**Title**
Evaluation and Improvement of Hydrological Simulations and Forecasts in the Western U.S.



**Author**
Su, Lu

**Publication Date**
2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

# Evaluation and Improvement of Hydrological Simulations and Forecasts in the Western U.S.

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Geography

by

Lu Su

2023

ABSTRACT OF THE DISSERTATION

Evaluation and Improvement of Hydrological Simulations and Forecasts in the Western U.S.

by

Lu Su

Doctor of Philosophy in Geography

University of California, Los Angeles, 2023

Professor Dennis P. Lettenmaier, Chair

Droughts and floods are among the most catastrophic yet least understood weather and climate threats. Accurate forecasting of droughts and floods is crucial due to their significant financial and human impacts. Concurrently, precise streamflow simulation is critical for effective water management and disaster prevention. The subseasonal drought forecast, vital for water management and disaster mitigation, has been under-studied due to a lack of appropriate meteorological forecast databases until recently. The NOAA's National Water Model (NWM), anchored by its hydrological core Noah Multi-parameterization (Noah-MP), needs a comprehensive comparison with SAC-SMA model-based River Forecast Center (RFC) forecasts to evaluate its flood forecasting efficiency in the Western United States. Accurate daily streamflow predictions are crucial, and a comprehensive, calibrated Land Surface Model (LSM) parameter set is important for reliable streamflow predictions.

This dissertation explores evaluations and enhancements of hydrological simulations and forecasts in the Western U.S., with a focus on three key aspects: a) subseasonal forecast accuracy for drought onset and termination using NOAA's Climate Testbed Subseasonal Experiment (SubX) reforecasts, b) the flood forecasting capabilities of the Noah-MP in comparison to current RFC forecasts, and c) the development of high-resolution calibrated parameters for two notable LSMs, the Variable Infiltration Capacity (VIC) model and Noah-MP. For the first aspect, I employ SubX to drive Noah-MP and produce drought forecasts of different severity and for lead weeks 1-4. I find significant drought termination and onset forecast skill within the initial two weeks and limited skill or no skill at week 4 regardless of drought severity. I find that skill is generally higher for drought termination than for onset for all drought events and that drought prediction skill generally decreases from north to south for all drought events. For the second aspect, I start with selection of appropriate physics options for Noah-MP and calibration of parameters in seven watersheds that form a transect along the U.S. Pacific Coast. I find promising flood prediction capacities of Noah-MP in northern basins but requires refinement for southern basins both in terms of bias and variability. For the third aspect, I develop calibrated and regionalized hydrologic parameters for VIC and Noah-MP at a precision of 1/16° latitude-longitude resolution across 4816 HUC-10 basins in the Western U.S., aiming to enhance the accuracy of hydrological modeling and predictions. In summary, the dissertation provides important contributions to understanding and improving the hydrological simulation and forecast in the Western U.S.

The dissertation of Lu Su is approved.

Kyle C. Cavanaugh

Alton Park Williams

Shraddhanand Shukla

Rong Fu

Yongkang Xue

Dennis P. Lettenmaier, Committee Chair

University of California, Los Angeles

2023

*Dedicated to my family.*

**TABLE OF CONTENTS**

## LIST of FIGURES

# LIST OF TABLES

ACKNOWLEDGEMENTS

I extend my deepest gratitude to Dennis P. Lettenmaier for his consistent guidance, support, and encouragement throughout my academic journey at UCLA. His invaluable insights have been instrumental in my growth, teaching me that academic excellence requires not only passion and curiosity but also unwavering dedication and discipline.

Heartfelt thanks to my committee members - Kyle Cavanaugh, Park Williams, Rong Fu, Shraddhanand Shukla, and Yongkang Xue. Their insightful feedback and collaborative efforts have enriched my academic experience.

A special nod to my lab colleagues at UCLA: Mu Xiao, Zhaoxin Ban, Xiaoyu Ma, Qian Cao, Huilin Huang, Ye Liu, Dongyue Li, Solomon Vimal, Emilie Tarouilly, Kim Wang, and Ruth Engel. Their camaraderie, encouragement, and shared moments of joy have been invaluable. I'm grateful to Kasi McMurray, Jenee Misraje, Brian Won, and the entire departmental staff for their unwavering assistance and support.

My family, especially my parents and sister, deserve immense appreciation. Their understanding and support, especially during the challenging times of the Covid pandemic, have been my anchor. Their faith in my decisions, even when thousands of miles apart, has been heartening.

To my husband, Yuhao Chin, my deepest affection and gratitude. As my steadfast partner, confidant, and teammate, he's been my beacon during challenging times, bringing both joy and motivation into my life.

Finally, I’d like to express my heartfelt gratitude to my friends, Yuehong Wang and Hongyin Wang. Their consistent support and the memories we've shared during both the high and low moments of my life have been invaluable to me.

# Curriculum Vitae

**EDUCATION**

C. Phil. Geography, University of California, Los Angeles 2020

MSc. Global Environmental Change, Beijing Normal University 2018

B.E. Water Conservancy and Hydropower Engineering (major), 2015
B.M. Accounting (minor), Wuhan University

**PROFESSIONAL EXPERIENCES**

Graduate Research Assistant: University of California, Los Angeles 2018-current

**SELECTED PUBLICATIONS**

**Su, L.**, Lettenmaier, D.P., Hartman, R.K., 2023: Improving Noah-MP Flood Forecast Skills over Coastal Western U.S. River Basins, in preparation.

**Su, L.**, Lettenmaier, D.P., Pan, M., Bass, B., 2023: Improving Runoff Simulation in the Western United States with Noah-MP and VIC, EGUsphere, 2023, pp.1-38.

**Su, L.**, Cao, Q., Shukla, S. Pan, M., Lettenmaier, D. P., 2023: Evaluation of Subseasonal Drought Forecast Skill over the Coastal Western U.S., Journal of Hydrometeorology https://doi.org/10.1175/JHM-D-22-0103.1.

**Su, L.**, Cao, Q., Xiao, M., Mocko, D. M., Barlage, M., Li, D., Peters-Lidard, C.D. and Lettenmaier, D. P., 2021: Drought Variability over the Conterminous United States for the Past Century. Journal of Hydrometeorology, 22(5), 1153-1168.

**Su, L.**, Miao, C. and Gou, J., 2021: Long-term Trends in Songhua River Basin Streamflow and its Multivariate Relationships with Meteorological Factors. Environmental Science and Pollution Research, 28, 64206-64219.

Peters-Lidard, C.D., Mocko, D. M., **Su, L.**, Lettenmaier, D. P., Gentine, P. and Barlage, M., 2021: Advances in Land Surface Models and Indicators for Drought Monitoring and Prediction. Bulletin of the American Meteorological Society, 102(5): E1099-E1122.

Pierce, D.W., **Su, L.**, Cayan, D. R., Risser, M. D., Livneh, B. and Lettenmaier, D. P., 2021: An Extreme-Preserving Long-Term Gridded Daily Precipitation Dataset for the Conterminous United States, Journal of Hydrometeorology, 22(7), 1883-1895.

**Su, L.**, Miao, C., Duan, Q., Lei, X. and Li, H., 2019: Multiple-wavelet Coherence of World's Large Rivers with Meteorological Factors and Ocean Signals. Journal of Geophysical Research: Atmospheres, 124(9), 4932-4954.

**Su, L.**, Miao, C., Kong, D., Duan, Q., Lei, X., Hou, Q. and Li, H., 2018: Long-term Trends in Global River Flow and the Causal Relationships between River Flow and Ocean Signals. Journal of hydrology, 563,818-833.

**Su, L.**, Miao, C., Borthwick, A. G., Duan, Q., 2017: Wavelet-based Variability of Yellow River Discharge at 500-, 100-, and 50-year Timescales. Gondwana Research, 49,94-105.

Yang, T., Tao, Y., Li, J., Zhu, Q., **Su, L.**, He, X., Zhang, X., 2017: Multi-criterion Model Ensemble of CMIP5 Surface Air Temperature over China. Theoretical and Applied Climatology, 1-16.

Miao, C., **Su, L.**, Sun, Q., Duan, Q., 2016: A Nonstationary Bias-correction Technique to Remove Bias in GCM Simulations. Journal of Geophysical Research: Atmospheres, 121(10), 5718-5735.

**SELECTED PRESENTATION**

**Su, L.**, Lettenmaier, D. P., Bass, B., 2022: High Resolution VIC Calibration in California Forced by ERA5-WRF Downscaled and Bias-corrected Meteorology, AGU Fall Meeting, Chicago, U.S., Dec 2022

**Su, L.**, Shukla, S., Cao, Q., Lettenmaier, D. P., 2021: Evaluation of Subseasonal Drought Forecast Skill over the Coastal Western U.S., AGU Fall Meeting, New Orleans, U.S., Dec 2021

Peters-Lidard, C. D., Mocko, D. M., **Su, L.**, Lettenmaier, D. P., Gentine, P., Barlage, M., 2021: Advances in Land Surface Models and Indicators for Drought Monitoring and Prediction (Invited), AGU Fall Meeting, New Orleans, U.S., Dec 2021

**Su, L.**, Cao, Q., Xiao, M., Lettenmaier, D. P., Li, D., Barlage, M., Mocko, D. M., 2020: Drought Variability and Trends over the Conterminous United States over the Past Century2, AGU Fall Meeting, Online, U.S. Dec 2020

Peters-aLidard, C. D., Mocko, D. M., **Su, L.**, Lettenmaier, D. P., Gentine, P., Barlage, M., 2020: Recent Advances in Indicators for Drought Monitoring and Prediction, AGU Fall Meeting, Online, U.S., Dec 2020

Pierce, D. P., **Su, L.**, Cayan, D. R., Risser, M. D., Livneh, B., Lettenmaier, D. P., 2020: An Extreme-Preserving Long-Term Gridded Daily Precipitation Data Set for the Conterminous United States, online, AGU Fall Meeting, U.S., Dec 2020

**Su, L.**, Cao, Q., Xiao, M., Lettenmaier, D. P., Li, D., Barlage, M., Mocko, D. M., 2019: Drought Variability and Trends over the Conterminous United States over the Past Century1, AGU Fall Meeting, San Francisco, U.S., Dec 2019

Kaenel, M. von, Alam, S., Vimal, S., **Su, L.**, Margulis, S. A., Lettenmaier, D. P., 2019: The Role of Soil Moisture Memory in Spring Runoff Predictability in Western US River Basins, AGU Fall Meeting, San Francisco, U.S., Dec 2019

**Su, L.**, Miao, C., Borthwick, A. G., Duan, Q., 2017: Wavelet-based Variability of Yellow River Discharge at 500-, 100-, and 50-year Timescales, EGU Meeting, Vienna, Austria, Apr 2017

**HONORS**

| | | |
|---|---|---|
| Conference Travel Stipend | UCLA Department of Geography | 2018 |
| Excellent graduate student | Ministry of Education, Beijing, China | 2018 |
| National Scholarship | Ministry of Education, P. R. China (Top 2%) | 2017 |
| First Academic Scholarship | Beijing Normal University (Top 5%) | 2016 |
| Elite Freshmen Scholarship | Beijing Normal University (Top 5%) | 2015 |
| Honor of Outstanding Student | Wuhan University (Top 5%) | 2013 |
| National Encouragement Scholarship | Ministry of Education, P. R. China (Top 5%) | 2012 &2013 |
| Merit Student | Wuhan University (Top 5%) | 2012 |

## Chapter 1 Introduction

Droughts and floods stand as two of the most devastating and least comprehended weather and climate hazards (Pulwarty and Sivakumar, 2014). Between 1980 and 2020, the U.S. experienced 27 major drought events, leading to an estimated cumulative cost of $291 billion (inflation-adjusted to 2023 values), with an average cost of about $10.8 billion for each occurrence (NOAA 2020). Between 1984 and 2013, flood-related damages had an average economic impact of $10.25 billion annually (in 2023 inflation-adjusted terms) and led to an average of 85 fatalities per year (National Weather Service, 2014). The financial and human repercussions of these events highlight the critical importance of accurate drought and flood predictions. Apart from predicting these severe hydrological events, the role of streamflow simulation in hydrological modeling is paramount. Precise streamflow simulations are essential for guiding water resource strategies, shaping infrastructure development, hydroelectric power plants, preserving ecosystems, and mitigating disasters (Raff et al. 2013; Anghileri et al. 2016; Maidment 2017; Federal Institute of Hydrology 2020).

As climate change amplifies the water cycle, floods and droughts are projected to become more frequent and/or severe in large parts of the world including North America especially with a projected 1.5°C global temperature increase (IPCC6 2023). Accurate predictions of these events, and relative streamflow predictions, are crucial for proactive disaster preparedness.

Current drought forecasting techniques, such as the North American Multi-Model Ensemble (NMME) project (Kirtman et al. 2014), have their limitations—as is evident from their inability to forecast the end of the 2013-2016 California drought (Wanders et al. 2017). Yet, what is unpredictable at the seasonal time scale can become predictable at the subseasonal-to-seasonal (S2S) (two weeks to a month or two) (Wang et al. 2017). Despite its potential, the subseasonal

time scale—critical for proactive water management and disaster mitigation—has been under-researched due to the unavailability of suitable meteorological forecast databases until very recently (Mariotti et al. 2018; Vitart and Robertson 2018; Vitart et al. 2017). NOAA's Climate Testbed Subseasonal Experiment (SubX) (Pegion et al. 2019) project introduced a S2S database that includes both operational and research models, providing real-time data access. However, research on the hydrological implications of this dataset remains scarce. Exploring its potential can pave the way for improved drought management techniques.

In the realm of flood forecasting, the National Water Model (NWM) introduced by NOAA in 2016 (NOAA 2016) is a promising advance in hydrologic prediction capabilities. Its core, the Noah Multi-parameterization (Noah-MP) (Niu et al. 2011), has yet to be compared with the operational forecasts (e.g., those produced by the California Nevada River Forecast Center (CNRFC) and the Northwest River Forecast Center (NWRFC)) for the U.S. West Coast. These forecasts are based on the SAC-SMA model (Burnash et al. 1973). A comparative analysis will help evaluate the NWM's operational viability in flood forecasting. Previous research on Noah-MP's flood forecasts is sparse, with most studies focusing on isolated events. A holistic examination, accounting for multiple floods, parameterization variations, and the benefits of automatic calibration, is sorely needed. This would determine how Noah-MP measures up to existing SAC-based forecasts in the Western U.S.

