecast asked the same question as before. Despite the decrease in stakeholder response to this question, the number of those who would take action increased with 10 taking action and only 1 answering that they would not take action. They attributed this to the probabilistic forecast

highlighting an area of higher risk and the specificity in location(s) of maximum potential impact of the extreme precipitation period improved their decision making processes.

To understand how this probabilistic forecast influences the stakeholders' deterministic actions, PRES²iP team members probed to see which actions would be taken in association with which probabilities. Unsurprisingly, as probabilities of an extreme precipitation increased, the actions that the participants said they would take became more serious. Once probabilities reached 50%, the stakeholders said that they would sharpen their focus on the event and its timing. At 70%, the consensus was that preparatory actions would be taken. When considering lower probabilities, stakeholders were more likely to take action if the upcoming period was given historical context. In contrast, most participants indicated that being given the probability of the extreme precipitation period without further context, such as comparison to long-term climatology, was most helpful as it avoided confusion. Contrarily, however, one of the practitioners present in Workshop 3 (see below) but not Workshop 2 appreciated probabilities framed in historical context, further demonstrating the wide variety of preferences and needs among end-users.

Overall, participants agreed that the display of uncertainties gave this product more appeal as it allowed them to apply greater focus to the zone(s) with the greatest potential impact. However, several stakeholders commented on the difficulty in making deterministic actions from probabilistic forecasts. Furthermore, communicating the forecast and its uncertainty remained challenging to convey to other officials or the public. Despite the ability for the zone(s) to change as time progresses, they still favored this product as they can account for forecast trends when making their decision to take action. Multiple stakeholders indicate an inherent built-in expectation for spatial and/or temporal shifts in the forecast. Indeed, one stakeholder commented that shifts in the impact zone by 50 miles, for example, are relatively un-impactful, suggesting that an object-oriented verification framework would be more useful than traditional process-based evaluation metrics.

7.3.3 Object-oriented verification

The activity began with ranking theoretical forecast-observation pairs (Fig. 7.3) to introduce participants to forecast and observation contiguous regions. Each workshop participant was asked to rank the forecast quality from best to worst in the context of subseasonal extreme precipitation. 3 of the 7 stakeholders agreed with the original authors that panel (Fig. 7.3a) was the best, yet the remaining 4 participants put panel (Fig. 7.3e) as their top rank. Conversely, most stakeholders ranked panels (Fig. 7.3b) and (Fig. 7.3d) in the bottom 2 slots.

The activity transitioned next to presenting forecast-observation scenarios on an actual map. Participants were given hypothetical forecasts with a lead time of 3 days. The observed polygon employed was the true extreme precipitation region whereas forecast objects were adapted from various U-Net predictions (see Chapter 6). The first set of maps (Fig. 7.4) were centered on the spatial proximity of forecast-observation pairs and utilized polygons that had roughly similar total areas. In Fig. 7.4a, the distance between the two polygons' centroids was about 150 miles and only a relatively small area in central Missouri saw overlap between the forecast and observed regions. Nevertheless, all participants agreed the forecast was a “hit”. When the forecast was shifted further north and the polygons touched with minimal overlap (Fig. 7.4b), participants were more hesitant to say the forecast was a hit yet only 1 definitively said it was a miss. One stakeholder described (Figs. 7.4a,b) both as “really good” forecasts with others making comments such as “the

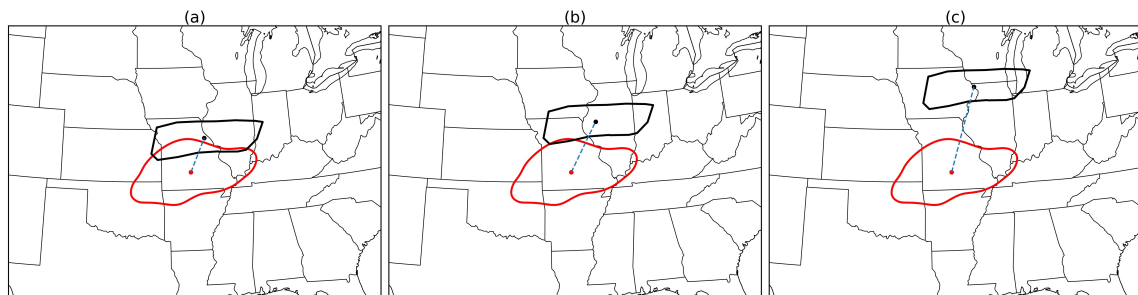


Figure 7.3: Schematic examples of hypothetical forecast-observation combinations. Taken from (Davis et al. 2019, Fig. 1)

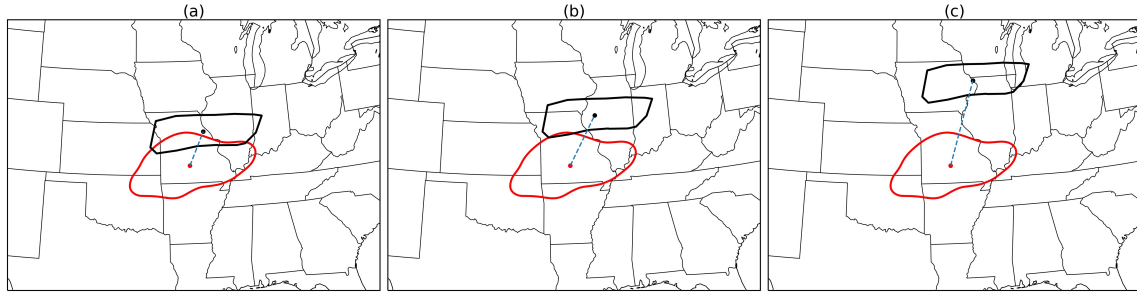


Figure 7.4: Various pairings of forecasts (black) and truth (red). Colored dots represent the polygon centroid. Distances between centroids are about (a) 150 miles, (b) 220 miles, and (c) 350 miles.

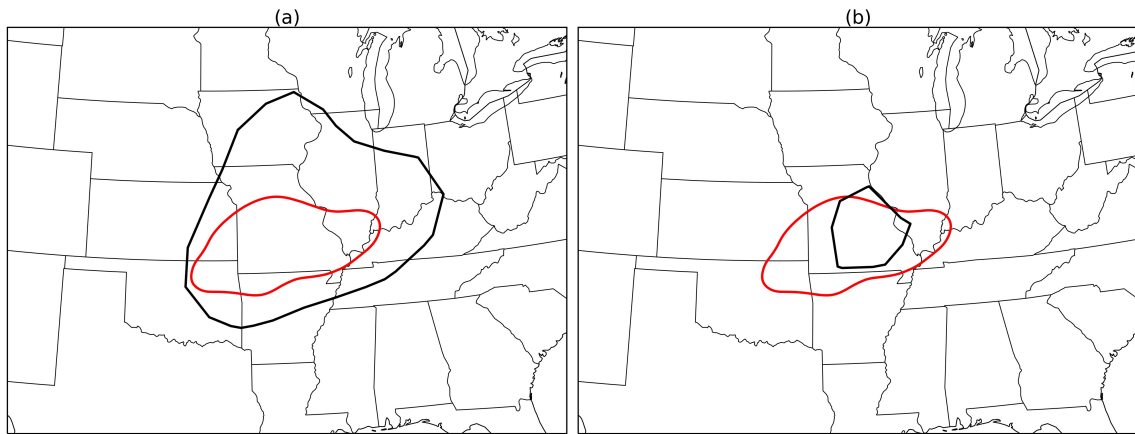


Figure 7.5: As in Fig. 7.4 but for scenarios with similar centers but different spatial scales.

forecast was not far off”, “good on a state level”, and “the forecast generally covers the same area...it is ok”. The furthest northward shift in the forecast made the centroid distance between polygons grow to about 350 mi (Fig. 7.4c) at which point all participants agreed the forecast was too far from the observed extreme period. Participants also noted that, all else being equal, the direction of spatial errors were unimportant, however changes in antecedent conditions, for example, could affect their tolerance to translations in the forecast.

The final part of the object-oriented forecast error activity compared the relative sizes of the forecast and observation regions whilst holding the distance between centroids relatively constant (~ 100 mi, Fig. 7.5). As before, the stakeholders were asked to ask if each forecast was considered a “hit”. In total, six out of seven labeled as a “hit” and generally preferred the larger forecast region in Fig. 7.5a, however one practitioner commented on Fig. 7.5b saying would use forecasts with higher precision if it was associated with higher likelihood. The general consensus was that a forecast like that in Fig. 7.5a would be a useful planning tool before using forecasts at shorter lead times to become more precise.

7.4 Conclusions

These workshops not only benefited the stakeholders and the PRES²iP team but also myself. By participating in the creation and production of multiple workshops, I was able to gain valuable skills that can be applied both in research and development settings. Specifically, by working with people from varying meteorological backgrounds, these workshops have improved my communication skills at all levels. I also learned how to receive input from those who use a product and apply the input to make the product more effective. My experiences on the PRES²iP team ultimately made me stand out compared to most graduate students and I am excited to continue pushing the boundaries of S2S prediction whilst ensuring any forecast advancements are actionable for end-users.

Chapter 8

Summary and Conclusions

Extreme precipitation affects life and property on a wide range of timescales and the present work laid some of the initial groundwork in subseasonal extreme precipitation. I first defined subseasonal extreme precipitation with two criteria: (1) its total accumulation must exceed the long-term 99th percentile, and (2) at least half of the period must have above normal daily precipitation. Criterion (2) was employed specifically to isolate extreme periods that were truly *subseasonal* and likely had relatively larger synoptic- and hemispheric-scale forcings. Grid points that met both criteria were grouped into a coherent extreme period region using KDE and a resulting polygon is considered to be a subseasonal extreme if its areal extent exceeded 200 000 km². An observationally-based database of 14-day subseasonal extreme precipitation periods was then generated with a period of record of 1915–2018. One of the strengths of my approach was that it did not define regions *a priori*, instead opting to use *k*-means clustering on the extreme periods to generate the regions more organically after the database was generated. Defining regions before identifying events ignores the processes that cause the event, whereas my method is more likely to group events that have similar large-scale characteristics. I also developed a novel index to objectively determine the optimal number of clusters and has the advantage of removing samples that do not fit.