Beyond these extremes, daily streamflow forecasts remain foundational to our lives. Given gaps in observed streamflow observations (even over the relatively well-observed conterminous U.S.), Land Surface Models (LSMs) are indispensable for simulating runoff and streamflow. Their capabilities are paramount for adept water management and understanding climate trajectories. However, accurate modeling requires meticulous calibration—a computationally intense process.

Historically, research on hydrologic model calibration has been focused narrowly, as evidenced by studies like Mascaro et al (2023) and Gou et al (2020). A regionally comprehensive, calibrated LSM parameter set is imperative for reliable streamflow predictions.

In light of the above background, my dissertation delves into three pivotal areas concerning the evaluation and enhancement of hydrological predictions and models in the Western U.S. Specifically:

(1) I investigate, in Chapter 2, the subseasonal forecast accuracy for drought onset and termination using SubX reforecasts.

(2) I assess, in Chapter 3, the flood forecasting capabilities of the Noah-MP and compare it with current RFC forecasts.

(3) In chapter 4, I develop high-resolution calibrated parameters for two widely used LSMs: the Variable Infiltration Capacity (VIC) model and Noah-MP.

I address these areas in the following three core chapters of this dissertation (Chapters 2-4).

Chapter 2 focuses on the subseasonal forecast skill for drought onset and termination in the coastal Western U.S. at lead times of 1-4 weeks. Initially, I enhance the spatial resolution of the SubX reforecasts from their native 1 degree to a finer 1/16 degree, aligning with the hydrological model's high spatial granularity. Using these downscaled and bias-corrected SubX reforecasts as forcings, I run the Noah-MP model over the coastal Western U.S.. Based on the model output soil moisture, I assess the proficiency of SubX-based drought forecasts, taking into account of geographical variations and lead times.

Chapter 3 evaluates the performance of Noah-MP (NWM) for flood forecasting. I identify the most suitable physical parameterizations for the model and calibrate it across seven river basins spanning the coastal Western U.S. By juxtaposing the Noah-MP flood reforecasts with archived

operational forecasts from CNRFC and NWRFC, I provide a comprehensive evaluation of Noah-MP's potential in enhancing forecast accuracy compared to existing NWS/RFC techniques.

In Chapter 4, I develop and implement a method for calibrating the parameters of two widely used hydrological models (Noah-MP and VIC) across the Western U.S.. I provide a detailed account of the calibration method employed across 263 monitored basins in the Western U.S.. I further assess the models' efficacy in these basins, exploring factors that might affect simulation performance. Additionally, I extend the calibrated parameters to ungauged basins, examining the effectiveness of the donor-basin regionalization technique in this context.

To summarize, this dissertation sheds light on the potential of subseasonal drought and flood forecast capabilities across the Western U.S., focusing mostly on the NWM. Furthermore, it offers calibrated parameter sets for two prominent hydrological models, promising advancements in hydrological forecasting and modeling endeavors.

**References**

Anghileri, D., N. Voisin, A. Castelletti, F. Pianosi, B. Nijssen, and D.P. Lettenmaier. 2016: Value of Long-Term Streamflow Forecasts to Reservoir Operations for Water Supply in Snow-Dominated River Catchments. Water Resources Research 52: 4209–25.

Burnash, R., and R. Ferral, 1973: A Generalized Streamflow Simulation System. U.S. Department of Commerce, National Weather Service, and State of California.

Cook, B. I., A. P. Williams, J. S. Mankin, R. Seager, J. E. Smerdon, and D. Singh, 2018: Revisiting the leading drivers of Pacific coastal drought variability in the contiguous United States. Journal of Climate, 31(1), 25-43.

Federal Institute of Hydrology. 2020: "SOSRHINE." http://sosrhine.euporias.eu/en/sosrhine_overview.

Gou, J., C. Miao, Q. Duan, Q. Tang, Z. Di, W. Liao, J. Wu, and R. Zhou, 2020: Sensitivity analysis-based automatic parameter calibration of the VIC model for streamflow simulations over China. Water Resources Research, 56(1), e2019WR025968.

IPCC, 2023: Climate Change 2023: Synthesis Report. A Report of the Intergovernmental Panel on Climate Change. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. H. Lee and J. Romero (eds.) IPCC, Geneva, Switzerland, (in press)

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, 585–601, https://doi.org/10.1175/BAMS-D-12-00050.1.

Maidment, D.R. 2017: Conceptual Framework for the National Flood Interoperability Experiment. Journal of the American Water Resources Association 53: 245–57.

Mariotti, A., P. M. Ruti, and M. Rixen, 2018: Progress in subseasonal to seasonal prediction through a joint weather and climate community effort. npj Climate Atmos. Sci., 1, 4, https://doi.org/10.1038/s41612-018-0014-z.

Mascaro, G., A. Hussein, A. Dugger, and D. J. Gochis, 2023: Process-based calibration of WRF-Hydro in a mountainous basin in southwestern US. JAWRA Journal of the American Water Resources Association, 59(1), 49-70.

National Weather Service, 2014: United States Flood Loss Report - Water Year 2014. https://www.nws.noaa.gov/os/ water/Flood%20Loss%20Reports/WY14%20Flood%20Loss%20Summary.pdf.

Niu, G. Y., Z. L. Yang, K. E. Mitchell, F. Chen, M. B. Ek, M. Barlage, and M. Tewari, 2011: The community Noah land surface model with multiparameterization options (Noah MP): 1. Model description and evaluation with local scale measurements. Journal of Geophysical Research: Atmospheres, 116.

NOAA. 2016: "National Water Model." Improving NOAA's Water Prediction Service. https://water.noaa.gov/docum ents/wrn-national-water-model.pdf.

NOAA, 2020: U.S. Billion-Dollar Weather and Climate Disasters. NOAA/NCEI, https://www.ncdc.noaa.gov/billions/.

Pegion, K., and Coauthors, 2019: The Subseasonal Experiment (SubX): A multimodel subseasonal prediction experiment. Bulletin of the American Meteorological Society, 100(10), pp.2043-2060.

Pulwarty, R.S. and M.V. Sivakumar, 2014: Information systems in a changing climate: Early warnings and drought risk management. Weather Clim. Extrem., High Level Meeting on National Drought Policy 3, 14–21.

Raff, D., L. Brekke, K. Werner, A. Wood, and K. White. 2013: Short-Term Water Management Decisions: User Needs for Improved Climate, Weather, and Hydrologic Information. U.S. Bureau of Reclamation. https://www.usbr.gov/research/st/roadmaps/WaterSupply.pdf.

Vitart, F., and A. W. Robertson, 2018: The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. npj Climate Atmos. Sci., 1, 3, https://doi.org/10.1038/s41612-018-0013-0.

Vitart, F., and Coauthors, 2017: The subseasonal to seasonal (S2S) prediction project database. Bulletin of the American Meteorological Society, 98(1), 163-173.

Wanders, N., and Coauthors, 2017: Forecasting the hydroclimatic signature of the 2015/16 El Niño event on the western United States. J. Hydrometeor., 18, 177–186, https://doi.org/10.1175/JHM-D-16-0230.1.

Wang, S., A. Anichowski, M. K. Tippett, and A. H. Sobel, 2017: Seasonal noise versus subseasonal signal: Forecasts of California precipitation during the unusual winters of 2015–2016 and 2016–2017. Geophys. Res. Lett., 44, 9513–9520, https://doi.org/10.1002/2017GL075052.

## Chapter 2 Evaluation of Subseasonal Drought Forecast Skill over the Coastal Western U.S.

This chapter has been published in its current form in the Journal of Hydrometeorology.

 The supplemental material for this chapter is provided in Appendix A.

Su, L., Q. Cao, S. Shukla, M. Pan, and D. P. Lettenmaier, 2023: Evaluation of Subseasonal Drought Forecast Skill over the Coastal Western United States. Journal of Hydrometeorology, 24(4), 709-726. https://doi.org/10.1175/JHM-D-22-0103.1

**Abstract**

Predictions of drought onset and termination at subseasonal (from two weeks to one month) lead times could provide a foundation for more effective and proactive drought management. We used reforecasts archived in NOAA's Subseasonal Experiment (SubX) to force the Noah Multi-parameterization (Noah-MP), which produced forecasts of soil moisture from which we identified drought levels D0-D4. We evaluated forecast skill of major and more modest droughts, with leads from one to four weeks, and with particular attention to drought termination and onset. We find usable drought termination and onset forecast skill at leads one and two weeks for major D0-D2 droughts; and limited skill at week three for major D0-D1 droughts, with essentially no skill at week four regardless of drought severity. Furthermore, for both major and more modest droughts, we find limited skill or no skill for D3-D4 droughts. We find that skill is generally higher for drought termination than for onset for all drought events. We also find that drought prediction skill generally decreases from north to south for all drought events.

### 2.1 Introduction

Drought is among the most damaging, and least understood, of all weather and climate hazards (Pulwarty and Sivakumar 2014). Droughts are usually incremental and can span from a

few weeks to decades temporally and from a few hundred $km^2$ to hundreds of thousands of $km^2$ spatially (Pendergrass et al. 2020). Droughts' creeping development is often neglected in the early stages and the changes accumulate and trigger more severe direct or indirect impacts. Eventually, the unattended creeping development leads to urgent crises that are more costly to deal with (Glantz, 2004). The impacts can persist even after the drought itself ends. Therefore, drought is often a 'hidden' natural disaster and its risk is underestimated (UNDRR 2019; Pendergrass et al. 2020).

During the past decade, nearly all of the contiguous United States (CONUS) from Colorado to the Pacific coast has suffered from moderate to exceptional droughts (Cook et al. 2018). This includes the continuation of multiyear events (2009-2011 and 2013-2016) in California (Griffin and Anchukaitis 2014; Seager et al. 2015; Williams et al. 2015) and the U.S. Southwest (Delworth et al. 2015; Seager and Hoerling 2014), and the emergence of significant drought conditions across the Pacific Northwest (Oregon and Washington) in 2015 (Mote et al. 2016). Drought episodes were especially severe in the coastal Western U.S. (including California, Oregon, and Washington). The prolonged severe droughts have stressed water resources management at the regional level (Mann and Gleick 2015; Engström et al 2020).

As the climate warms, an argument has evolved as to whether drought duration and intensity are increasing (Christensen et al. 2007; Seneviratne et al. 2012; Pendergrass et al. 2020). If so, more foresighted responses that adopt proactive risk mitigation strategies may be necessary (Pulwarty and Verdin 2013; Wilhite et al. 2014). Drought forecast systems in this context would be especially useful (Arsenault et al. 2020; Carrão et al. 2018; Hao et al. 2018). Predictions of drought onset and termination (although evasive to date) in addition to other drought characteristics could provide a foundation for effective proactive drought management.

Seasonal climate forecast systems including the North American Multi-Model Ensemble (NMME) project (Kirtman et al. 2014; Wanders et al. 2017) consistently predicted a false wet 2015/2016 winter and forecasted a false signal for California drought termination. In contrast, the forecasts and reforecasts from the ECMWF and NCEP CFSv2 models, at the subseasonal-to-seasonal (S2S) (weeks to a month or two) time scale, were able to predict the correct sign of precipitation anomalies (Wang et al. 2017). Wang et al. (2017) shows that what is unpredictable at the seasonal time scale can become predictable at the subseasonal time scale. Recently there has been surging interest in 'flash droughts', which are characterized by their sudden onset and rapid intensification and severe impacts (Otkin et al. 2018). While many drought prediction products are updated at monthly time scales, these predictions are of limited value for flash droughts which develop on shorter time scales (Pendergrass et al. 2020), nor are they useful in determining, for instance, whether individual storms (which can be forecast with potentially usable accuracy at lead times of one to several weeks) will terminate a drought. This further motivates the need for incorporation of S2S forecasts into drought monitoring and prediction systems. Our study aims to fill a gap in the literature on drought forecast skill to incorporate subseasonal forecasts. Like seasonal drought prediction systems, such as the NOAA Climate Prediction Center's (CPC) seasonal drought outlook, subseasonal drought forecasts derive their skill from knowledge of weather/climate information and initial hydrologic conditions (IHCs) at the onset of the forecast period (Shukla et al. 2012). While subseasonal precipitation forecast skill is generally lower than the skill of forecasts for temperature for the same location and lead time (Monhart et al. 2018; Pegion et al. 2019; Cao et al. 2021), these studies show that there nonetheless is potentially usable precipitation forecast skill to leads of 2-3 weeks. Furthermore, Land Surface Models (LSMs) provide estimates of IHCs that are critical for drought forecasts, particularly when (as in the case

of agricultural drought) soil moisture is the metric used to identify droughts (Shukla and Lettenmaier 2011; Shukla et al. 2012). In this respect, the work we report here extends this earlier work to utilize S2S forecasts which better exploit precipitation (and hence, soil moisture) forecast skill at lead times of one to several weeks.

The subseasonal forecasting time scale (the terms subseasonal and Subseasonal-to-Seasonal (S2S) are used interchangeably here) is typically defined by lead times ranging from two weeks to one (or two) months. This is a critical lead time window for proactive disaster mitigation efforts such as water resource management for drought mitigation (Mariotti et al. 2018; Vitart and Robertson 2018). However, research on hydrological application of forecasts has not paid much attention to subseasonal lead times until very recently due to a lack of subseasonal meteorological forecast databases (Vitart et al. 2017). Multimodel ensemble approaches have proved to be a successful tool for improving forecast quality for weather and seasonal predictions (Krishnamurti et al. 1999; Krishnamurti et al. 2000). They have the advantage of exploiting complementary skill from different models and allow for better estimation of forecast uncertainty (Hao et al. 2018).

Thanks to joint efforts between the weather and climate communities, several subseasonal forecast databases have been developed to bridge the weather-climate prediction gap in the S2S range (Mariotti et al. 2018; Merryfield et al. 2020). These include the World Weather Research Programme (WWRP)/World Climate Research Program (WCRP) S2S Prediction Project (Vitart et al. 2017) and the NOAA/Climate Testbed Subseasonal Experiment (SubX) project (Pegion et al. 2019). Recent studies have found that the prediction skill for precipitation and the application to streamflow forecasts of the WWRP/WCRP S2S database varied among predictor combinations, catchments and dates of prediction; and the skill is frequently less than climatology beyond two weeks lead time (Lin et al. 2018; Pan et al. 2019; Schick et al. 2019).

NOAA's SubX project is different from the WWRP/WCRP reforecasts by including both operational and research models. Furthermore, it is available in near real-time (Pegion et al. 2019). To our knowledge, little research has been done to evaluate the hydrological application of subseasonal forecasts based on the newly developed SubX dataset. A thorough investigation of the hydrological usefulness of subseasonal drought forecasts based on the SubX dataset could form the foundation of a proactive drought management system.

SubX provides forecasts of climate variables like precipitation and temperature, but not all of them provide hydrologic variables like soil moisture and runoff. However, hydrologic forecasts based on SubX can be produced by using the SubX precipitation (and other surface variables) forecasts to drive a land surface model (see e.g., Cao et al. 2021). Here, we drive hydrological forecasts from SubX with the Noah Multi-parameterization (Noah-MP, V4.0.1) (Niu et al. 2011). We adopted the WRF-HYDRO recommended physical options and details are in Appendix A Text S1. Noah-MP is a state-of-the-art LSM originally intended to be the land surface scheme in numerical weather prediction (NWP) models. It is currently used for physically based, spatially distributed hydrologic simulations within the construct of NOAA's National Water Model (NWM). Noah-MP extends the capabilities of the Noah LSM (Chen et al. 1996; Chen and Dudhia 2001) and incorporates multiple options for key land-atmosphere interaction processes, such as surface water infiltration, runoff, groundwater transfer, and channel routing (Niu et al. 2007; Niu et al. 2011). Noah-MP has been widely used for predicting seasonal climate, weather, droughts and floods within and beyond CONUS (Zheng et al. 2019).

Given this background, our objectives here are to examine: 1) subseasonal forecast skill (at 1–4-week lead times) of drought onset and termination driven by downscaled SubX reforecasts in the coastal Western U.S.; 2) how forecast skill for drought onset and termination vary

geographically and with lead times. To achieve these objectives, we first downscaled the SubX reforecasts to a finer spatial resolution (1/16 degree) from their coarse native resolution (1 degree), in consideration of the high spatial resolution of our hydrological model. We then implemented the Noah-MP hydrology model over the coastal Western US using downscaled and bias-corrected SubX reforecasts as forcings. Based on the model outputs, we evaluated the SubX-based drought forecasts skill (All of the ‘‘forecasts’’ in this paper technically are reforecasts).

## 2.2 Study Domain and Dataset

### 2.2.1 Study Domain

Our study domain is the coastal Western U.S., consisting of all of California (CA), as well as coastal Oregon (OR) and Washington (WA) (Figure 2.1).

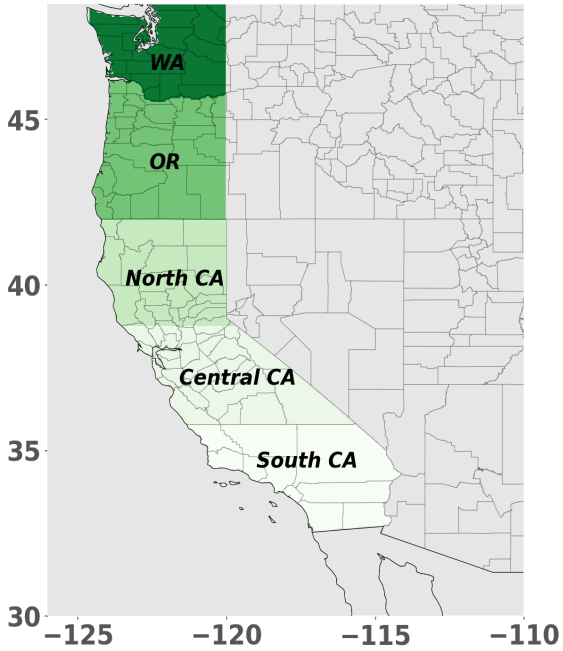



Figure 2.1 Study domain: the coastal Western U.S.

### 2.2.2 SubX Database

We used six models from the SubX database with 30 ensemble members in total (Table 2.1) over the reforecast period Jan 1999-Dec 2016. The initialization interval of each model is at least

once a week and the lead time is at least 32 days. The temporal resolution of the SubX output is daily and the raw spatial resolution is 1°× 1°. We downscaled and bias corrected the SubX output to 1/16°× 1/16° as described in section 2.3.1.

Table 2.1 List of SubX models used in the research. Community column indicates target users for each model (SEAS for seasonal prediction community and NWP for numerical weather prediction community).

| Model | Members | Initialization Day | Forecast Length (days) | Community | Reference(s) |
|---|---|---|---|---|---|
| NCEP-CFSv2 | 4 | W | 45 | SEAS | Saha et al. (2014) |
| GMAO-GEOS_V2p1 | 4 | Varies | 45 | SEAS | Koster et al. (2000), Molod et al. (2012), Reichle and Liu (2014), and Rienecker et al. (2008) |
| RSMAS-CCSM4 | 3 | Su | 45 | SEAS | Infanti and Kirtman (2016) |
| EMC-GEFS | 11 | W | 35 | NWP | Zhou et al. (2016, 2017) and Zhu et al. (2018) |
| ECCC-GEPS6 | 4 | Th | 32 | NWP | Lin et al. (2016) |
| ESRL-FIMr1p1 | 4 | W | 32 | NWP | Sun et al. (2018a, b) |

## 2.3 Methods

### 2.3.1 Downscaling and Bias Correction

We downscaled the raw SubX output (forcings to Noah-MP) using a statistical downscaling method, bias correction and spatial downscaling (BCSD; Wood et al. 2004). We applied daily BCSD since it has been shown to be an effective approach for removing bias (e.g., Monhart et al. 2018; Baker et al. 2019, Cao et al. 2021) in atmospheric model output. By using this method, we constrained the precipitation temporal variability (wet/dry days) to be the same as in the raw data, which we view as desirable (in contrast to methods like localized constructed analogs (LOCA;

Pierce et al. 2014) that attempt to reproduce realistic wet/dry sequences). We applied daily BCSD to precipitation, maximum daily temperature (Tmax), minimum daily temperature (Tmin), and wind speed following the steps in Cao et al (2021), which can be summarized as follows: (1) we applied spatial (bilinear) interpolation of the 1°× 1° daily SubX forecasts to 1/16°× 1/16°; (2) we bias corrected the outputs from step (1) by each grid point using the daily empirical quantile mapping (QM) method (Wood et al. 2002; Cao et al. 2021). The training dataset we used here is the gridded observation dataset of Livneh et al. (2013) (extended to 2018 as described in Su et al. 2021).

### 2.3.2 Evaluation of SubX Precipitation and Temperature

We evaluated SubX forecast skill for precipitation and temperature at different lead times before and after bias correction with BCSD. The skill of forecasts at S2S time scales is typically evaluated in terms of anomalies or differences from the climatology. Following Pegion et al. (2019) and Cao et al. (2021), we used the anomaly correlation coefficient (ACC; Wilks 2006). ACC provides information about how well the variability of the forecasted anomalies matches the observed variability. It is calculated as the temporal correlation of anomalies at each grid cell (details of the ACC calculation procedures are as in Cao et al. 2021). To evaluate the performance of downscaling methods, we also compared the relative biases for both precipitation and temperature before and after the implementation of BCSD.

### 2.3.3 Hydrological Model implementation

We implemented Noah-MP over the coastal Western U.S., which consists of all of CA, as well as coastal OR and WA. Noah-MP requires meteorological forcings including specific humidity, surface pressure, downward solar and longwave radiation in addition to precipitation, wind speed, air temperature. We calculated the first four variables based on the Mountain

Microclimate Simulation Model (MTCLIM) algorithms (implemented as in Bohn et al. 2013; Cao et al. 2021; and Su et al. 2021) and disaggregated the daily output to 3-hourly (Liang et al. 1994; Bennett et al. 2020).

The prediction skill of subseasonal hydrological forecasts depends on both the IHCs at the time of forecast and the accuracy of forecasts of hydrologic model forcings during the forecast period (Arnal et al. 2017; Li et al. 2009). Before we implemented Noah-MP using SubX forcings, we first ran the model using the Livneh et al. (2013) forcings for the period 1951-2016 and repeated twice. We cropped out the 1961-2016 period from the second repetition to serve as a baseline run and also to provide assumed perfect IHCs at forecast initiation time for forecasts made over the period 1999-2016. The initialization interval for most SubX models is seven days, but different models have different initiation days. We output baseline run model states for all the SubX initiation dates and these states served as the IHCs. For each SubX ensemble member and each identified initialization, we ran Noah-MP for 28 days (4-week forecast).

To assess the hydrological model dependency and the effects of calibration, we also implemented Variable Infiltration Capacity (VIC) V4.1.2.d (Liang et al. 1994) before and after calibration (details in Appendix A Text S2). Overall, those results show that, while there are some differences between models (Noah-MP and VIC) and VIC before and after calibration, our results are not strongly dependent on model and calibration. This is consistent with Mo et al. (2012) who found that differences in soil moisture percentiles during drought periods are modest among different LSMs.

### 2.3.4 Assessment of Drought Forecast Skill

#### 2.3.4.1 Identification of Drought Events

Soil moisture is an important drought indicator, especially for agricultural droughts. We archived the total column soil moisture and calculated the soil moisture percentile (relative to that grid cell's and that week's total column soil moisture history of all the ensembles of the model) to identify drought events equivalent to D0 to D4 droughts as used by the U.S. Drought Monitor (https://droughtmonitor.unl.edu/About/WhatistheUSDM.aspx) (see also Table 2.2)

Table 2.2 Drought categories, descriptions and percentiles.

| Category | **D0** | **D1** | **D2** | **D3** | **D4** |
|---|---|---|---|---|---|
| Description | Abnormally Dry | Moderate Drought | Severe Drought | Extreme Drought | Exceptional Drought |
| Percentiles | <30 | <20 | <10 | <5 | <3 |

#### 2.3.4.2 Evaluation Skill

We evaluated the probabilistic drought forecast skill of all six SubX models using 30 ensemble members. The evaluation metrics we used include 1) debiased Brier skill score (Weigel et al. 2007); 2) Bias score (BS), Probability of detection (POD), False alarm ratio (FAR), Equitable threat score (ETS) and Heidke skill score (HSS). We discuss these skill measures and our applications briefly below.

i. BSS

The Brier Skill Score (Wilks, 2006) is widely used to measure the mean squared error of probability forecasts for binary events. It is, however, sensitive to small ensemble sizes. To overcome this issue, we used the debiased Brier skill score (BSS) which incorporates a correction term in the denominator of the Brier Score (DeFlorio et al. 2019). BSS is calculated as follows:

$$BSS = 1 - \frac{BS}{BS_{ref+D}}, \quad (1)$$

$$BS = \frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2, \quad (2)$$

$$BS_{ref} = \frac{1}{N}\sum_{i=1}^{N}(P_{clim} - O_i)^2, \quad (3)$$

$$D = \frac{1}{M}P_{clim}(1 - P_{clim}), \quad (4)$$

where $P_i$ is the forecast skill for drought onset/termination and is determined by the fraction of the ensemble members that predicted drought onset/termination for a single reforecast; $O_i$ shows whether the observed drought onset/termination occurs (1 if yes, 0 if no); N is the number of reforecast droughts for the grid cell/region (varies for each grid cell/region); M is the ensemble size (30 here); and Pclim is the probability of the reference climatology. BSS ranges from negative infinity to one. Positive values indicate that the reforecast skill is higher than the climatological forecast skill.

ii. Contingency Table

We evaluated the forecast of drought onset/termination, where a dichotomous forecast indicates whether an event will happen or not. To verify this type of forecast we start with a contingency table that shows the frequency of "yes" and "no" forecasts and occurrences. The four combinations of forecasts (yes or no) and observations (yes or no), are:

Table 2.3 Contingency Table

| | ***Observed*** | | |
|---|---|---|---|
| ***Forecast*** | ***Yes*** | ***No*** | ***Total*** |
| ***Yes*** | Hits | False alarms | Forecast Yes |
| ***No*** | Misses | Correct Negatives | Forecast No |
| | Observed Yes | Observed No | |

1. hit - event forecast to occur, and it did occur

2. miss - event forecast not to occur, but did occur

3. false alarm - event forecast to occur, but did not occur

4. correct negative - event forecast not to occur, and did not occur

We calculated a variety of categorical statistics from the elements in the contingency table to describe particular aspects of forecast performance.

iii. Bias Score (BIAS)

$$BIAS = \frac{hits+false\ alarms}{hits+misses} \tag{5}$$

Bias score indicates how the forecasted frequency of "yes" events compared to the observed frequency of "yes" events. It ranges from 0 to ∞ with 1 a perfect score. It indicates whether the forecast system tends to underforecast ($BIAS < 1$) or overforecast ($BIAS > 1$) events. It only measures relative frequencies and does not measure how well the forecast corresponds to the observations,

iv. Probability of detection (POD, also known as hit rate)

$$POD = \frac{hits}{hits+misses} \tag{6}$$

Probability of detection tells us what fraction of the observed "yes" events were correctly forecasted. It ranges from 0 to 1 with 1 a perfect score. POD is sensitive to the climatological frequency of the event and is most informative for rare events.

v. False alarm ratio (FAR)

$$FAR = \frac{false\ alarms}{hits+false\ alarms} \tag{7}$$

FAR gives the fraction of predicted "yes" events that actually did not occur (i.e., were false alarms). It ranges from 0 to 1 with 0 a perfect score. FAR is sensitive to false alarms but ignores misses and should be used in conjunction with POD (above).

vi. Equitable threat score (ETS, also known as Gilbert skill score)

$$ETS = \frac{hits - hits_{random}}{hits + misses + false\ alarms - hits_{random}} \tag{8}$$

where $$hits_{random} = \frac{(hits + misses)(hits + false\ alarms)}{total} \tag{9}$$

ETS measures the fraction of observed events that were correctly predicted, adjusted for hits associated with random chance (for example, it is easier to correctly forecast precipitation occurrence in a wet climate than in a dry climate). It ranges from -1/3 to 1; 0 indicates no skill and 1 is a perfect score. *ETS* is often used in the verification of precipitation in NWP models because its "equitability" allows scores to be compared more fairly across different regimes.

vii. Heidke skill score (HSS, also known as Cohen's k)

$$HSS = \frac{(hits + correct\ negative) - (expected\ correct)_{random}}{N - (expected\ correct)_{random}} \tag{10}$$

where $$(expected\ correct)_{random} = \frac{1}{N}[A + B]$$

$$A = (hits + misses)(hits + false\ alarms)$$

$$B = (correct\ negatives + misses)(correct\ negatives + false\ alarms)$$

$$N = hits + misses + false\ alarms + correct\ negatives$$

HSS measures the fraction of correct forecasts after eliminating those forecasts which could be correct due purely to random chance. It ranges from -1 to 1; 0 indicates no skill and 1 is a perfect score. HSS is used in NOAA's climate prediction center (https://www.cpc.ncep.noaa.gov/products/predictions/90day/skill_exp.html).