In Chapter 3, I showed how the framework could seamlessly be applied to other datasets using ERA5 and the CNRM-CM6.1-HR historical simulation from the CMIP6 suite. Nearly the entirety of the framework discussed in Chapter 2 remained unchanged, demonstrating the resilience of my algorithm. The main changes involved the KDE step in making extreme period polygons and a linear bandwidth function based on the grid spacing of the

precipitation dataset was proposed. For datasets that have relatively large horizontal grid spacing, the KDE approach may become less viable and future works may consider using techniques such as kriging to develop extreme period polygons.

Although databases from all 3 datasets generally depict a local maximum in the Northwest U.S., the maxima in ERA5 and CNRM are located further west and north (Fig. 3.6b,c), however it is likely the geopolitical boundaries affecting Livneh are the major cause of this disagreement. CNRM also had a secondary maximum in North-Central California whereas the other two databases generally had decreasing frequencies as move moves along the West Coast. East of the Rocky Mountains, the 3 databases diverged more as ERA5 did not have a clear local maximum in the CONUS and CNRM had a large number of extreme periods in the Midwest. ERA5 also depicted a relatively few number of extreme periods across much of the Great Plains. Consistent with a plethora of previous literature (e.g., Kunkel et al. 1999; Groisman et al. 2004; Hoerling et al. 2016; Stegall and Kunkel 2019; Kirchmeier-Young and Zhang 2020; Crossett et al. 2023), my observationally-based database depicted an increasing trend in subseasonal extreme precipitation, particularly in the Northern Great Plains, Northeast US, and in western parts of the Southeast US. The CNRM historical run replicated the increasing trends in the Northern Great Plains and the Southeast reasonably well, yet had small, insignificant trends in the Northeast. ERA5 also showed the increasing trend in the Southeast, but otherwise had little association with the full period-of-record Livneh database. However, trends in many locations, such as along the Oregon coast and across the Midwest, changed appreciably after only considering 1950–2018 in the Livneh database (Fig. 3.8a), emphasizing the sensitivity of trends in subseasonal extreme precipitation to the chosen temporal period. Trends in CNRM over North America are not correlated with either map using the Livneh database (Figs. 3.7c and 3.8c), indicating the need to analyze a greater number of CMIP6 models.

The Livneh and CNRM databases were employed next to examine the large-scale characteristics that force subseasonal extreme precipitation. Trough-ridge dipoles were seen in composites for all regions of the CONUS (Fig. 4.1), although the position and orientation

were spatially dependent. The deepest trough was seen in the WC with a broad region of $< 1\sigma$ located in the eastern Pacific. Anomalous Rossby wave trains were commonly observed as composites showed a wavenumber 4 pattern in the MW, NP, SP, GL, and SE. Anomalous troughs were frequently located in parts of the Rockies with the exact positioning of both the immediate trough and ridge controlling the precipitation region. For instance, the NP and SP both have anomalous ridging downstream extending from Hudson Bay south into the Southeast US with the SP trough digging slightly further south. Similarly, the GL and SE regions have troughs in the easternmost parts of the Rockies co-located with anomalous high pressure in the northern Caribbean and off the US coast. Strong, anomalous specific humidity was also observed to support prolific and sustained precipitation periods (Figs. 4.5-4.6). Moisture anomalies are stronger and more centered on the region at 850 hPa compared to 250 hPa, however anomalies were relatively strong at both isobaric levels. In the WC, onshore flow is exceptionally strong and anomalous moisture is seen across much of the western CONUS. Similarly strong flow into each region another common feature across the clusters as the winds both supply anomalous moisture and support large-scale ascent through ageostrophic divergence aloft. Although the clusters developed herein were not exactly the same as in Jennrich et al. (2020) and Schroers (2020), two of the first studies on subseasonal extreme precipitation, the large-scale features were strikingly similar, lending confidence to the results. Furthermore, simulated extreme periods by CNRM were forced by similar patterns in the midlatitudes.

To address a lack of forecast products on subseasonal extreme precipitation, my work examined the ability of two ML models in predicting the extreme periods identified in the Livneh database. First, random forests were trained to forecast the likelihood of an subseasonal extreme precipitation day in a specific cluster. The forests, overall, were characterized by relatively low skill owing to a large number of false alarms. There was better predictability seen on the West Coast compared to the Great Plains and Ohio River Valley with an SEDS of nearly 0.5 in the testing dataset. After employing a relaxed verification scheme to account for “close” periods, the skill in each cluster rose with all models now

having an SEDS of >0.5 , although CSI scores remained <0.23 . Using detrended anomalies reduced the number of false alarms in the Plains and Ohio River Valley compared to inputting standardized anomalies, suggesting trends in the predictors (e.g., precipitable water) may impact the frequency and location of extreme periods moving forward. One major downside to the random forests in this framework, though, is its output: practitioners would desire more information than the random forests currently supply. Accordingly, the second predictive method employed was a type of CNN known as a U-Net (Ronneberger et al. 2015). U-Nets are an architecture that excel at receiving images as input and classifying each pixel in the image to a class. Here, following the preferences of stakeholders learned in Workshop 2 (Chapter 7), the output of the U-Net is the probability of subseasonal extreme precipitation. I tested a number of different architectures including the base U-Net, improvements suggested in the ML literature (He et al. 2016) and medical literature (Ok-tay et al. 2018), and my own developed block that uses differential depthwise convolutions (DDC). Although additions such as residual convolutions are substantially helpful in traditional image recognition, the results were mixed when applied to subseasonal extreme precipitation. There was no clear “best” architecture as each model had its own strengths and weaknesses. The U-Nets, evaluated using an object-oriented scheme, tended to fall in a regime with CSIs between 0.2 and 0.3, although the types of errors were not uniform. On one extreme, models like the base-bilin U-Net had a large number of false alarms yet had PODs > 0.6 when using both the 50% and 60% forecast probability thresholds; conversely, the residual-bilin U-Net traded a relatively high number of misses for better success ratios. The U-Net with the best CSI score utilized my DDC module and I hypothesize more complex improvements onto the U-Nets shown here will best work if using the DDC block. Forecasts from select U-Nets were analyzed for a number of case studies of different regions and forcing mechanisms. Using a 2-week composite as input, the models were shown to yield promising results, successfully identifying areas with increased likelihood of extreme precipitation.

To ensure that the work I have done will be useful in mitigating the damages done by S2S extreme precipitation, I had the chance to interact with decision-makers to whom accurate long-term precipitation forecasts would be useful. Over the course of multiple workshops, I and the rest of the PRES²iP team learned that stakeholders value context both within their education (fact-sheets) and their forecasts. The workshops also taught the team that the definition of successful forecast was highly application dependent, emphasizing the importance of understanding the target audience when making applied forecasts. The workshops proved the importance of conveying meteorological information in a variety of ways to address the variety of meteorological backgrounds of end-users.

Although the work here has provided a solid foundation, there is still a great deal to learn about subseasonal extreme precipitation, and the databases documented herein provide the perfect starting point for a plethora of studies. Firstly, the framework documented in Chapter 2 was designed for all timescales across the S2S scale and future works should continue to examine the large-scale precursors and characteristics during extreme periods of other durations. Secondly, the biggest weakness in the present versions of the U-Nets currently is in their ability to accurately predict subseasonal extreme snowfall events, particularly in the Mountain West. Heavy snowfall events in this region may be hard to resolve adequately on a 1.5° horizontal grid. A two-pronged ML approach which downscales subseasonal forecast output prior to an architecture like those presented in Chapter 6 may be a viable approach to improve snowfall forecasts for future experiments. Similarly, U-Nets dedicated to specific seasons (e.g., NDJFM) may better learn cool-season patterns producing subseasonal extreme precipitation, although the trade-off would be decreasing the number of extreme periods for the networks to be trained on.

Furthermore, architectures which explicitly consider all 14 days in the 2-week window instead of using 2-wk composite patterns may allow the U-Nets to be more spatiotemporally precise, although such models will be more complex and harder to interpret. One possible solution would be to change the 2D convolutions into 3D convolutions with learned weights for the temporal dimension; a second method utilizes so-called “time-distributed”

convolutions, which apply the same 2D kernel weights across all timesteps. As the duration of the extreme period increases, the importance of explicitly including temporal information also increases since inputting an average daily pattern will likely be much less useful for an extreme season compared to an extreme 2 weeks. Expanding the predictor set is another area of future work, especially as longer durations are considered. For instance, soil moisture is one important variable that has been left out here and its impact on modulating extremes may become more important as the length increases. Adding lagged predictors into the set may also bring increases in skill, although a domain increase may be needed to fully realize those teleconnections.