## 2.4 Results

### a. Evaluation of SubX reforecasts

#### 1) Precipitation and temperature skill

We examined the precipitation and temperature skill of the individual SubX models (raw data, 1° resolution), as well as the multimodel ensemble mean (denoted as "Multimodel"), at lead times of 1-4 weeks averaged over the coastal Western U.S. for each month during the Oct-Mar period separately (see Figure 2.2). We chose to focus our evaluation on the cool season months Oct-Mar as precipitation is generally much lower over most of our domain in the warm season. Figure 2.2a shows that precipitation skill (as measured by ACC) drops rapidly by approximately 40% after week 1. Almost all models have positive ACC in all months; but by week 3, some models show almost zero ACC in certain months. Among individual models, NCEP-CFSv2 performs best in weeks 1-2, with skill similar to Multimodel. However, the model performance at longer lead times varies by months.

Figure 2.2b shows the forecast skill for temperature (the pattern for Tmin is similar, so we only show Tmax here). The temperature of SubX models individually as well as their multimodel mean shows statistically significant (different from zero) skill for all lead times in most conditions. Similar to precipitation, Tmax skill drops quickly after week 1. Tmax shows higher skill than precipitation for all leads and shows fewer negative ACC values in weeks 3–4. Overall, multimodel shows consistently statistically significant ACC across all lead times for both precipitation and temperature. The precipitation and temperature skill we found is consistent with previous studies of SubX (Cao et al. 2021; DeAngelis et al. 2020).

**2) Performance of Daily BCSD**

The difference in precipitation and temperature skill (as measured by ACC) before and after applying daily BCSD is small. This meets our expectation since the QM is performed in a lead time-dependent manner. Figures 2.3-2.5 show the average relative bias ((model - observation)/observation %) for precipitation forecasts and bias (model - observation) for temperature forecasts before and after applying daily BCSD, averaged over October–March. Before applying daily BCSD, the absolute relative biases of precipitation were up to 80% across models and over weeks 1–4. They were reduced to below 6% after applying BCSD. The biases in temperature were also reduced from up to 3.5℃ to below 0.5℃ after applying BCSD (Figure 2.3). The bias maps before and after BCSD also show that the biases were essentially removed after applying BCSD (Figures. 2.4 and 2.5).

**b. Hydrologic model evaluation**

We examined model performance of the baseline run, forced by the Livneh et al. (2013) data with hourly disaggregation. We evaluated California drought area history for various drought levels (D0-D4 drought based on USDM) compared with the USDM. The drought area time series in baseline run and USDM are highly consistent with correlation coefficients ranging from about 0.8 for D0 to 0.6 for D4 (Figure 2.6). We further compared the drought area time series for different drought levels in five subregions (coastal Washington, coastal Oregon, northern California, central California and southern California from north to south, see Figure 2.7). We found that drought duration becomes longer, and drought spatial coverage becomes larger from north to south. There are more small drought events in the north while the droughts in the south are more prolonged.

It is important to note that our results are from the Noah-MP model with the Livneh forcing as the truth. Use of observed soil moisture was not feasible because soil moisture observations are

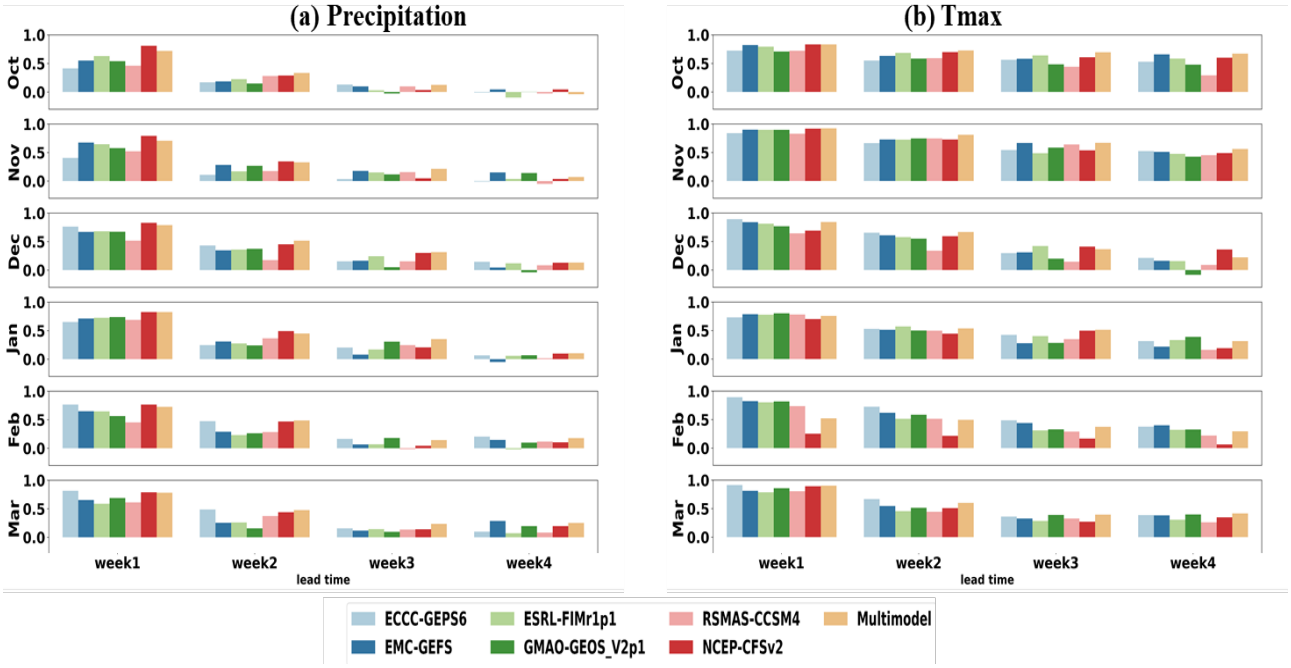



Figure 2.2 Precipitation (a) and Tmax (b) prediction skill (as measured by the anomaly correlation coefficient (ACC)) of SubX models averaged over the coastal Western US for leads1-4 weeks without bias correction).

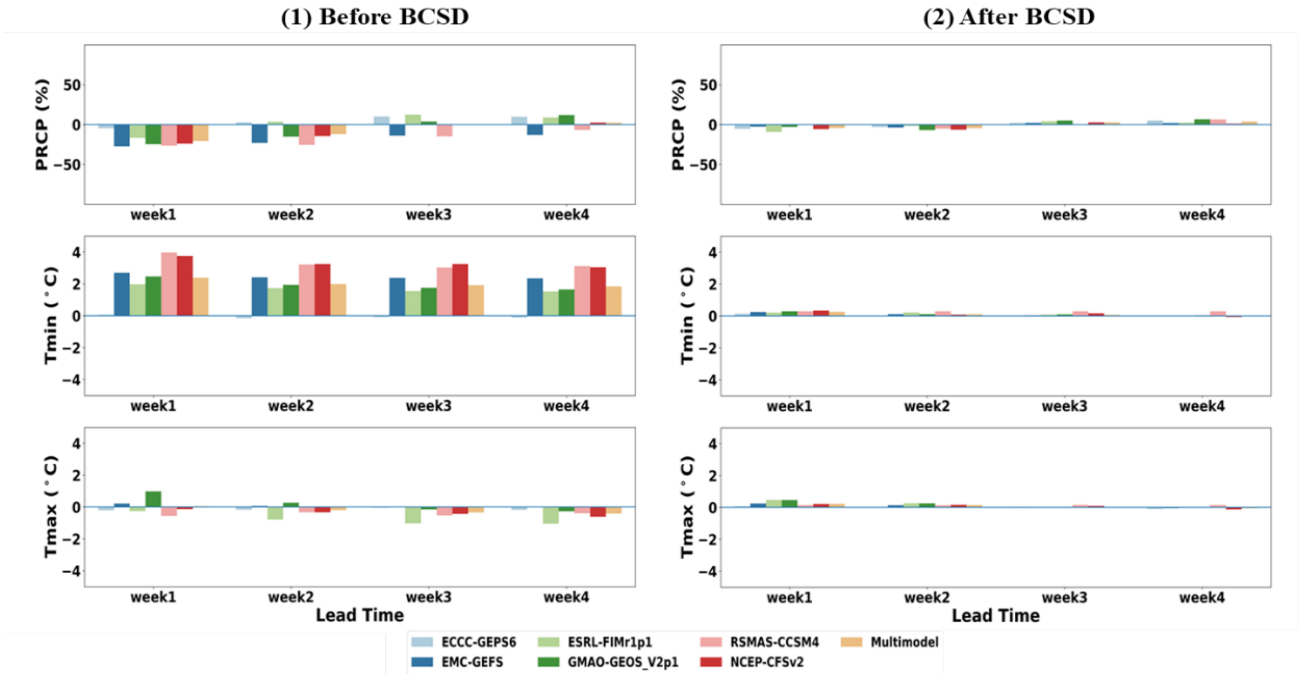



Figure 2.3 Precipitation, Tmax and Tmin bias of SubX models averaged over representative basins and over October–March before and after bias correction.

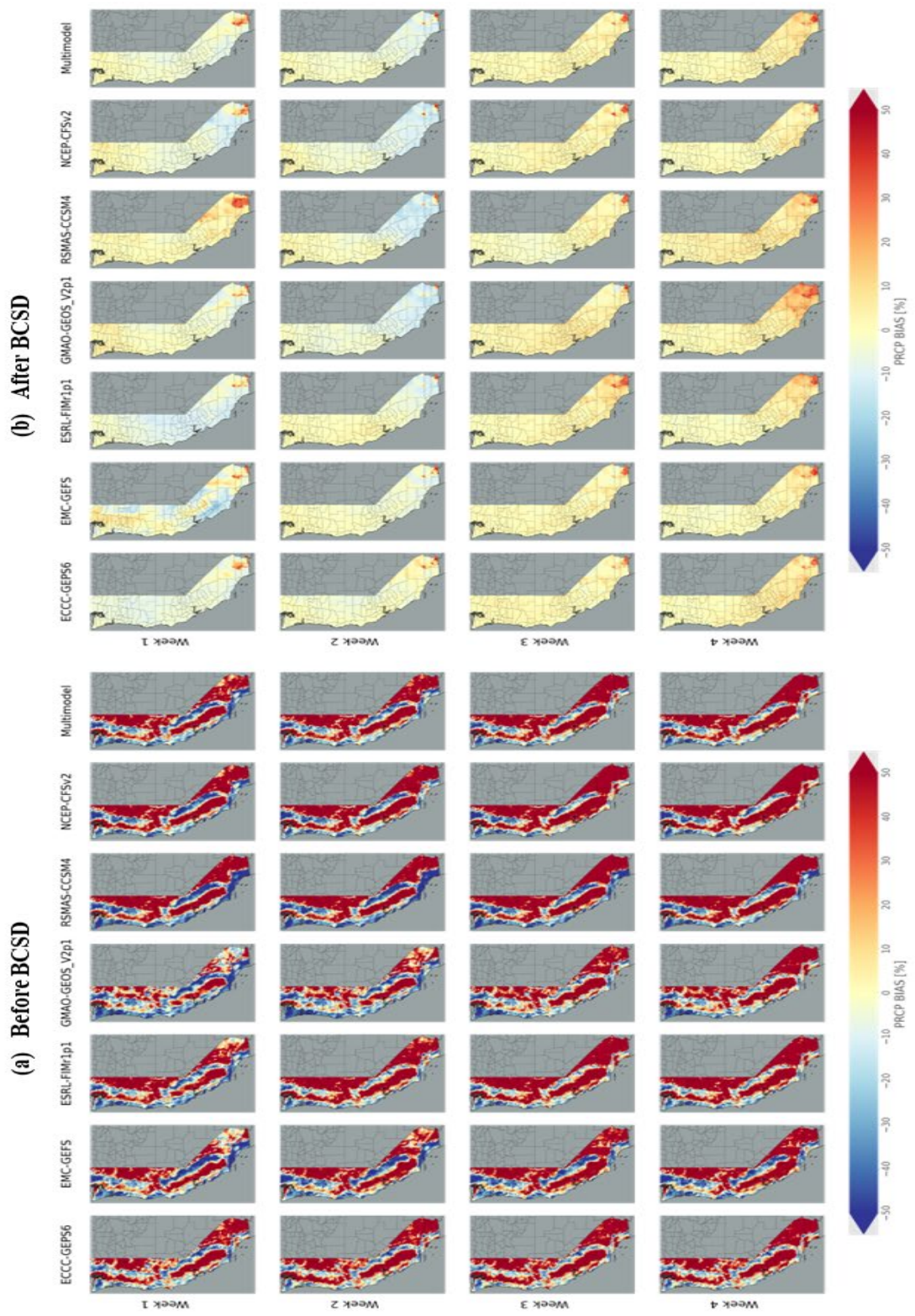



Figure 2.4 Spatial distribution of precipitation bias of SubX models over October–March before (1) and after (2) bias correction.

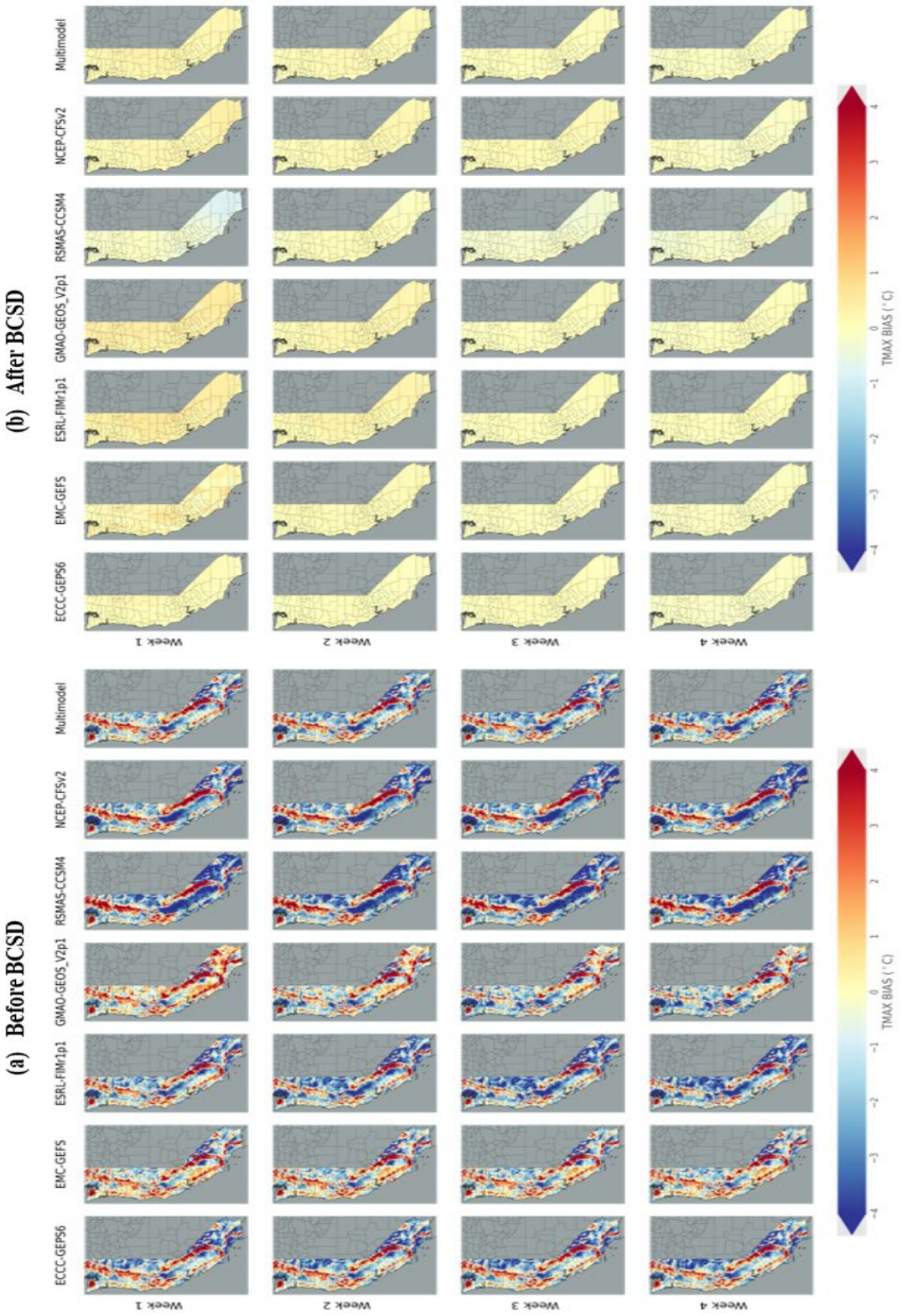



Figure 2.5 Spatial distribution of TMAX bias of SubX models over October–March before and after bias correction.

sparsely distributed and in most cases are only available for a decade or so at most. We nonetheless argue that use of model output soil moisture is plausible based on our past work and work of others. For instance, Su et al. (2021) compared the Livneh et al. (2013) forced Noah-MP simulated soil moisture with observed soil moisture from USDA/NRCS SCAN (Soil Climate Analysis Network) across CONUS. Their results showed in general that the spatial patterns of abnormally low soil moisture in the Noah-MP model constructions are similar to those in the observations. Furthermore, as shown in Appendix A Text S2 and noted in section 3.3, our comparison here of Noah-MP soil moisture with VIC soil moisture yielded similar results. We might, alternatively, have used soil from one of several coupled land-atmosphere reanalyses, e.g., ERA-5 (ECMWF, 2017). ERA-5 soil moisture was found to have the highest skill among reanalysis products compared to in situ observations of soil moisture by Alessi et al. (2022) and Li et al. (2020). However, it was less accurate than soil moisture produced by the LSM-based North American Land Data Assimilation System (NLDAS), and in particular the Noah LSM (Xia et al., 2012; Alessi et al, 2022). We opted therefore not to use reanalysis soil moisture (e.g., ERA5) in consideration of the above studies, and also because of root-zone soil moisture discontinuities issues at the transition points of some of the ERA5 production streams (Hersbach et al. 2020).

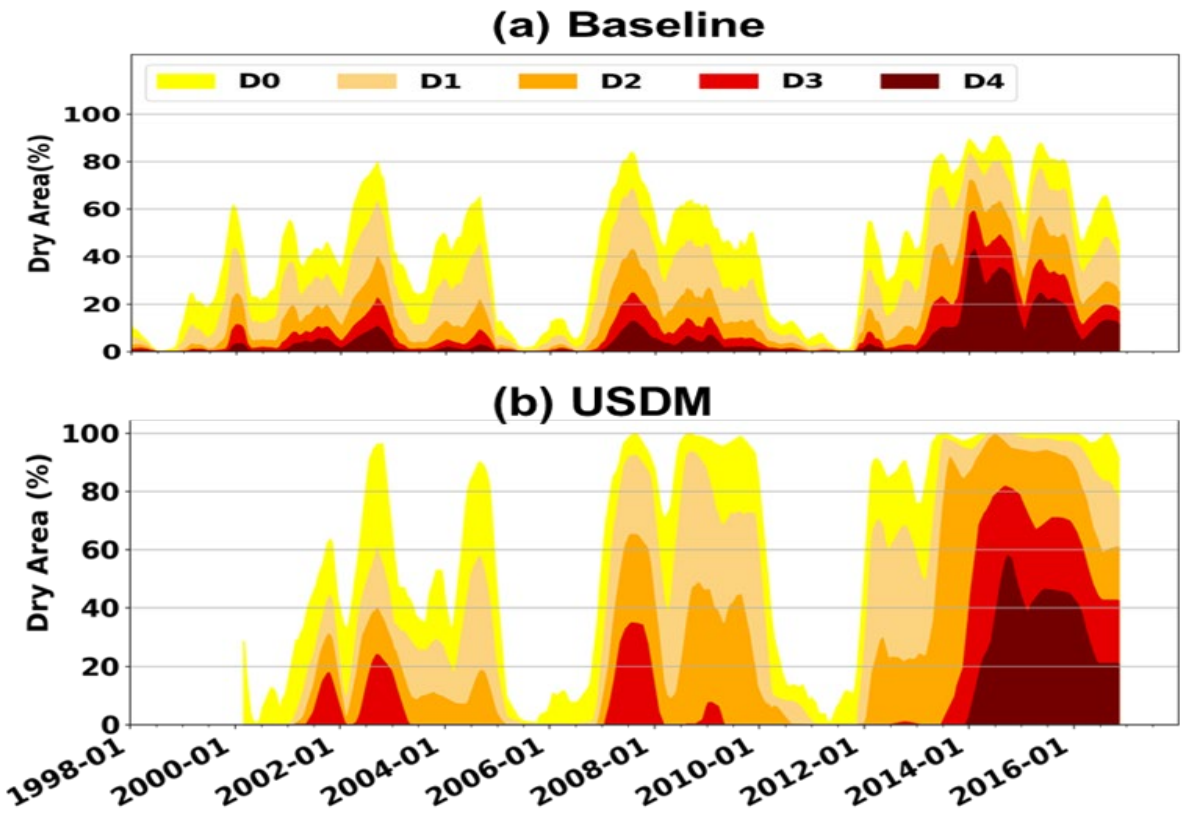



Figure 2.6 California drought (D0-D4) area time series for different drought levels from (a) baseline (driven by Livneh et al. (2013) forcing) and (b) USDM.

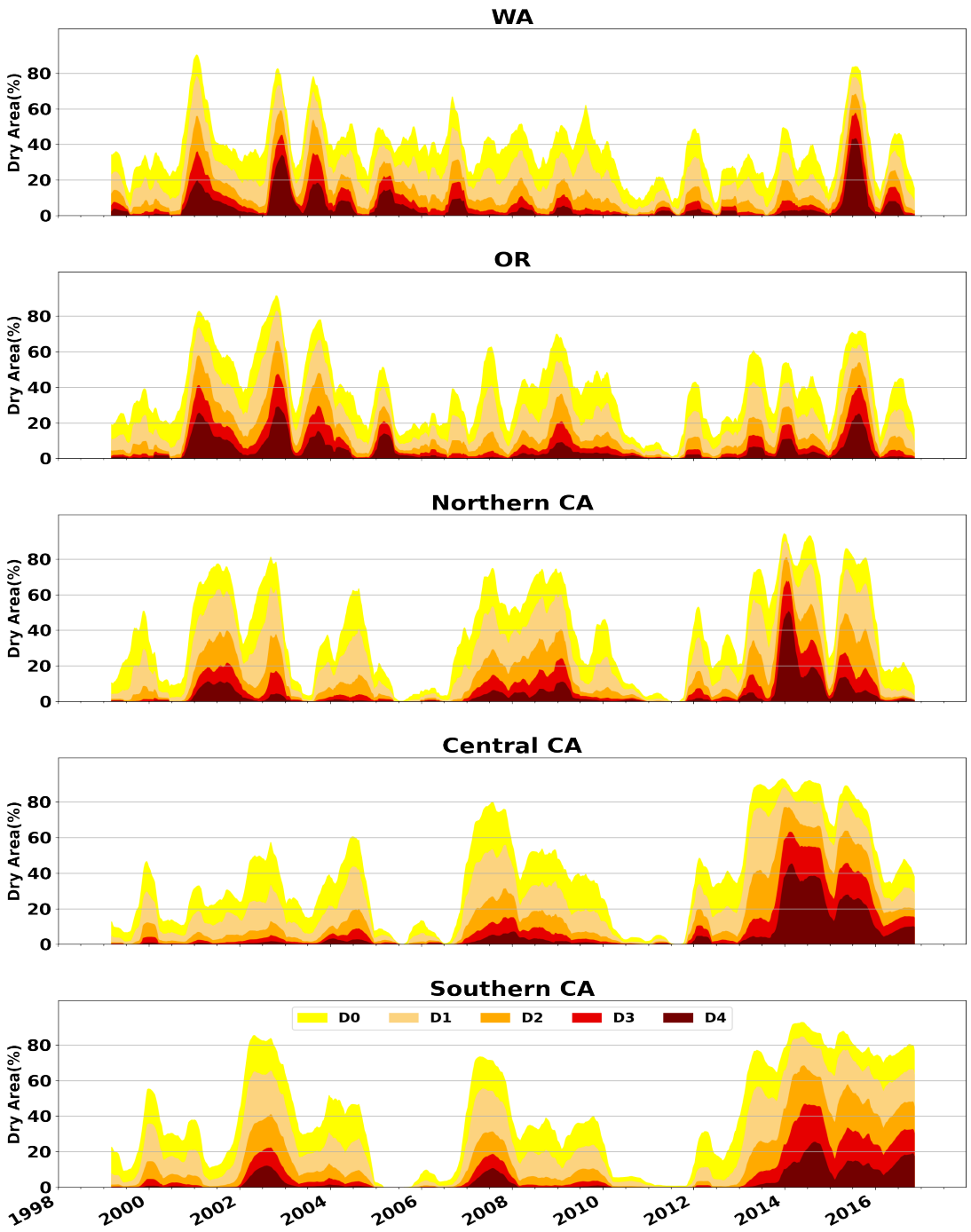



Figure 2.7 Baseline drought area time series for different drought levels for five subregions (coastal Washington, coastal Oregon, northern California, central California and southern California from north to south).

**c. Assessment of Drought Forecast Skill**

Figure 2.8 shows the SubX-based BSS values for major drought termination at leads week 1–4. Here we define major droughts at the grid cell level as a) drought period > 50 days, and b) the drought event is separated by at least 30 days from any other drought. The drought termination and onset forecast is defined as a hit when the forecasted date and the observed date fall within a one week window. We found that drought termination skill is highest for D0 drought and lead week 1. Here we show median results of the 30 ensembles. At lead week 1, we see widespread high skill (BSS score higher than 0.4~0.5) for droughts D0 - D2 (except for southern CA for D2, Figure 2.8). The skill drops to negative for D3 in large parts of southern and central CA and part of OR. The decreasing skill spreads further in CA and OR for D4. At lead week 2, the skills for D0-D2 are still relatively high (BSS score around 0.2~0.3 for most part, except for southern CA for D2). We see more widespread negative skill in D2-D4 compared with at week 1. At lead week 3, there is some limited skill for D0 -D2. At week 4, most of our study domain shows no skill for D0-D4 (except a small part of inland southern CA and WA). Overall, the skill decreases as the drought severity increases and also as the lead time increases. From a spatial perspective, skill decreases from north to south. Figure 2.9 shows the SubX-based BSS values for drought onset at lead week 1-4. We see usable onset skill in lead week 1 and 2 for droughts D0-D2 over most of WA, OR northern and central CA. Overall, onset skill is a little lower than termination skill. The onset skill also decreases with drought severity and lead time and decreases from north to south. To reduce noise spatially, we averaged the soil moisture for the subregions shown in Figure A5 and assessed the drought forecast skills at different subregions (see Appendix A Text S3 for details). The skills at the subregion level are generally consistent with what we found from grid cell-based skills.