Reference List

- Abadi, M., and Coauthors, 2015: TensorFlow: Large-scale machine learning on heterogeneous systems. URL <https://www.tensorflow.org/>, software available from tensorflow.org.
- Agel, L., and M. Barlow, 2020: How Well Do CMIP6 Historical Runs Match Observed Northeast U.S. Precipitation and Extreme Precipitation–Related Circulation? *J. Climate*, **33**, 9835–9848, <https://doi.org/10.1175/JCLI-D-19-1025.1>.
- Agel, L., M. Barlow, S. Feldstein, and W. Gutowski Jr., 2018: Identification of large-scale meteorological patterns associated with extreme precipitation in the US northeast. *Climate Dyn.*, **50**, 1819–1839, <https://doi.org/10.1007/s00382-017-3724-8>.
- Agel, L., M. Barlow, J. Qian, F. Colby, E. Douglas, and T. Eichler, 2015: Climatology of Daily Precipitation and Extreme Precipitation Events in the Northeast United States. *J. Hydrometeor.*, **16**, 2537–2557, <https://doi.org/10.1175/JHM-D-14-0147.1>.
- Arthur, D., and S. Vassilvitskii, 2006: k-means++: The Advantages of Careful Seeding. Technical Report 2006-13, Stanford InfoLab. URL <http://ilpubs.stanford.edu:8090/778/>.
- Barlow, W., and Coauthors, 2019: North American extreme precipitation events and related large-scale meteorological patterns: A review of statistical methods, dynamics, modeling, and trends. *Climate Dyn.*, <https://doi.org/10.1007/s00382-019-04958-z>.
- Barnes, E., J. Hurrell, I. Ebert-Uphoff, C. Anderson, and D. Anderson, 2019: Viewing Forced Climate Patterns Through an AI Lens. *Geophys. Res. Lett.*, **46**, 13 389–13 398, <https://doi.org/10.1029/2019GL084944>.
- Baxter, S., and S. Nigam, 2015: Key Role of the North Pacific Oscillation–West Pacific Pattern in Generating the Extreme 2013/14 North American Winter. *J. Climate*, **28**, 8109–8117, <https://doi.org/10.1175/JCLI-D-14-00726.1>.
- Becker, E., B. Kirtman, and K. Pegion, 2020: Evolution of the North American Multi-Model Ensemble. *Geophys. Res. Lett.*, **47**, <https://doi.org/10.1029/2020GL087408>.
- Bluestein, H., 1992: *Synoptic-Dynamic Meteorology in Midlatitudes: Volume 1, Principles of Kinematics and Dynamics*.
- Bradley, A. A., and J. A. Smith, 1994: The hydrometeorological environment of extreme rainstorms in the Southern Plains of the United States. *J. Appl. Meteor. Climatol.*, **33**, 1418–1431, [https://doi.org/10.1175/1520-0450\(1994\)033<1418:THEOER>2.0.CO;2](https://doi.org/10.1175/1520-0450(1994)033<1418:THEOER>2.0.CO;2).
- Breiman, L., 2001: Random Forests. *Machine Learning*, **45**, 5–32, <https://doi.org/10.1023/A:1010933404324>.

- Bunker, R., 2020: Radar analysis of the physics of extreme rainfall events. M.S. thesis, School of Meteorology, University of Oklahoma.
- Cannon, A., S. Sobie, and T. Murdock, 2015: Bias Correction of GCM Precipitation by Quantile Mapping: How Well Do Methods Preserve Changes in Quantiles and Extremes? *J. Climate*, **28**, 6938–6959, <https://doi.org/10.1175/JCLI-D-14-00754.1>.
- Cavalcanti, I., N. Barreto, M. Alvarez, M. Osman, and C. Coelho, 2021: Teleconnection patterns in the Southern Hemisphere represented by ECMWF and NCEP S2S project models and influences on South America precipitation. *Meteorological Applications*, **28**, <https://doi.org/10.1002/met.2011>.
- Chase, R., D. Harrison, A. Burke, G. Lackmann, and A. McGovern, 2022: A Machine Learning Tutorial for Operational Meteorology. Part I: Traditional Machine Learning. *Wea. Forecasting*, **37**, 1509–1529, <https://doi.org/10.1175/WAF-D-22-0070.1>.
- Clark, A., G. J. W.A., M. Xue, and F. Kong, 2009: A Comparison of Precipitation Forecast Skill between Small Convection-Allowing and Large Convection-Parameterizing Ensembles. *Wea. Forecasting*, **24**, 1121–1140, <https://doi.org/10.1175/2009WAF22222.1>.
- CPC, 2023: Week 3-4 Outlooks. NOAA Climate Prediction Center, <https://www.cpc.ncep.noaa.gov/products/predictions/WK34/>.
- Cressie, N., 1990: The origins of kriging. *Mathematical Geology*, **22**, 239–252.
- Crossett, C., L.-A. Dupigny-Giroux, K. Kunkel, A. Betts, and A. Bombalies, 2023: Synoptic Typing of Multiduration, Heavy Precipitation Records in the Northeastern United States: 1895–2017. *J. Appl. Meteor. Climatol.*, **62**, 721–736, <https://doi.org/10.1175/JAMC-D-22-0091.1>.
- Daly, C., M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, and P. P. Pasteris, 2008: Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *Int. J. Climatol.*, **28**, 2031–2064, <https://doi.org/10.1002/joc.1688>.
- Daly, C., R. P. Neilson, and D. L. Phillips, 1994: A Statistical-Topographic Model for Mapping Climatological Precipitation over Mountainous Terrain. *J. Appl. Meteor.*, **33**, 140–158, [https://doi.org/10.1175/1520-0450\(1994\)033<0140:ASTMFM>2.0.CO;2](https://doi.org/10.1175/1520-0450(1994)033<0140:ASTMFM>2.0.CO;2).
- Davis, C., B. Brown, and R. Bullock, 2019: Object-Based Verification of Precipitation Forecasts. Part I: Methodology and Application to Mesoscale Rain Areas. *Mon. Wea. Rev.*, **134**, 1772–1784, <https://doi.org/10.1175/MWR3145.1>.
- Davis, J., and M. Goadrich, 2006: The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning*, 233–240.

- de Andrade, F., C. Coelho, and I. Cavalcanti, 2019: Global precipitation hindcast quality assessment of the Subseasonal to Seasonal (S2S) prediction project models. *Climate Dyn.*, **52**, 5451–5475, <https://doi.org/10.1007/s00382-018-4457-z>.
- de Andrade, F., M. Young, D. MacLeod, L. Hiron, S. Woolnough, and E. Black, 2021: Subseasonal Precipitation Prediction for Africa: Forecast Evaluation and Sources of Predictability. *Wea. Forecasting*, **36**, 265–284, <https://doi.org/10.1175/WAF-D-20-0054.1>.
- Dickinson, T., M. Richman, and J. Furtado, 2021: Subseasonal to Seasonal Extreme Precipitation Events in the Contiguous United States: Generation of a Database and Climatology. *J. Climate*, **34**, 7571–7586, <https://doi.org/10.1175/JCLI-D-20-0580.1>.
- Domeisen, D., and A. Butler, 2020: Stratospheric drivers of extreme events at the Earth’s surface. *Commun Earth Environ*, **1**, <https://doi.org/10.1038/s43247-020-00060-z>.
- Domeisen, D., A. Butler, K. Frohlich, M. Bittner, W. Muller, and J. Baehr, 2015: Seasonal Predictability over Europe Arising from El Niño and Stratospheric Variability in the MPI-ESM Seasonal Prediction System. *J. Climate*, **28**, 256–271, <https://doi.org/10.1175/JCLI-D-14-00207.1>.
- Domeisen, D., and Coauthors, 2019: The Role of the Stratosphere in Subseasonal to Seasonal Prediction: 2. Predictability Arising From Stratosphere-Troposphere Coupling. *JGR Atmospheres*, **125**, <https://doi.org/10.1029/2019JD030920>.
- Efron, B., 1987: Better bootstrap confidence intervals. *Journal of the American Statistical Association*, **82**, 171–185, <https://doi.org/10.1080/01621459.1987.10478410>.
- Epstein, E. S., and A. G. Barnston, 1990: A precipitation climatology of 5-day periods. *J. Climate*, **3**, 218–236, [https://doi.org/10.1175/1520-0442\(1990\)003<0218:APCOPD>2.0.CO;2](https://doi.org/10.1175/1520-0442(1990)003<0218:APCOPD>2.0.CO;2).
- Ester, M., H.-P. Kriegel, J. Sander, and X. Xu, 1996: A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Vol. 96, 226–231.
- Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, **9** (5), 1937–1958, <https://doi.org/10.5194/gmd-9-1937-2016>.
- Fan, Y., V. Krasnopolsky, H. van den Dool, C.-Y. Wu, and J. Gottschalck, 2023: Using Artificial Neural Networks to Improve CFS Week-3–4 Precipitation and 2-m Air Temperature Forecasts. *Wea. Forecasting*, **38**, 637–654, <https://doi.org/10.1175/WAF-D-20-0014.1>.
- Flanagan, P. X., J. B. Basara, J. C. Furtado, and X. Xiao, 2018: Primary atmospheric drivers of pluvial years in the United States Great Plains. *J. Hydrometeor.*, **19**, 643–658, <https://doi.org/10.1175/JHM-D-17-0148.1>.

- Forgy, E., 1965: Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *Biometrics*, **12**, 768–769.
- Gagne, D., S. Haupt, D. Nychka, and G. Thompson, 2019: Interpretable Deep Learning for Spatial Analysis of Severe Hailstorms. *Mon. Wea. Rev.*, **147**, 2827–2845, <https://doi.org/10.1175/MWR-D-18-0316.1>.
- Gagne, D., A. McGovern, and M. Xue, 2014: Machine Learning Enhancement of Storm-Scale Ensemble Probabilistic Quantitative Precipitation Forecasts. *Wea. Forecasting*, **29**, 1024–1043, <https://doi.org/10.1175/WAF-D-13-00108.1>.
- Gaspari, G., and S. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, **125**, 734–757, <https://doi.org/10.1002/qj.49712555417>.
- Gichamo, T., and C. Draper, 2022: An Optimal Interpolation–Based Snow Data Assimilation for NOAA’s Unified Forecast System (UFS). *Wea. Forecasting*, **37**, 2209–2221, <https://doi.org/10.1175/WAF-D-22-0061.1>.
- Gong, X., and M. B. Richman, 1995: On the application of cluster analysis to growing season precipitation data in North America east of the Rockies. *J. Climate*, **8** (4), 897–931, [https://doi.org/10.1175/1520-0442\(1995\)008<0897:OTAOCA>2.0.CO;2](https://doi.org/10.1175/1520-0442(1995)008<0897:OTAOCA>2.0.CO;2).
- Grimm, A., L. Hakoyama, and L. Scheibe, 2021: Active and break phases of the South American summer monsoon: MJO influence and subseasonal prediction. *Climate Dyn.*, **56**, 3603–3624, <https://doi.org/10.1007/s00382-021-05658-3>.
- Groisman, P. Y., R. W. Knight, T. R. Karl, D. R. Easterling, B. Sun, and J. H. Lawrimore, 2004: Contemporary changes of the hydrological cycle over the contiguous United States: Trends derived from *in situ* observations. *J. Hydrometeorol.*, **5**, 64–85, [https://doi.org/10.1175/1525-7541\(2004\)005<0064:CCOTHC>2.0.CO;2](https://doi.org/10.1175/1525-7541(2004)005<0064:CCOTHC>2.0.CO;2).
- Grotjahn, R., and G. Faure, 2008: Composite predictor maps of extraordinary weather events in the Sacramento, California, region. *Wea. Forecasting*, **23**, 313–335, <https://doi.org/10.1175/2007WAF2006055.1>.
- Guan, H., and Coauthors, 2022: GEFSv12 Reforecast Dataset for Supporting Subseasonal and Hydrometeorological Applications. *Mon. Wea. Rev.*, **150**, 647–665, <https://doi.org/10.1175/MWR-D-21-0245.1>.
- Hamill, T., D. Stovern, and L. Smith, 2023: Improving National Blend of Models Probabilistic Precipitation Forecasts Using Long Time Series of Reforecasts and Precipitation Reanalyses. Part I: Methods. *Mon. Wea. Rev.*, **151**, 1521–1534, <https://doi.org/10.1175/MWR-D-22-0308.1>.
- He, K., X. Zhang, S. Ren, and J. Sun, 2016: Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

- Henn, B., A. J. Newman, B. Livneh, C. Daly, and J. D. Lundquist, 2018: An assessment of differences in gridded precipitation datasets in complex terrain. *Journal of Hydrology*, **556**, 1205–1219, <https://doi.org/10.1016/j.jhydrol.2017.03.008>.
- Herman, G. R., and R. S. Schumacher, 2018: Money Doesn't Grow on Trees, but Forecasts Do: Forecasting Extreme Precipitation with Random Forests. *Mon. Wea. Rev.*, **146**, 1571–1600, <https://doi.org/10.1175/MWR-D-17-0250.1>.
- Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, <https://doi.org/10.1002/qj.3803>.
- Hill, A., and R. Schumacher, 2021: Forecasting Excessive Rainfall with Random Forests and a Deterministic Convection-Allowing Model. *Wea. Forecasting*, **36**, 1693–1711, <https://doi.org/10.1175/WAF-D-21-0026.1>.
- Hoerling, M., J. Eischeid, J. Perlwitz, X.-W. Quan, K. Wolter, and L. Cheng, 2016: Characterizing recent trends in U.S. heavy precipitation. *J. Climate*, **29**, 2313–2332, <https://doi.org/10.1175/JCLI-D-15-0441.1>.
- Hogan, R., E. O'Connor, and A. Illingworth, 2009: Verification of cloud-fraction forecasts. *Quart. J. Roy. Meteor. Soc.*, **135**, 1494–1511, <https://doi.org/10.1002/qj.481>.
- Hu, Y.-F., F.-K. Yin, and W.-M. Zhang, 2021: Deep learning-based precipitation bias correction approach for Yin-He global spectral model. *Meteorological Applications*, **28**, <https://doi.org/10.1002/met.2032>.
- Huang, Z., and Coauthors, 2022: A seven-parameter Bernoulli-Gamma-Gaussian model to calibrate subseasonal to seasonal precipitation forecasts. *J. Hydrology*, **610**, <https://doi.org/10.1016/j.jhydrol.2022.127896>.
- Ikeda, K., M. Steiner, J. Pinto, and C. Alexander, 2013: Evaluation of Cold-Season Precipitation Forecasts Generated by the Hourly Updating High-Resolution Rapid Refresh Model. *Wea. Forecasting*, **28**, 921–939, <https://doi.org/10.1175/WAF-D-12-00085.1>.
- Ines, A., and J. Hansen, 2006: Bias correction of daily GCM rainfall for crop simulation studies. *Agric. For. Meteorol.*, **138**, 44–53, <https://doi.org/10.1016/j.agrformet.2006.03.009>.
- Ioffe, S., and C. Szegedy, 2015: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proceedings of the 32nd International Conference on Machine Learning*, <https://doi.org/10.48550/arXiv.1502.03167>.
- Iyer, E. R., A. J. Clark, M. Xue, and F. Kong, 2016: A comparison of 36–60-h precipitation forecasts from convection-allowing and convection-parameterizing ensembles. *Wea. Forecasting*, **31**, 647–661, <https://doi.org/10.1175/WAF-D-15-0143.1>.
- Jennrich, G. C., J. C. Furtado, J. B. Basara, and E. R. Martin, 2020: Synoptic characteristics of 14-Day extreme precipitation events across the United States. *J. Climate*, **33**, 6423–6440, <https://doi.org/10.1175/JCLI-D-19-0563.1>.

- Johnson, A., X. Wang, Y. Wang, A. Reinhart, A. Clark, and I. Jirak, 2020: Neighborhood- and Object-Based Probabilistic Verification of the OU MAP Ensemble Forecasts during 2017 and 2018 Hazardous Weather Testbeds. *Wea. Forecasting*, **35**, 169–191, <https://doi.org/10.1175/WAF-D-19-0060.1>.
- Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77**, 437–472, [https://doi.org/10.1175/1520-0477\(1996\)077<0437:TNYRP>2.0.CO;2](https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2).
- Kautz, L.-A., O. Martius, S. Pfahl, J. Pinto, A. Ramos, P. Sousa, and T. Woollings, 2022: Atmospheric blocking and weather extremes over the Euro-Atlantic sector – a review. *Weather and Climate Dynamics*, **3**, 1819–1839, <https://doi.org/10.5194/wcd-3-305-2022>.
- Keim, B. D., 1996: Spatial, synoptic, and seasonal patterns of heavy rainfall in the Southeastern United States. *Physical Geography*, **17**, 313–328, <https://doi.org/10.1080/02723646.1996.10642588>.
- Kingma, D., and J. Ba, 2015: Adam: A Method for Stochastic Optimization. *3rd International Conference for Learning Representations, San Diego*.
- Kirchmeier-Young, M. C., and X. Zhang, 2020: Human influence has intensified extreme precipitation in North America. *Proc. Natl. Acad. Sci. (USA)*, <https://doi.org/10.1073/pnas.1921628117>.
- Kirkyla, K. I., and S. Hameed, 1989: Harmonic analysis of the seasonal cycle in precipitation over the United States: A comparison between observations and a general circulation model. *J. Climate*, **2**, 1463–1475, [https://doi.org/10.1175/1520-0442\(1989\)002<1463:HAOTSC>2.0.CO;2](https://doi.org/10.1175/1520-0442(1989)002<1463:HAOTSC>2.0.CO;2).
- Kirtman, B. P., and Coauthors, 2014: The North American Multimodel Ensemble: Phase-1 seasonal-to-interannual prediction; Phase-2 toward developing intraseasonal prediction. *Bull. Amer. Meteor. Soc.*, **95** (4), 585–601, <https://doi.org/10.1175/BAMS-D-12-00050.1>.
- Klemm, T., and R. A. McPherson, 2017: The development of seasonal forecasting for agricultural producers. *Agricultural and Forest Meteorology*, **232**, 384–399, <https://doi.org/10.1016/j.agrformet.2016.09.005>.
- Knapp, A. K., and Coauthors, 2008: Consequences of more extreme precipitation regimes for terrestrial ecosystems. *BioScience*, **58**, 811–821, <https://doi.org/10.1641/B580908>.
- Koenker, R., and G. Bassett Jr., 1978: Regression Quantiles. *Econometrica*, **46** (1), 33–50.
- Kolachian, R., and B. Saghaian, 2019: Deterministic and probabilistic evaluation of raw and post processed sub-seasonal to seasonal precipitation forecasts in different precipitation regimes. *Theor. Appl. Climatol.*, **137**, 1479–1493, <https://doi.org/10.1007/s00704-018-2680-5>.