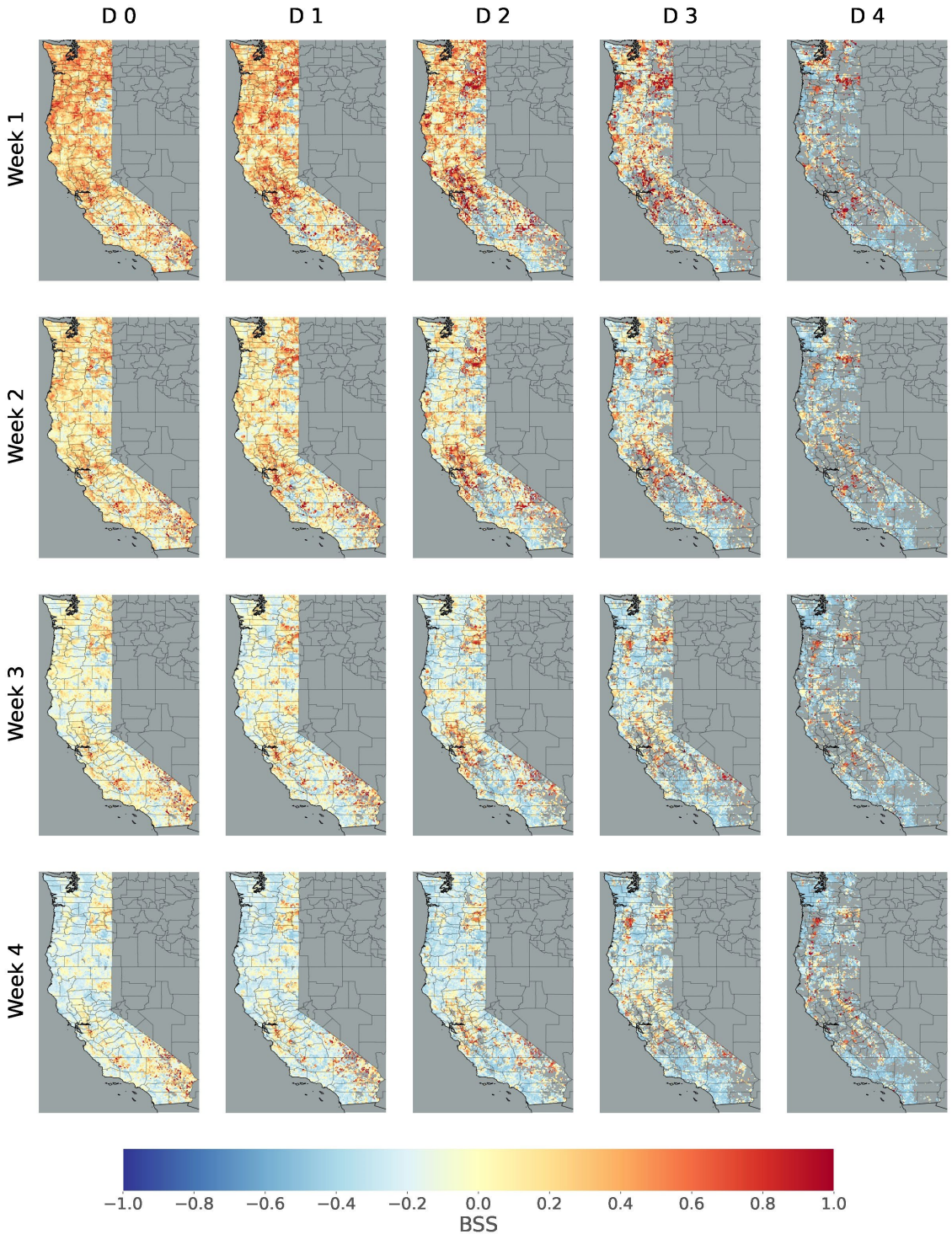



Figure 2.8 SubX-based debiased Brier skill score (BSS) for lead weeks 1-4 for drought termination. The columns show results for drought levels D0-D4; the rows show leads from week1 to week4. Blank areas denote no drought at this level in this location.

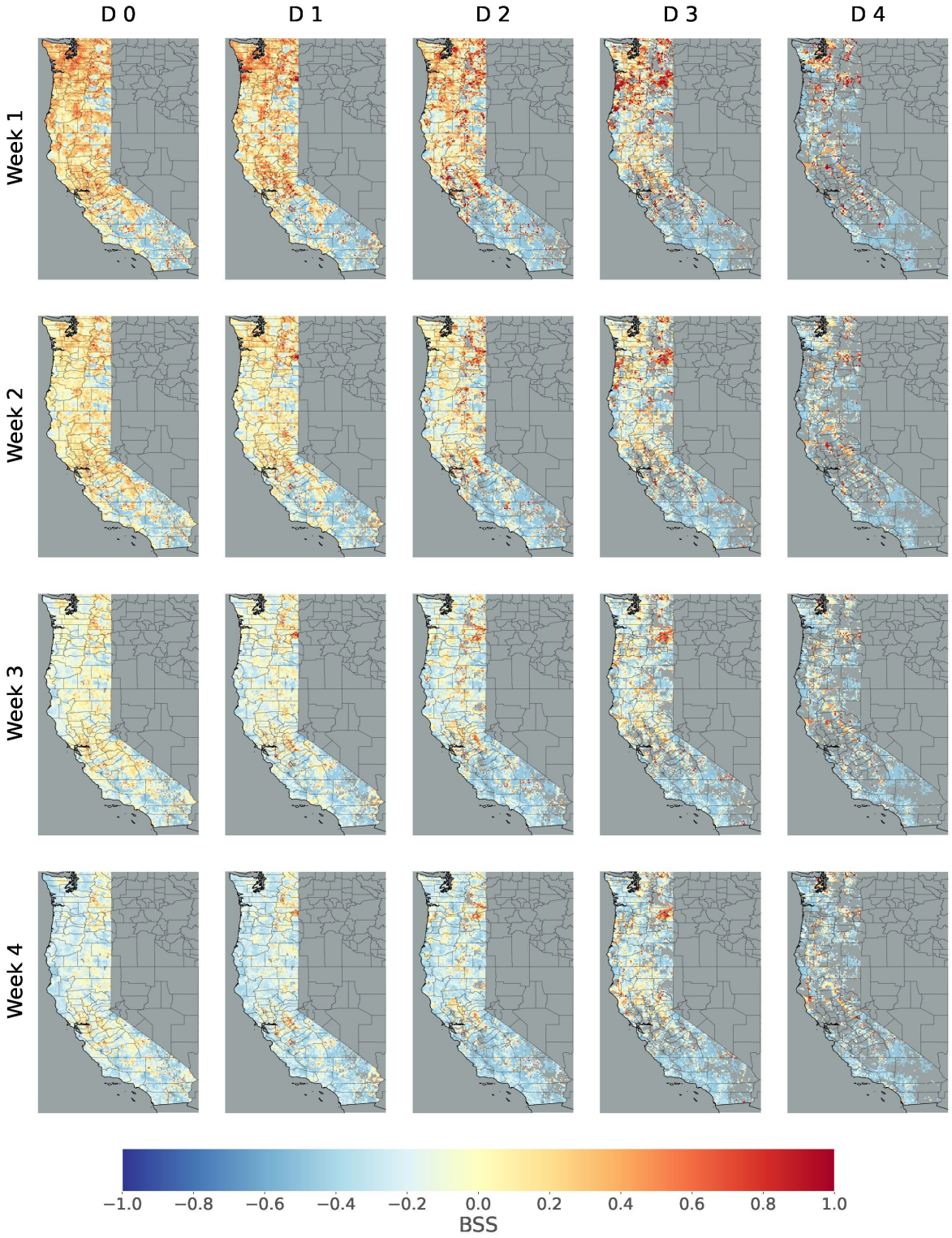



Figure 2.9 SubX-based debiased Brier skill score (BSS) for lead weeks 1-4 for drought onset. Columns show drought levels D0-D4; rows show leads from week1 to week4.

The drought forecast skill is highly related to precipitation forecast skill. Li et al. (2021) found a similar degradation pattern of SubX precipitation forecast skill from north to south over the coastal Western U.S. for most of the models and at all lead times (weeks). This might explain the north to south decreasing drought forecast skill we found here. Atmospheric rivers (AR) play a critical role as a common cause of the end of droughts on the West Coast (Dettinger et al. 2013). The high skill of drought termination at lead week 3-4 in southern CA and WA might be related to the high AR forecast skill in these regions. DeFlorio et al (2019) found isolated positive skill over these locations at weeks 3–4 lead for strong AR activities in some of the SubX models.

Figure 2.10 shows forecast POD for major D1 drought continuance, termination and onset for the five subregions and for different models at 2-week lead time. We summarized the POD (hit rates) based on the percent detection at the grid cell level. A forecast of drought continuance is counted as hit when the drought remains through the predicted period. The forecast of continuance is evaluated relative to persistence, defined as drought conditions assumed to persist through the period (if there is no drought in the beginning, then it's assumed no drought in the end; if there is drought in the beginning, then it's assumed drought in the end). The figure shows that skill for forecasts of continuance is consistently high in all regions and across all models, Skill for forecasts of termination is higher in the north than in the south. Except for forecasts of termination in WA, which have skill comparable to persistence, all other regions' onset and termination forecast skill are lower than persistence. We see very low forecast termination skill and very high continuance skill in southern CA. The reason might be that (a) the precipitation forecast skill in southern CA is comparatively lower (figure 2.4), which leads to lower soil hydrological forecast skill; and (b) drought events in southern CA are very prolonged and the drought event pool is small particularly

during the SubX time period. Fewer events give a false prediction more weight in the calculation of POD and this may reduce apparent drought termination skill.

The previous analyses all examined major droughts. We also want to know if the patterns for major droughts are similar to those for more modest drought events. Thus we also examined all drought events without restrictions on drought length. We calculated the ETS, HSS, POD FAR and BIAS score for drought termination, at grid cell scale at 2-week lead time (Figure 2.11). Using all 30 ensembles, we evaluated the best condition and the median condition among all ensemble members. For ETS, HSS and POD, positive values indicate skill. ETS for drought termination is ~0.3 in coastal WA and OR and southern and central CA in the best condition and ~0.2 in the median condition. HSS and POD are as high as high ~0.4 - ~0.6 in the above locations in the best condition and ~0.2 - ~0.3 in the median condition. These metrics all show the lowest skill in southern CA. FAR results show higher false alarms in the south (especially southern CA) and lower in the north. The bias score is almost 1 in most of our study area in the best condition, indicating almost no bias in this case. We see scattered high bias (overforecast, mostly in inland southern CA and inland WA) and low bias (underforecast, mostly in CA and OR) in the median condition. In summary, all the metrics show the same general trend as for major droughts: higher skill in the north and lower in the south. We repeated the same procedure for drought onset (Figure 12) and found similar patterns from north to south, however the overall forecast skill for onset is lower than for drought termination.

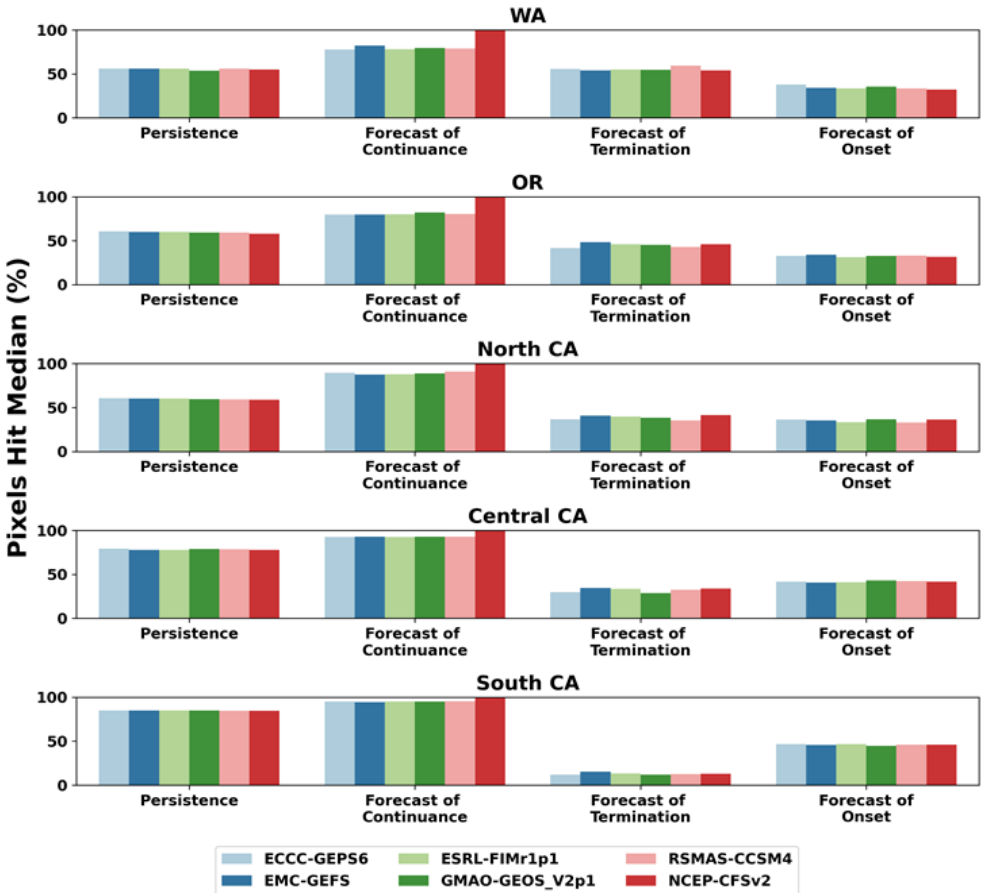



Figure 2.10 Drought persistence, continuance, termination and onset forecast skill for D1 drought at 2-week lead time by subregions and by SubX models.

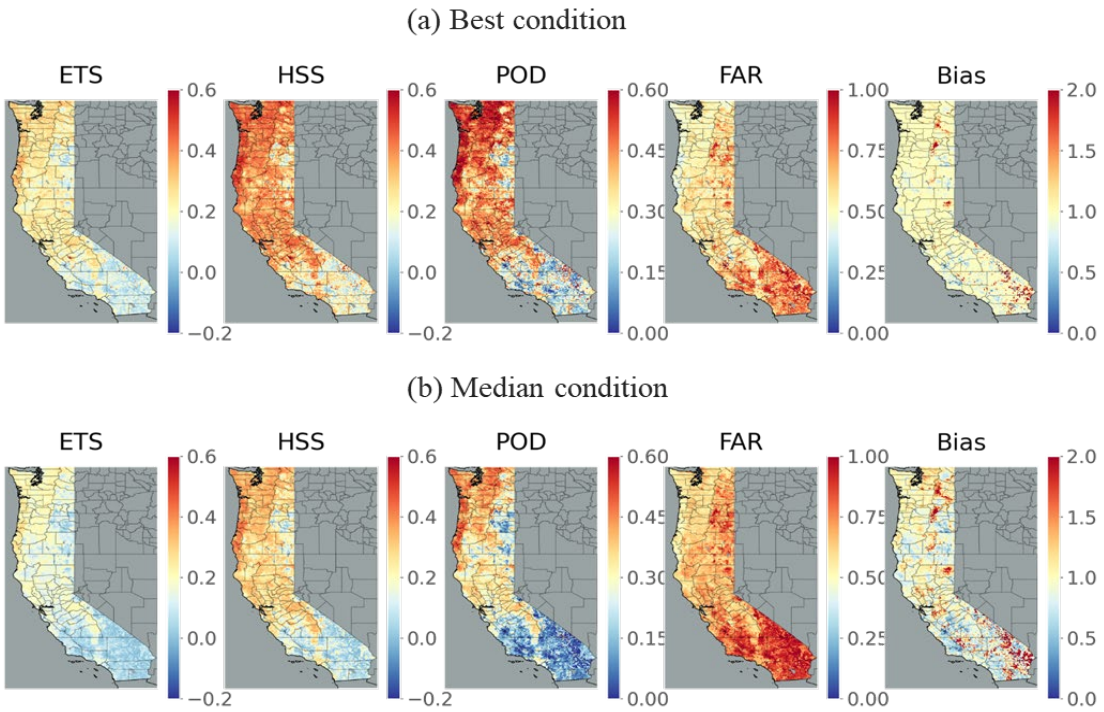



Figure 2.11 ETS, HSS, POD, FAR and Bias Score for drought termination in (a) best condition, (b) median condition across all ensembles at 2-week lead time.