- Konrad, C. E., 2001: The most extreme precipitation events over the eastern United States from 1950 to 1996: Considerations of scale. *J. Hydrometeor.*, **2** (3), 309–325, [https://doi.org/10.1175/1525-7541\(2001\)002<0309:TMEPEO>2.0.CO;2](https://doi.org/10.1175/1525-7541(2001)002<0309:TMEPEO>2.0.CO;2).
- Krishnamurphy, V., and Coauthors, 2021: Sources of Subseasonal Predictability over CONUS during Boreal Summer. *J. Climate*, **34**, 3273–3294, <https://doi.org/10.1175/JCLI-D-20-0586.1>.
- Kunkel, K. E., K. Andsager, and D. R. Easterling, 1999: Long-term trends in extreme precipitation events over the conterminous United States and Canada. *J. Climate*, **12**, 2515–2527, [https://doi.org/10.1175/1520-0442\(1999\)012<2515:LTTIEP>2.0.CO;2](https://doi.org/10.1175/1520-0442(1999)012<2515:LTTIEP>2.0.CO;2).
- Kunkel, K. E., D. R. Easterling, D. A. Kristovich, B. Gleason, L. Stoecker, and R. Smith, 2012: Meteorological causes of the secular variations in observed extreme precipitation events for the conterminous United States. *J. Hydrometeor.*, **13**, 1131–1141, <https://doi.org/10.1175/JHM-D-11-0108.1>.
- Lagerquist, R., A. McGovern, and D. Gagne, 2019: Deep Learning for Spatially Explicit Prediction of Synoptic-Scale Fronts. *Wea. Forecasting*, **34**, 1137–1160, <https://doi.org/10.1175/WAF-D-18-0183.1>.
- Lagerquist, R., A. McGovern, C. Homeyer, D. Gagne, and T. Smith, 2020: Deep learning on three-dimensional multiscale data for next-hour tornado prediction. *Mon. Wea. Rev.*, **148**, 2837–2861, <https://doi.org/10.1175/MWR-D-19-0372.1>.
- Lagerquist, R., J. Stewart, I. Ebert-Uphoff, and C. Kumler, 2021: Using Deep Learning to Nowcast the Spatial Coverage of Convection from Himawari-8 Satellite Data. *Mon. Wea. Rev.*, **149**, 3897–3921, <https://doi.org/10.1175/MWR-D-21-0096.1>.
- Lee, C., S. Sheridan, G. Dusek, and D. Pirhalla, 2021: Atmospheric pattern-based predictions of S2S sea-level anomalies for two selected US locations. *Artif. Intell. Earth Syst.*, 1–38, <https://doi.org/10.1175/AIES-D-22-0057.1>.
- Lemos, M., C. Kirchhoff, and V. Ramprasad, 2012: Narrowing the climate information usability gap. *Nature Clim Change*, **2**, 789–794, <https://doi.org/10.1038/nclimate1614>.
- Li, Q., S. Yang, T. Wu, and X. Liu, 2017: Subseasonal Dynamical Prediction of East Asian Cold Surges. *Wea. Forecasting*, **32**, 637–654, <https://doi.org/10.1175/WAF-D-16-0209.1>.
- Li, Y., Z. Wu, H. He, Q. Wang, H. Xu, and G. Lu, 2021: Post-processing sub-seasonal precipitation forecasts at various spatiotemporal scales across China during boreal summer monsoon. *J. Hydrology*, **598**, <https://doi.org/10.1016/j.hydrol.2020.125742>.
- Li, Y., Z. Wu, H. He, and H. Yin, 2022: Probabilistic subseasonal precipitation forecasts using preceding atmospheric intraseasonal signals in a Bayesian perspective. *Hydrology and Earth System Sciences*, **26**, 4975–4994, <https://doi.org/10.5194/hess-26-4975-2022>.

- Lin, T., P. Goyal, R. Girshick, K. He, and P. Dollár, 2017: Focal loss for dense object detection. *Proceedings of the IEEE international conference on computer vision*, 2980–2988.
- Livneh, B., E. A. Rosenberg, C. Lin, B. Nijssen, V. Mishra, K. M. 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: Update and Extensions. *J. Climate*, **26**, 9384–9392, <https://doi.org/10.1175/JCLI-D-12-00508.1>.
- Lloyd, S. P., 1982: Least squares quantization in CPM. *IEEE Transactions on Information Theory*, **28**, 129–136.
- Loken, E., A. Clark, A. McGovern, M. Flora, and K. Knopfmeier, 2019: Postprocessing Next-Day Ensemble Probabilistic Precipitation Forecasts Using Random Forests. *Wea. Forecasting*, **34**, 2017–2044, <https://doi.org/10.1175/WAF-D-19-0109.1>.
- Lopez, M. G., H. Wennerström, L.-a. Nordén, and J. Seibert, 2015: Location and density of rain gauges for the estimation of spatial varying precipitation. *Geografiska Annaler: Series A, Physical Geography*, **97** (1), 167–179, <https://doi.org/10.1111/geoa.12094>.
- MacLachlan, C., and Coauthors, 2015: Global Seasonal forecast system version 5 (GloSea5): a high-resolution seasonal forecast system. *Quart. J. Roy. Meteor. Soc.*, **141**, 1072–1084, <https://doi.org/10.1002/qj.2396>.
- Maddox, R., C. Chappell, and L. Hoxit, 1979: Synoptic and Meso-alpha Scale Aspects of Flash Flood Events. *Bull. Amer. Meteor. Soc.*, **60**, 115–123, <https://doi.org/10.1175/1520-0477-60.2.115>.
- Mayer, K., and E. Barnes, 2021: Subseasonal Forecasts of Opportunity Identified by an Explainable Neural Network. *Geophys. Res. Lett.*, **48**, <https://doi.org/10.1029/2020GL092092>.
- McAfee, M., Devin, E. Martin, and J. Furtado, 2023: Bias Correction and Evaluation of S2S Model Performance for Forecasting US Extreme Precipitation Events. *103rd Annual Meeting*, Denver, CO, Amer. Meteor. Soc., S151, <https://ams.confex.com/ams/103ANNUAL/meetingapp.cgi/Paper/412553>.
- McGovern, A., K. Elmore, D. Gagne, S. Haupt, C. Karstens, R. Lagerquist, T. Smith, and J. Williams, 2017: Using Artificial Intelligence to Improve Real-Time Decision-Making for High-Impact Weather. *Bull. Amer. Meteor. Soc.*, **98**, 2073–2090, <https://doi.org/10.1175/BAMS-D-16-0123.1>.
- Meadow, A., D. Ferguson, Z. Guido, A. Horangic, G. Owen, and T. Wall, 2015: Moving toward the Deliberate Coproduction of Climate Science Knowledge. *Weather, Climate, and Society*, **7**, 179–191, <https://doi.org/10.1175/WCAS-D-14-00050.1>.
- Mercer, A. E., C. M. Shafer, C. A. Doswell, L. M. Leslie, and M. B. Richman, 2012: Synoptic composites of tornadic and nontornadic outbreaks. *Mon. Wea. Rev.*, **140** (8), 2590–2608, <https://doi.org/10.1175/MWR-D-12-00029.1>.

- Mesinger, F., and Coauthors, 2006: North American Regional Reanalysis. *Bull. Amer. Meteor. Soc.*, **87**, 343–360, <https://doi.org/10.1175/BAMS-87-3-343>.
- Milrad, S., E. Atallah, J. Gyakum, and G. Dookhie, 2014: Synoptic Typing and Precursors of Heavy Warm-Season Precipitation Events at Montreal, Québec. *Wea. Forecasting*, **29**, 419–444, <https://doi.org/10.1175/WAF-D-13-00030.1>.
- Moore, B. J., K. M. Mahoney, E. M. Sukovich, R. Cifelli, and T. M. Hamill, 2015: Climatology and environmental characteristics of extreme precipitation events in the Southeastern United States. *Mon. Wea. Rev.*, **143**, 718–741, <https://doi.org/10.1175/MWR-D-14-00065.1>.
- Moore, B. J., P. J. Neiman, F. M. Ralph, and F. E. Barthold, 2012: Physical processes associated with heavy flooding rainfall in Nashville, Tennessee, and vicinity during 1–2 May 2010: The role of an atmospheric river and mesoscale convective systems. *Mon. Wea. Rev.*, **140**, 358–378, <https://doi.org/10.1175/MWR-D-11-00126.1>.
- NCEI, 2020: U.S. Billion-Dollar Weather and Climate Disasters. NOAA National Center for Environmental Information, <https://www.ncdc.noaa.gov/billions/>.
- Nguyen, P., M. Ombadi, S. Sorooshian, K. Hsu, A. AghaKouchak, D. Braithwaite, H. Ashouri, and A. R. Thorstensen, 2018: The PERSIANN family of global satellite precipitation data: a review and evaluation of products. *Hydrology and Earth System Sciences*, **22**, 5801–5816, <https://doi.org/10.5194/hess-22-5801-2018>.
- NOAA, 2018: Subseasonal and Seasonal Forecasting Innovation: Plans for the Twenty-First Century. Tech. rep., National Oceanic and Atmospheric Administration. Available online at https://www.weather.gov/media/sti/S2S/Annotated%20Outline%20for%20Public%20Release_7_12_18.pdf.
- Odena, A., V. Dumoulin, and C. Olah, 2016: Deconvolution and Checkerboard Artifacts. *Distill*, <https://doi.org/10.23915/distill.00003>, URL <http://distill.pub/2016/deconv-checkerboard>.
- Oktay, O., and Coauthors, 2018: 1st Conference on Medical Imaging with Deep Learning (MIDL 2018). *1st Conference on Medical Imaging with Deep Learning (MIDL 2018)*, <https://doi.org/10.48550/arXiv.1804.03999>.
- Orth, R., and S. Seneviratne, 2014: Using soil moisture forecasts for sub-seasonal summer temperature predictions in Europe. *Climate Dyn.*, **43**, 3403–3418, <https://doi.org/10.1007/s00382-014-2112-x>.
- Osborne, A., J. Zhang, M. Simpson, K. Howard, and S. Cocks, 2022: Application of Machine Learning Techniques to Improve Multi-Radar Multi-Sensor (MRMS) Precipitation Estimates in the Western United States. *Artif. Intell. Earth Syst.*, **2**, <https://doi.org/10.1175/AIES-D-22-0053.1>.
- Pan, B., K. Hsu, A. AghaKouchak, S. Sorooshian, and W. Higgins, 2019: Precipitation prediction skill for the West Coast United States: From short to extended range. *J. Climate*, **32**, 161–182, <https://doi.org/10.1175/JCLI-D-18-0355.1>.

- Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.
- Pegion, K., and Coauthors, 2019: The Subseasonal Experiment (SubX): A Multi-model Subseasonal Prediction Experiment. *Bull. Amer. Meteor. Soc.*, **100**, 2043–2060, <https://doi.org/10.1175/BAMS-D-18-0270.1>.
- Piao, S., and Coauthors, 2010: The impacts of climate change on water resources and agriculture in China. *Nature*, **467**, 43–51, <https://doi.org/10.1038/nature09364>.
- Richman, M. B., and I. Adrianto, 2010: Classification and regionalization through kernel principal component analysis. *Physics and Chemistry of the Earth, Parts A/B/C*, **35** (9), 316–328, <https://doi.org/10.1016/j.pce.2010.02.001>.
- Rigby, R., and D. Stasinopoulos, 2005: Generalized additive models for location, scale and shape. *Appl. Statist.*, **54**, 507–554, <https://doi.org/10.1111/j.1467-9876.2005.00510.x>.
- Rodney, M., H. Lin, and J. Derome, 2013: Subseasonal Prediction of Wintertime North American Surface Air Temperature during Strong MJO Events. *Mon. Wea. Rev.*, **141**, 2897–2909, <https://doi.org/10.1175/MWR-D-12-00221.1>.
- Ronneberger, O., P. Fischer, and T. Brox, 2015: U-Net: Convolutional Networks for Biomedical Image Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241.
- Rosenblatt, M., 1956: Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, **27** (3), 832–837, <https://doi.org/10.1214/aoms/1177728190>.
- Rousseeuw, P. J., 1987: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53–65, [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).
- Saha, S., and Coauthors, 2014: The NCEP Climate Forecast System Version 2. *J. Climate*, **27**, 2185–2208, <https://doi.org/10.1175/JCLI-D-12-00823.1>.
- Saint-Martin, D., and Coauthors, 2020: Evaluation of CMIP6 DECK Experiments With CNRM-CM6-1. *J. Advances in Modeling Earth Systems*, **13**, <https://doi.org/10.1029/2020MS002190>.
- Schaefer, J. T., 1990: The Critical Success Index as an Indicator of Warning Skill. *Wea. Forecasting*, **5**, 570–575, [https://doi.org/10.1175/1520-0434\(1990\)005<0570:TCSIAA>2.0.CO;2](https://doi.org/10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2).
- Schaeffer, R., and Coauthors, 2012: Energy sector vulnerability to climate change: A review. *Energy*, **38**, 1–12, <https://doi.org/10.1016/j.energy.2011.11.056>.
- Scheuerer, M., M. Switanek, R. Worsnop, and T. Hamill, 2020: Using Artificial Neural Networks for Generating Probabilistic Subseasonal Precipitation Forecasts over California. *Mon. Wea. Rev.*, **148**, 3489–3506, <https://doi.org/10.1175/MWR-D-20-0096.1>.

- Schroers, M., 2020: Potential Predictability and Impacts of Subseasonal Extreme Precipitation Events in the United States. M.S. thesis, School of Meteorology, University of Oklahoma.
- Schroers, M., and E. Martin, 2022: Synoptic Connections and Impacts of 14-Day Extreme Precipitation Events in the United States. *J. Appl. Meteor. Climatol.*, **61**, 877–890, <https://doi.org/10.1175/JAMC-D-21-0174.1>.
- Schumacher, R. S., and R. H. Johnson, 2005: Organization and environmental properties of extreme-rain-producing mesoscale convective systems. *Mon. Wea. Rev.*, **133**, 961–976, <https://doi.org/10.1175/MWR2899.1>.
- Schumacher, R. S., and R. H. Johnson, 2006: Characteristics of U.S. extreme rain events during 1999–2000. *Wea. Forecasting*, **21** (1), 69–85, <https://doi.org/10.1175/WAF900.1>.
- Scott, C. M., and M. D. Shulman, 1979: An areal and temporal analysis of precipitation in the Northeastern United States. *J. Appl. Meteor. Climatol.*, **18**, 627–633, [https://doi.org/10.1175/1520-0450\(1979\)018<0627:AAATAO>2.0.CO;2](https://doi.org/10.1175/1520-0450(1979)018<0627:AAATAO>2.0.CO;2).
- Sha, Y., D. Gagne, G. West, and R. Stull, 2022: A Hybrid Analog-Ensemble-Convolutional-Neural-Network Method for Postprocessing Precipitation Forecasts. *Mon. Wea. Rev.*, **150**, 1495–1515, <https://doi.org/10.1175/MWR-D-21-0154.1>.
- Shah, R., A. Sahai, and V. Mishra, 2017: Short to sub-seasonal hydrologic forecast to manage water and agricultural resources in India. *Hydrol. Earth Syst. Sci.*, **21**, 707–720, <https://doi.org/10.5194/hess-21-707-2017>.
- Slinsky, E. A., P. C. Loikith, D. E. Waliser, and A. Goodman, 2019: An extreme precipitation categorization scheme and its observational uncertainty over the continental United States. *J. Climate*, **20**, 1029–1052, <https://doi.org/10.1175/JHM-D-18-0148.1>.
- Specq, D., and L. Batté, 2020: Improving subseasonal precipitation forecasts through a statistical-dynamical approach : application to the southwest tropical Pacific. *Climate Dynamics*, **55**, 1913–1927, <https://doi.org/10.1007/s00382-020-05355-7>.
- Srivastava, A., R. Grotjahn, and P. Ullrich, 2020: Evaluation of historical CMIP6 model simulations of extreme precipitation over contiguous US regions. *Weather and Climate Extremes*, **29**, <https://doi.org/10.1016/j.wace.2020.100268>.
- Stegall, S. T., and K. E. Kunkel, 2019: Simulation of daily extreme Precipitation over the United States in the CMIP5 30-Yr Decadal Prediction Experiment. *J. Appl. Meteor. Climatol.*, **58**, 875–886, <https://doi.org/10.1175/JAMC-D-18-0057.1>.
- Stevenson, S., and R. Schumacher, 2013: A 10-Year Survey of Extreme Rainfall Events in the Central and Eastern United States Using Gridded Multisensor Precipitation Analyses. *Mon. Wea. Rev.*, **142**, 3147–3162, <https://doi.org/10.1175/MWR-D-13-00345.1>.

- Stovern, D., T. Hamill, and L. Smith, 2023: Improving National Blend of Models Probabilistic Precipitation Forecasts Using Long Time Series of Reforecasts and Precipitation Reanalyses. Part II: Results. *Mon. Wea. Rev.*, **151**, 1535–1559, <https://doi.org/10.1175/MWR-D-22-0310.1>.
- Straaten, C., K. Whan, D. Coumou, B. Hurk, and M. Schmeits, 2023: Correcting sub-seasonal forecast errors with an explainable ANN to understand misrepresented sources of predictability of European summer temperatures. *Artif. Intell. Earth Syst.*, 1–49, <https://doi.org/10.1175/AIES-D-22-0047.1>.
- Sukovich, E. M., F. Ralph, F. E. Barthold, D. W. Reynolds, and D. R. Novak, 2014: Extreme quantitative precipitation forecast performance at the Weather Prediction Center from 2001 to 2011. *Wea. Forecasting*, **29**, 894–911, <https://doi.org/10.1175/WAF-D-13-00061.1>.
- Sun, L., H. Hoerling, J. Richter, A. Hoell, A. Kumar, and J. Hurrell, 2022: Attribution of North American Subseasonal Precipitation Prediction Skill. *Wea. Forecasting*, **37**, 2069–2085, <https://doi.org/10.1175/WAF-D-22-0076.1>.
- Thorndike, R. L., 1953: Who belongs in the family? *Psychometrika*, **18** (4), 267—276, <https://doi.org/10.1007/BF02289263>.
- VanBuskirk, O., T. Dickinson, M. Schroers, R. McPherson, and E. Martin, 2023: Listening to Stakeholders II: Adapting Research Products on Subseasonal to Seasonal Heavy Precipitation Events by Exploring Options with Users. *Bull. Amer. Meteor. Soc.*, **104**, E781–E791, <https://doi.org/10.1175/BAMS-D-22-0229.1>.
- Vitart, F., and Coauthors, 2008: The new VarEPS-monthly forecasting system: A first step towards seamless prediction. *Quart. J. Roy. Meteor. Soc.*, **134**, 1789–1799, <https://doi.org/10.1002/qj.322>.
- Vitart, F., and Coauthors, 2017: The Subseasonal to Seasonal (S2S) Prediction Project Database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, <https://doi.org/10.1175/BAMS-D-16-0017.1>.
- Voltaire, A., and Coauthors, 2019: Evaluation of CMIP6 DECK Experiments With CNRM-CM6-1. *J. Advances in Modeling Earth Systems*, **11**, 2177–2213, <https://doi.org/10.1029/2019MS001683>.
- Wallace, J., and D. Gutzler, 1981: Teleconnections in the Geopotential Height Field during the Northern Hemisphere Winter. *Mon. Wea. Rev.*, **109**, 784–812, [https://doi.org/10.1175/151520-0493\(1981\)109<0784:TITGHF>2.0.CO;2](https://doi.org/10.1175/151520-0493(1981)109<0784:TITGHF>2.0.CO;2).
- Warner, M. D., C. F. Mass, and E. P. Salathé Jr., 2012: Wintertime extreme precipitation events along the Pacific Northwest Coast: Climatology and synoptic evolution. *Mon. Wea. Rev.*, **140**, 2021–2043, <https://doi.org/10.1175/MWR-D-11-00197.1>.
- Weather Research and Forecasting Innovation Act of 2017, 2017: H.R.353, 115th Congress.