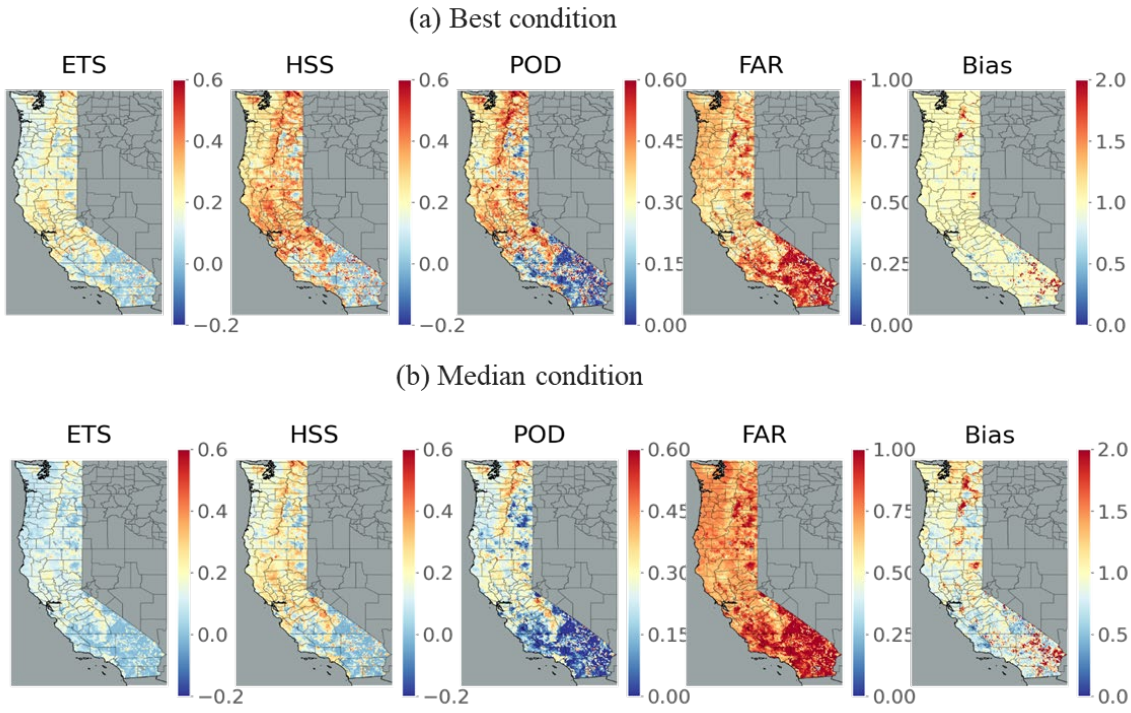



Figure 2.12 ETS, HSS, POD, FAR and Bias Score for drought onset in (a) best condition, (b) median condition across all ensembles at 2-week lead time.

## 2.5 Conclusions

We examined the performance of SubX-driven forecasts of droughts in the coastal Western U.S. with leads from 1 to 4 weeks. We first evaluated SubX reforecasts of precipitation and temperature. Our findings with respect to SubX precipitation and temperature skill are similar to previous studies (e.g., Cao et al. 2021). After statistical downscaling and bias correction of the forcings, we ran the Noah-MP LSM over the domain for the period 1999-2016. We then evaluated skill of SubX-based drought forecasts with a focus on drought termination and onset by using a variety of metrics. We evaluated both major droughts and more modest events.

Based on our analysis, we found usable drought termination and onset forecast skill at week 1 and 2 leads for major D0-D2 droughts; we found limited skill at week 3 for major D0-D1 droughts and essentially no skill at week 4. Drought prediction skills generally decline with increasing drought severity. We found the skill is generally higher on termination than for onset for both major and all drought events. We also found drought prediction skill generally increases from south to north for both major and all drought events.

S2S forecasting of meteorological and hydrologic variables is an active research topic that is attracting significant attention from both the research community (Vitart et al. 2017&2018; DeFlorio et al. 2019; Pan et al. 2019; Zhu et al. 2018) and the applications and stakeholders' communities (including public health, agriculture, emergency management and response sectors, and water resource management, e.g., White et al. 2017&2022; Robertson et al. 2020). We acknowledge, however, that S2S forecasting is still a maturing area. The drought forecast skill (in onset and termination) that we find is highly dependent on precipitation forecast skill. Precipitation forecast products with finer resolution and higher skill likely will improve drought forecast skill. Future studies could extend our work to more extreme events like floods and explore the usefulness

of including higher resolution of forecast products. Exploiting of the large-scale climate drivers might also benefit by identifying additional sources of skill (e.g., El Niño–Southern Oscillation (ENSO), the Madden–Julian Oscillation (MJO) and North Atlantic Oscillation (NAO)) (DeFlorio et al. 2019; White et al. 2022). Finally, employing artificial intelligence and machine learning techniques (e.g., Chapman et al. 2019; Bouaziz et al. 2021; Qian et al. 2021) may have the potential to improve S2S prediction skill.

**Data availability statement**

NCEP-CFSv2 output were obtained from https://www.ncdc.noaa.gov/data-access/model-data/ model-datasets/climate-forecast-system-version2-cfsv2, and output for the other models were obtained from the IRI data library http://iridl.ldeo.columbia.edu/SOURCES/.Models/.SubX/). The LSMs simulation results are available at https://doi.org/10.6084/m9.figshare.21508047.v1

**References**

Alessi, M. J., D. A. Herrera, C. P. Evans, A. T. DeGaetano and T. R. Ault, 2022: Soil Moisture Conditions Determine Land-Atmosphere Coupling and Drought Risk in the Northeastern United States. Journal of Geophysical Research: Atmospheres, 127(6), e2021JD034740.

Arnal, L., A. W. Wood, E. Stephens, H. L. Cloke, F. Pappenberger, 2017: An efficient approach for estimating streamflow forecast skill elasticity. Journal of Hydrometeorology, 18(6), 1715-1729.

Arsenault, K.R., and Coauthors, 2020: The NASA hydrological forecast system for food and water security applications. Bulletin of the American Meteorological Society, 101(7), E1007-E1025.

Baker, S. A., A. W. Wood, and B. Rajagopalan, 2019: Developing subseasonal to seasonal climate forecast products for hydrology and water management. J. Amer. Water Resour. Assoc., 55, 1024–1037, https://doi.org/10.1111/1752-1688.12746.

Bennett et al., 2020: MetSim: A Python package for estimation and disaggregation of meteorological data. Journal of Open Source Software, 5(47), 2042, https://doi.org/10.21105/joss.02042

Bohn, T. J., B. Livneh, J. W. Oyler, S. W. Running, B. Nijssen, and D. P. Lettenmaier, 2013: Global evaluation of MTCLIM and related algorithms for forcing of ecological and hydrological models. Agric. For. Meteor., 176, 38–49, https://doi.org/10.1016/j.agrformet.2013.03.003.

Bouaziz, M., E. Medhioub and E. Csaplovisc, 2021: A machine learning model for drought tracking and forecasting using remote precipitation data and a standardized precipitation index from arid regions. Journal of Arid Environments, 189, 104478.

Cao, Q., S. Shukla, M.J. DeFlorio, F.M. Ralph, and D.P. Lettenmaier, 2021: Evaluation of the Subseasonal Forecast Skill of Floods Associated with Atmospheric Rivers in Coastal Western US Watersheds. Journal of Hydrometeorology, 22(6), 1535-1552.

Carrão, H., G. Naumann, E. Dutra, C. Lavaysse, and P. Barbosa, 2018: Seasonal drought forecasting for Latin America using the ECMWF S4 forecast system. Climate, 6(2), 48.

Chapman, W., A. C. Subramanian, L. Delle Monache, S.P. Xie and F. M. Ralph, 2019: Improving atmospheric river forecasts with machine learning. Geophysical Research Letters, 46, 10627–10635. https://doi.org/10.1029/2019GL083662

Chen, F., K. Mitchell, J. Schaake, Y. Xue, H. Pan, V. Koren, Y. Duan, M. Ek, and A. Betts, 1996: Modeling of land-surface evaporation by four schemes and comparison with FIFE observations. Journal of Geophysical Research, 101, 7251-7268.

Chen, F. and J., Dudhia, 2001: Coupling an advanced land surface–hydrology model with the Penn State–NCAR MM5 modeling system. Part I: Model implementation and sensitivity. Monthly weather review, 129(4), 569-585.

Christensen, J. et al, 2007: The Physical Science Basis (eds Solomon, S. et al.) IPCC, Cambridge Univ. Press, Ch. 11,.

Cook, B. I., A. P. Williams, J. S. Mankin, R. Seager, J. E. Smerdon, and D. Singh, 2018: Revisiting the leading drivers of Pacific coastal drought variability in the contiguous United States. Journal of Climate, 31(1), 25-43.

DeAngelis, A.M., H. Wang, R.D. Koster, S.D. Schubert, Y. Chang, and J. Marshak, 2020: Prediction skill of the 2012 US Great Plains flash drought in subseasonal experiment (SubX) models. Journal of Climate, 33(14), pp.6229-6253.

DeFlorio, M.J. et al, 2019: Experimental subseasonal-to-seasonal (S2S) forecasting of atmospheric rivers over the western United States. J. Geophys. Res. Atmos., 124, 11 242–11 265, https://doi.org/10.1029/2019JD031200.

Delworth, T.L., F. Zeng, A. Rosati, G.A. Vecchi, and A.T. Wittenberg, 2015: A link between the hiatus in global warming and North American drought. Journal of Climate, 28(9), pp.3834-3845.

Dettinger, M. D. (2013). Atmospheric rivers as drought busters on the US West Coast. Journal of Hydrometeorology, 14(6), 1721-1732.

Engström, J., K. Jafarzadegan, and H. Moradkhani, 2020: Drought vulnerability in the United States: An integrated assessment. Water, 12(7), p.2033.

European Centre for Medium-Range Weather Forecasts (ECMWF). (2017). ERA5 Reanalysis. https://doi.org/10.5065/D6X34W69

Glantz, M., 2004: Early Warning Systems: Do′s and Don′ts. Workshop Report, 20–23 October 2003, Shanghai, China. (www.esig.ucar.edu/warning)

Hao, Z., V. P. Singh, and Y. Xia, 2018: Seasonal drought prediction: advances, challenges, and future prospects. Reviews of Geophysics, 56(1), 108-141.

Hersbach, H., et al, 2020: The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146(730), 1999-2049.

Infanti, J. M., and B. P. Kirtman, 2016: Prediction and predictability of land and atmosphere initialized CCSM4 climate forecasts over North America. J. Geophys. Res. Atmos., 121, 12 690–12 701, https://doi.org/10.1002/2016JD024932.

Koster, R. D., M. J. Suarez, A. Ducharne, M. Stieglitz, and P. Kumar, 2000: A catchment-based approach to modeling land surface processes in a general circulation model: 1. Model structure. J. Geophys. Res., 105, 24 809–24 822, https://doi.org/ 10.1029/2000JD900327.

Kirtman, Ben P., et al., 2014: The North American multimodel ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. Bulletin of the American Meteorological Society 95(4), 585-601.

Krishnamurti, T. N., and Coauthors, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. Science, 285(5433), 1548-1550.

Krishnamurti, T. N., and Coauthors, 2000: Multimodel ensemble forecasts for weather and seasonal climate. Journal of Climate, 13(23), 4196-4216.

Griffin, D., and K. J. Anchukaitis, 2014: How unusual is the 2012–2014 California drought? Geophys. Res. Lett., 41, 9017–9023, doi:10.1002/2014GL062433.

Global Assessment Report on Disaster Risk Reduction, UNDRR, 2019.

Li, H., L. Luo, E. F. Wood, and J., Schaake, 2009: The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting. Journal of Geophysical Research, 114, D04114, https://doi.org/0.1029/2008JD010969.

Li, M., P. Wu and Z. Ma, 2020: A comprehensive evaluation of soil moisture and soil temperature from third-generation atmospheric and land reanalysis data sets. International Journal of Climatology, 40(13), 5744–5766. https://doi.org/10.1002/joc.6549

Li, Y., D. Tian, and H. Medina, 2021. Multimodel Subseasonal Precipitation Forecasts over the Contiguous United States: Skill Assessment and Statistical Postprocessing. Journal of Hydrometeorology, 22(10), 2581-2600.Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J.

Burges, 1994: A Simple hydrologically Based Model of Land Surface Water and Energy Fluxes for GSMs, J. Geophys. Res., 99(D7), 14,415-14,428.

Lin, H., N. Gagnon, S. Beauregard, R. Muncaster, M. Markovic, B. Denis, and M. Charron, 2016: GEPS-based monthly prediction at the Canadian Meteorological Centre. Mon. Wea. Rev., 144, 4867–4883, https://doi.org/10.1175/MWR-D-16-0138.1.

Lin, H., R. Mo, F. Vitart, and C. Stan, 2018: Eastern Canada flooding 2017 and its subseasonal predictions. Atmos.–Ocean, 57, 195–207, https://doi.org/10.1080/07055900.2018.1547679.

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.