- Whan, K., and M. Schmeits, 2018: Comparing Area Probability Forecasts of (Extreme) Local Precipitation Using Parametric and Machine Learning Statistical Postprocessing Methods. *Mon. Wea. Rev.*, **146**, 3651–3673, <https://doi.org/10.1175/MWR-D-17-0290.1>.
- White, C. J., and Coauthors, 2017: Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteorological Applications*, **24**, 315–325, <https://doi.org/10.1002/met.1654>.
- Wilks, D., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed., Academic Press.
- Wolter, K., and M. Timlin, 2011: El Niño/Southern Oscillation behaviour since 1871 as diagnosed in an extended multivariate ENSO index (MEI.ext). *International Journal of Climatology*, **31**, 1074–1087, <https://doi.org/10.1002/joc.2336>.
- Yan, M., and Y. Guo, 2023: Influence of ENSO on the ECMWF subseasonal prediction of summer rainfall over the Yangtze River. *Climate Dyn.*, **145**, <https://doi.org/10.1007/s00382-023-06784-w>.
- Zhang, L., T. Kim, T. Yang, Y. Hong, and Q. Zhu, 2021: Evaluation of Subseasonal-to-Seasonal (S2S) precipitation forecast from the North American Multi-Model ensemble phase II (NMME-2) over the contiguous U.S. *J. Hydrology*, **603**, <https://doi.org/10.1016/j.jhydrol.2021.127058>.
- Zhao, S., Y. Deng, and R. X. Black, 2017: A dynamical and statistical characterization of U.S. extreme precipitation events and their associated large-scale meteorological patterns. *J. Climate*, **30**, 1307–1326, <https://doi.org/10.1175/JCLI-D-15-0910.1>.
- Zhuang, J., R. Dussin, A. Jüling, and S. Rasp, 2020: JiaweiZhuang/xESMF: v0.3.0 Adding ESMF.LocStream capabilities. URL <https://doi.org/10.5281/zenodo.3700105>.

1 Appendix A

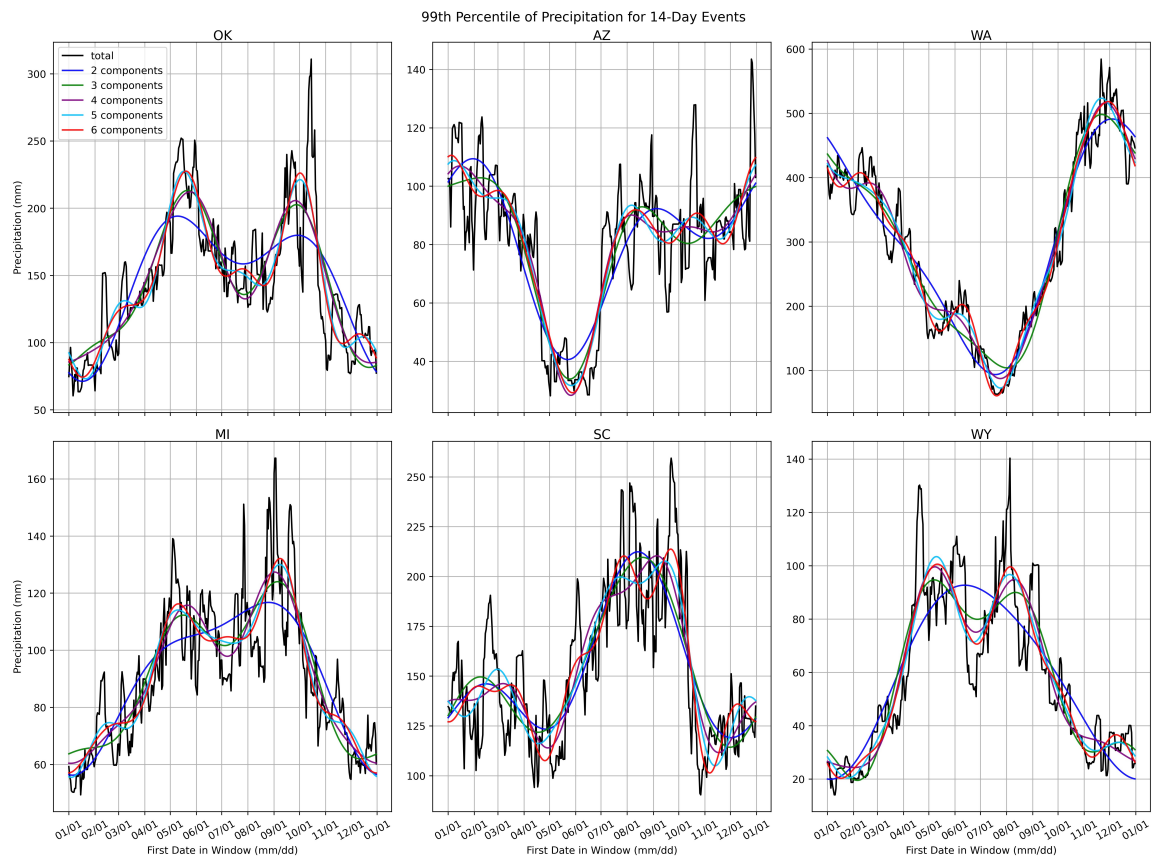


Figure A.1: As in Fig. 2.2, but with using (dark blue) 2 harmonics, (green) 3 harmonics, (purple) 4 harmonics, (light blue) 5 harmonics, and (red) 6 harmonics.

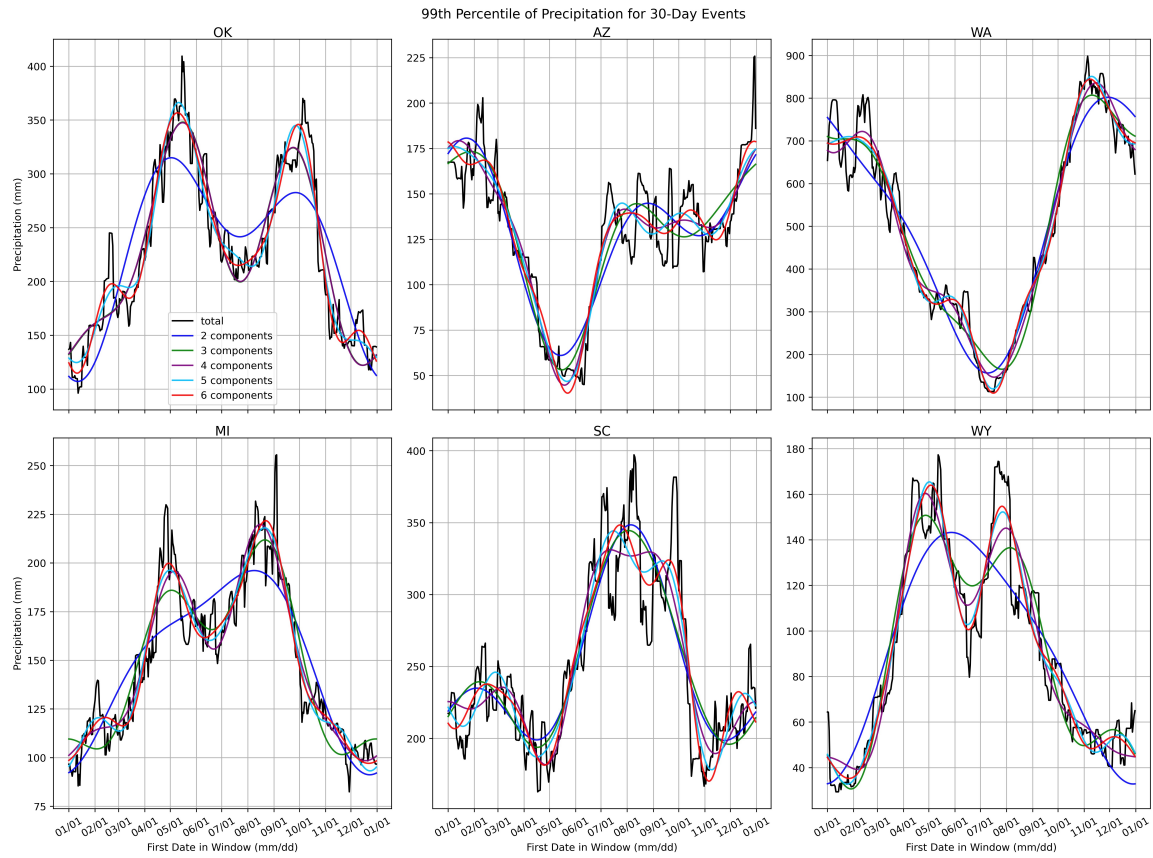


Figure A.2: As in Figure A.1, but for 30-day events.

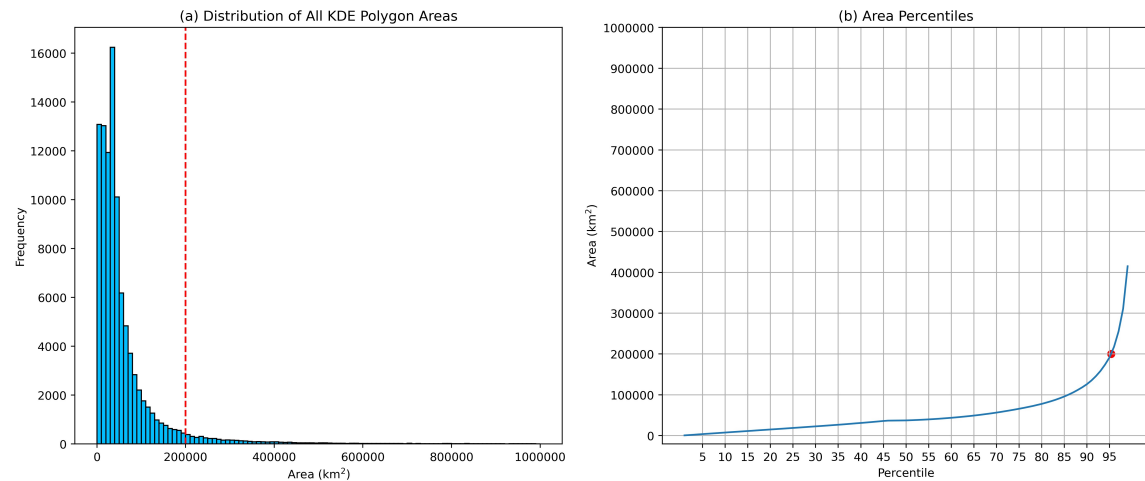


Figure A.3: (a) Histogram of the area of every developed polygon in the database period of record. (b) Polygon area as a function of percentile from (a). Dashed red line in (a) and red dot in (b) correspond to our chosen 200,000 km² threshold.

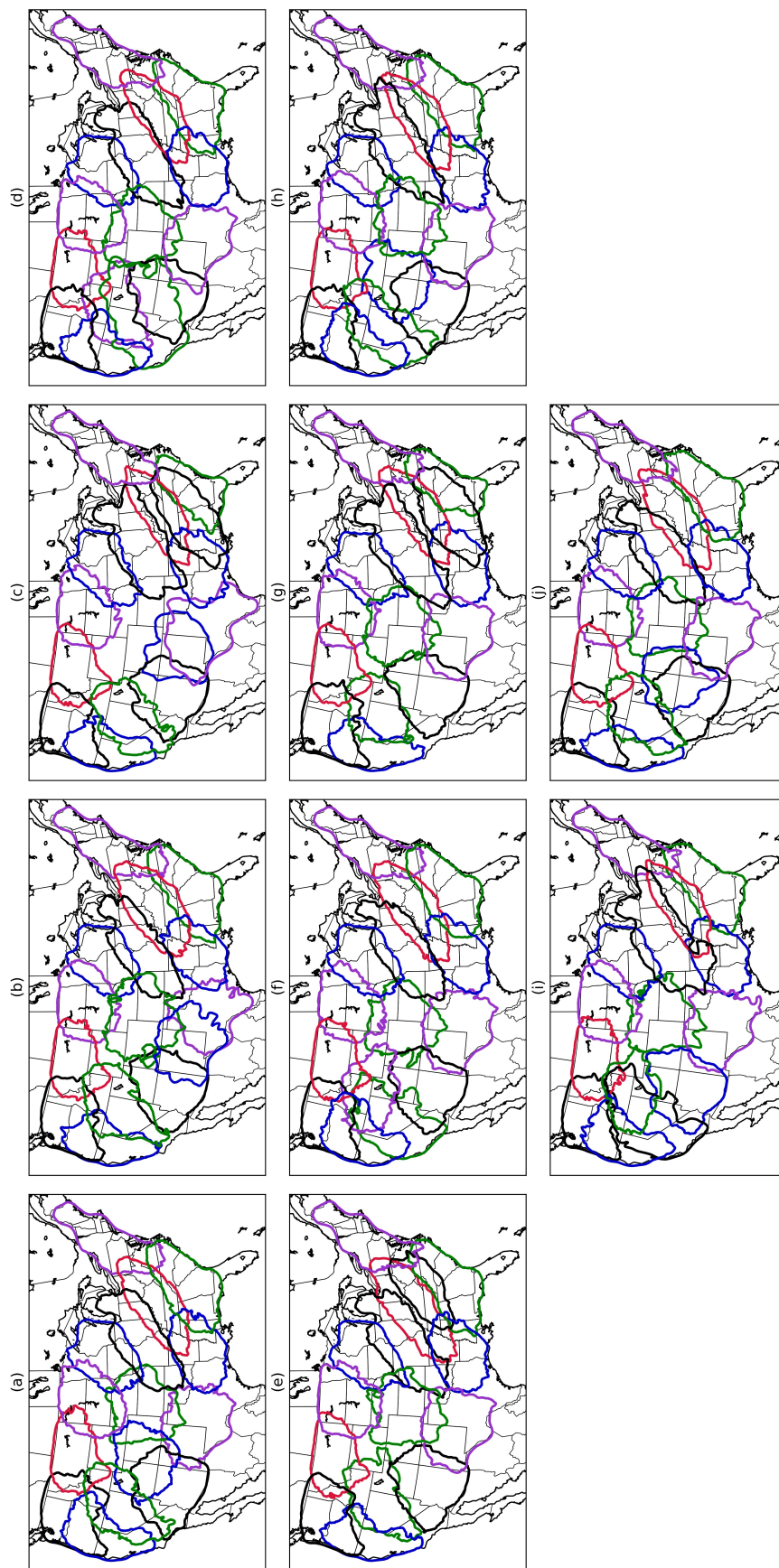


Figure A.4: As in Fig. 2.9, but for different seed initializations. Panel (a) is the seed used to produce Figure 9.

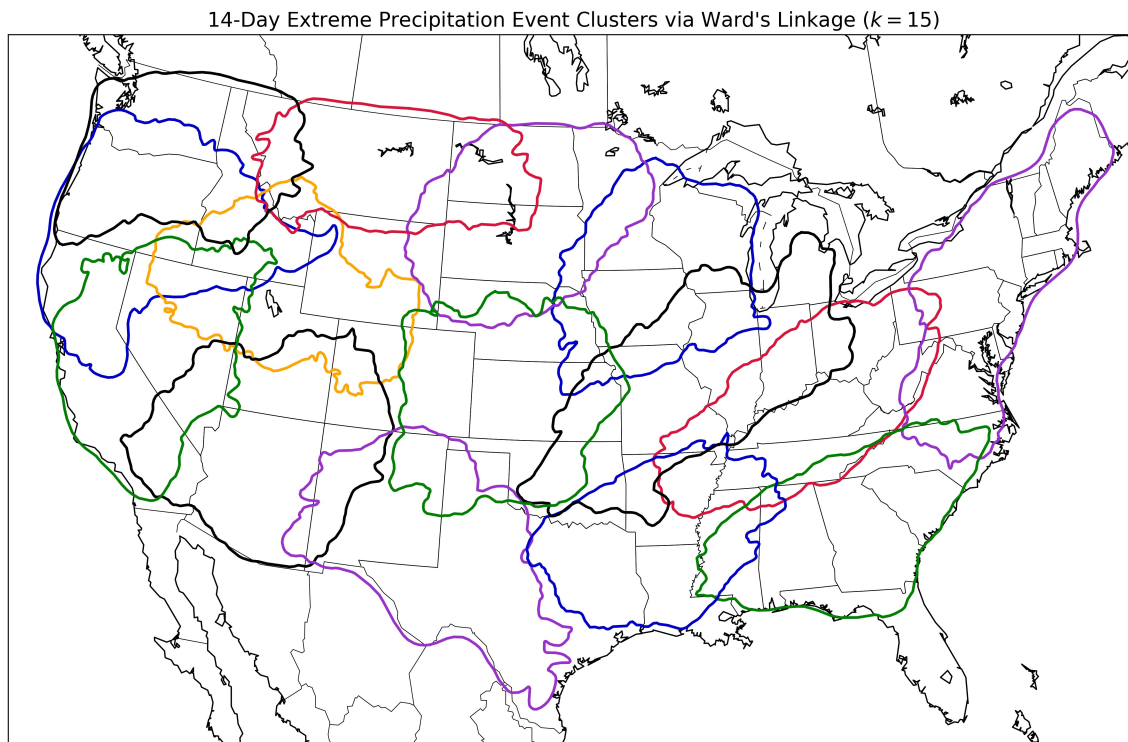


Figure A.5: As in Fig. 2.9, but using Ward's linkage.

2 Appendix B

Date	Cluster
3/25/2012 – 4/2/2012	Cluster 1
9/18/2013 – 9/23/2013	Cluster 1
10/12/2016 – 10/18/2016	Cluster 1
1/8/2017 – 1/10/2017	Cluster 4
3/27/2006 – 4/7/2006	Unclustered
12/17/2010 – 12/27/2010	Unclustered
Total hits = 48	

Table A.1: Standardized Anomalies for Central West Coast.

Date	Cluster
3/28/2012 – 4/2/2012	Cluster 1
9/19/2013 – 9/23/2013	Cluster 1
10/12/2016 – 10/18/2016	Cluster 1
1/7/2017 – 1/12/2017	Cluster 4
3/27/2006 – 4/5/2006	Unclustered
12/17/2010 – 12/29/2010	Unclustered
Total hits = 47	

Table A.2: Detrended Anomalies for Central West Coast.

Date	Cluster
5/31/2005 – 6/5/2005	Cluster 7
5/7/2015 – 5/19/2015	Cluster 7
10/7/2018 – 10/12/2018	Cluster 8
2/26/2004 – 3/2/2004	Cluster 10
7/4/2010 – 7/6/2004	Cluster 10
5/3/2007 – 5/6/2007	Unclustered
12/8/2007 – 12/11/2007	Unclustered
6/11/2010 – 6/14/2010	Unclustered
Total hits = 52	

Table A.3: Standardized Anomalies for Central Plains.

Date	Cluster
6/1/2005	Cluster 7
2/26/2004 – 3/1/2004	Cluster 10
5/4/2007 – 5/5/2007	Cluster 10 and Unclustered
6/11/2010 – 6/14/2010	Unclustered
5/24/2015 – 5/26/2015	Unclustered
Total hits = 14	

Table A.4: Detrended Anomalies for Central Plains.

Date	Cluster
5/5/2002 – 5/12/2002	Cluster 11
1/5/2005 – 1/13/2005	Cluster 11
4/29/2017 – 4/30/2017	Cluster 11
2/23/2001 – 2/24/2001	Cluster 13
2/22/2018 – 2/27/2018	Cluster 13
12/16/2008 – 12/21/2008	Cluster 15
2/26/2011	Cluster 15
10/27/2009 – 10/31/2009	Unclustered
2/16/2018 – 2/27/2018	Unclustered
Total hits = 45	

Table A.5: Standardized Anomalies for Ohio River Valley.

Date	Cluster
5/9/2002 – 5/12/2002	Cluster 11
1/6/2005 – 1/12/2005	Cluster 11
2/22/2018 – 2/27/2018	Cluster 13
10/28/2009	Unclustered
2/16/2018 – 2/27/2018	Unclustered
Total hits = 25	

Table A.6: Detrended Anomalies for Ohio River Valley.