Mariotti, A., P. M. Ruti, and M. Rixen, 2018: Progress in subseasonal to seasonal prediction through a joint weather and climate community effort. npj Climate and Atmospheric Science, 1, 4.

Mann, M.E. and P.H. Gleick, 2015: Climate change and California drought in the 21st century. Proceedings of the National Academy of Sciences, 112(13), pp.3858-3859.

Merryfield, W. J., and Coauthors, 2020: Current and emerging developments in subseasonal to decadal prediction. Bull. Amer. Meteor. Soc., 101, E869–E896, https://doi.org/10.1175/BAMS-D-19-0037.1.

Mo, K.C., L.C. Chen, S. Shukla, T.J. Bohn and D.P. Lettenmaier, 2012: Uncertainties in North American land data assimilation systems over the contiguous United States. Journal of Hydrometeorology, 13(3), 996-1009.

Molod, A., and Coauthors, 2012: The GEOS-5 atmospheric general circulation model: Mean climate and development from MERRA to Fortuna. Tech. Memo. NASA/TM-2012-104606, 115 pp., https:// ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20120011790.pdf.

Monhart, S., C. Spirig, J. Bhend, K. Bogner, C. Schär, and M. A.Liniger, 2018: Skill of subseasonal forecasts in Europe: Effect of bias correction and downscaling using surface observations.J. Geophys. Res. Atmo

Mote, P. W., and Coauthors, 2016: Perspectives on the causes of exceptionally low 2015 snowpack in the western United States. Geophys. Res. Lett., 43, 10 980–10 988, doi:10.1002/2016GL069965.

Niu, G. Y., Z. L. Yang, K. E. Mitchell, F. Chen, M. B. Ek, M. Barlage, and M. Tewari, 2011: The community Noah land surface model with multiparameterization options (Noah MP): 1. Model description and evaluation with local scale measurements. Journal of Geophysical Research: Atmospheres, 116.

Niu, G.-Y., Z. L. Yang, R. E. Dickinson, L. E. Gulden, and H. Su, 2007: Development of a simple groundwater model for use in climate models and evaluation with Gravity Recovery and Climate Experiment data. Journal of Geophysical Research, 112, D07103, https://doi.org/10.1029/2006JD007522.

Otkin, J. A. and Coauthors, 2018: Flash droughts: a review and assessment of the challenges imposed by rapid-onset droughts in the United States. Bull. Am. Meteorol. Soc. 99, 911–919.

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. Journal of Climate, 32(1), pp.161-182.

Pegion, K., and Coauthors, 2019: The Subseasonal Experiment (SubX): A multimodel subseasonal prediction experiment. Bulletin of the American Meteorological Society, 100(10), pp.2043-2060.

Pendergrass, A.G., and Coauthors, 2020: Flash droughts present a new challenge for subseasonal-to-seasonal prediction. Nat. Clim. Change 10, 191–199. https://doi.org/10.1038/s41558-020-0709-0

Pierce, D. W., D. R. Cayan, and B. L. Thrasher, 2014: Statistical downscaling using Localized Constructed Analogs (LOCA). J. Hydrometeor., 15, 2558–2585, https://doi.org/10.1175/JHM-D-14-0082.1.

Pulwarty, R. S. and J. P. Verdin, 2013: Measuring Vulnerability to Natural Hazards: Towards Disaster Resilient Societies 2nd edn (ed. Birkmann, J.) ,United Nations Univ. Press, 124–147.

Pulwarty, R.S. and M.V. Sivakumar, 2014. Information systems in a changing climate: Early warnings and drought risk management. Weather Clim. Extrem., High Level Meeting on National Drought Policy 3, 14–21. https://doi.org/10.1016/j.wace.2014.03.005

Qian, Q., Jia, X., Lin, H., & Zhang, R. (2021). Seasonal forecast of nonmonsoonal winter precipitation over the Eurasian continent using machine-learning models. Journal of Climate, 34(17), 7113-7129.

Reichle, R., and Q. Liu, 2014: Observation-corrected precipitation estimates in GEOS-5. Tech. Memo. NASA/TM-2014-104606, 18, https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20150000725.pdf.

Rienecker, M. M., and Coauthors, 2008: The GEOS-5 Data assimilation system-documentation of versions 5.0.1, 5.1.0, and 5.2.0. Tech. Memo. NASA/TM-2008-104606, 97., https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20120011955.pdf.

Robertson, A. W., F. Vitart, & S. J. Camargo, 2020: Subseasonal to seasonal prediction of weather to climate with application to tropical cyclones. Journal of Geophysical Research: Atmospheres, 125(6), e2018JD029375.

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.

Seager, R., M. Hoerling, S. Schubert, H. Wang, B. Lyon, A. Kumar, J. Nakamura, and N. Henderson, 2015: Causes and predictability of the 2011–2014 California drought. NOAA Drought Task Force, 40 pp., http://cpo.noaa.gov/Portals/0/Docs/MAPP/TaskForces/DTF/california_drought_report.pdf.

Seager, R., and M. Hoerling, 2014: Atmosphere and ocean origins of North American droughts. J. Climate, 27, 4581–4606, https://doi.org/10.1175/JCLI-D-13-00329.1.

Seneviratne, S. I., and Coauthors, 2012: Special Report on Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation (eds Field, C. B. et al.), IPCC, Cambridge Univ. Press, 109–230.

Schick, S., O. Rössler, and R. Weingartner, 2019: An evaluation of model output statistics for subseasonal streamflow forecasting in European catchments. J. Hydrometeor., 20, 1399–1416, https://doi.org/10.1175/JHM-D-18-0195.1.

Shukla, S., N. Voisin and D.P. Lettenmaier, 2012: Value of medium range weather forecasts in the improvement of seasonal hydrologic prediction skill. Hydrology and Earth System Sciences, 16(8), pp.2825-2838.

Shukla, S. and D.P. Lettenmaier, 2011. Seasonal hydrologic prediction in the United States: understanding the role of initial hydrologic conditions and seasonal climate forecast skill. Hydrology and Earth System Sciences, 15(11), pp.3529-3538.

Su, L., Q. Cao, M. Xiao, D.M. Mocko, M. Barlage, D. Li, C.D. Peters-Lidard, and D.P. Lettenmaier, 2021: Drought variability over the conterminous United States for the past century. J. Hydrometeor., 22, 1153–1168, https://doi.org/10.1175/ JHM-D-20-0158.1.

Sun, S., R. Bleck, S. G. Benjamin, B. W. Green, and G. A. Grell, 2018a: Subseasonal forecasting with an icosahedral, vertically quasi-Lagrangian coupled model. Part I: Model overview and evaluation of systematic errors. Mon. Wea. Rev., 146, 16011617, https://doi.org/10.1175/MWR-D-18-0006.1.

Sun, S., B. W. Green, R. Bleck, and S. G. Benjamin, 2018b: Subseasonal forecasting with an icosahedral, vertically quasiLagrangian coupled model. Part II: Probabilistic and deterministic forecast skill. Mon. Wea. Rev., 146, 1619–1639, https:// doi.org/10.1175/MWR-D-18-0007.1.

Vitart, F., and Coauthors, 2017: The subseasonal to seasonal (S2S) prediction project database. Bulletin of the American Meteorological Society, 98(1), 163-173.

Vitart, F. and A. W. Robertson, 2018: The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. npj Climate and Atmospheric Science, 1(1), 1-7.

Wang, S., A. Anichowski, M.K. Tippett, and A.H. Sobel, 2017: Seasonal noise versus subseasonal signal: Forecasts of California precipitation during the unusual winters of 2015–2016 and 2016–2017. Geophysical Research Letters, 44(18), pp.9513-9520.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2007: The discrete Brier and ranked probability skill scores. Mon. Wea. Rev., 135, 118–124, https://doi.org/10.1175/MWR3280.1.

White, C. J., Carlsen, H., Robertson, A. W., Klein, R. J. T., Lazo, J. K., Kumar, A., et al. (2017). Potential applications of subseasonal-toseasonal (S2S) predictions. Meteorological Applications, 24(3), 315–325. https://doi.org/10.1002/met.1654

White, C. J., et al., 2022: Advances in the application and utility of subseasonal-to-seasonal predictions. Bulletin of the American Meteorological Society, 103(6), E1448-E1472.

Wilhite, D. A., M. V. K. Sivakumar, and R. Pulwarty, 2014: Managing drought risk in a changing climate: the role of national drought policy. Weather Clim. Extrem. 3, 4–13.

Wilks, D. S., 2006. Statistical Methods in the Atmospheric Sciences. 2nd ed. International Geophysics Series, Vol. 100, Academic Press, 648 pp.

Williams, A. P., R. Seager, J. T. Abatzoglou, B. I. Cook, J. E. Smerdon, and E. R. Cook, 2015: Contribution of anthropogenic warming to California drought during 2012–2014. Geophys. Res. Lett., 42, 6819–6828, doi:10.1002/2015GL064924.

Wanders, N. N., et al., 2017: Forecasting the hydroclimatic signature of the 2015/16 El Niño event on the Western United States. Journal of Hydrometeorology, 18, 177–186. https://doi.org/10.1175/JHM-D-16-0230.1

Wood, L. R. Leung, V. Sridhar, and D. P. Lettenmaier, 2004: Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. Climatic Change, 62, 189–216, https://doi.org/10.1023/B:CLIM.0000013685.99609.9e.

Wood, E. P. Maurer, A. Kumar, and D. P. Lettenmaier, 2002: Long-range experimental hydrologic forecasting for the eastern United States. J. Geophys. Res., 107, 4429, https://doi.org/10.1029/2001JD000659.

Xia, Y., et al., 2012: Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System Project Phase 2 (NLDAS-2): 1. Intercomparison and application of model products. Journal of Geophysical Research, 117, D03109. https://doi.org/10.1029/2011JD016048

Zheng, H., and Coauthors, 2019: On the sensitivity of the precipitation partitioning into evapotranspiration and runoff in land surface parameterizations. Water Resources Research, 55(1), pp.95-111.

Zhou, X., Y. Zhu, D. Hou, and D. Kleist, 2016: A comparison of perturbations from an ensemble transform and an ensemble Kalman filter for the NCEP Global Ensemble Forecast System. Wea. Forecasting, 31, 2057–2074, https://doi.org/ 10.1175/WAF-D-16-0109.1.

Zhou, X., Y. Zhu, D. Hou, Y. Luo, J. Peng, and R. Wobus, 2017: Performance of the new NCEP Global Ensemble Forecast System in a parallel experiment. Wea. Forecasting, 32, 1989–2004, https://doi.org/10.1175/WAF-D-17-0023.1.

Zhu, Y., and Coauthors, 2018: Toward the improvement of subseasonal prediction in the national centers for environmental prediction global ensemble forecast system. J. Geophys. Res., 123, 6732–6745, https://doi.org/10.1029/2018JD028506.

## Chapter 3 Improving National Water Model Flood Forecast Skills over Coastal Western U.S. River Basins

This chapter will be submitted to the Journal of Hydrometeorology as

Lu Su, Dennis P. Lettenmaier, Robert K. Hartman, Ming Pan, 2023: Improving Noah-MP Flood Forecast Skills over Coastal Western U.S. River Basins. Journal of Hydrometeorology, (in preparation).

The supplemental material for this chapter is provided in Appendix B.


### Abstract

Flooding is one of the deadliest and costliest of natural hazards. In 2016, NOAA launched the National Water Model (NWM), a comprehensive hydrological modelling system intended for use across the U.S.. Noah-MP is the hydrologic core of NWM and provides new technology for producing flood forecasts. However, there have been reports that Noah-MP-based flood forecasts are less accurate in the Western U.S. than are current methods and that additional effort needs to be devoted to selection of Noah-MP physics options and improving its calibration. Here, we identify the best Noah-MP physics options and calibrate the model's parameters in seven watersheds that form a transect along the U.S. Pacific Coast. Our results show that when using the default free drainage option, the resulting (baseline) flood simulations achieved the best performance across the study basins. We then calibrated six parameters that control soil moisture, runoff and groundwater using the Dynamically Dimensioned Search (DDS) automatic calibration method. After calibration, simulation performance improved greatly. We then constructed reforecasts for the largest flood events of the past 7-20 years by running Noah-MP with the calibrated land surface parameters and Quantitative Precipitation Forecast (QPF) precipitation (the latter produced by the two coastal Western U.S. River Forecast Centers (RFCs). We compared

Noah-MP flood reforecasts with RFC archived forecasts and found that for both POT 3 floods and major floods, Noah-MP's forecasting performance was comparable with NWRFC in the three northern basins but was inferior to RFC in the four southern basins, especially in terms of timing and magnitude variability. While both models tended to underestimate flood peaks, Noah-MP's discrepancies increased more rapidly with longer lead times and typically predicted earlier peaks. Meanwhile, for the largest floods, Noah-MP and RFC forecasts were largely comparable in magnitude, but Noah-MP often predicted earlier peaks, with its performance varying event by event. This research highlights the potential of Noah-MP for flood forecasting, particularly in the northern basins, provided that suitable parameter selection and calibration are employed while improvements in the southern (relatively dry) river basins.

### 3.1 Introduction

Flooding is one of the deadliest and costliest natural hazards in the U.S., with an annual average cost of $10.25 billion (adjusted to 2023 inflation) and 85 fatalities per year on average between 1984 and 2013 (National Weather Service 2014). The flood cost in water year 2017 alone was 76 billion (adjusted to 2023 inflation, National Weather Service 2017), marking the greatest amount in the previous twenty years. While the frequency of these events has remained steady over the last hundred years, their consequences have evolved owing to societal growth and swift societal transitions (Institute for Business & Home Safety 2001). This evolution can be observed in instances such as the increase in population residing in flood-prone peripheral areas (National Hydrologic Warning Council 2002). Effective flood forecasting is crucial to the country's ability to mitigate flood damages, reduce associated costs, and save lives (Rogers and Tsirkunov, 2011; National Hydrologic Warning Council 2002; Kundzewicz and Kaczmarek, 2000). Therefore,

improving flood forecasting is a priority for reducing these impacts and enhancing the resilience of potentially flood-affected communities across the country.

The California Nevada River Forecast Center (CNRFC) and the Northwest River Forecast Center (NWRFC) have long been responsible for issuing flood forecasts within California and Nevada, and the Pacific Northwest regions of the U.S. Both entities are branches of the U.S. National Weather Service (NWS) which itself is a part of the National Oceanic and Atmospheric Administration (NOAA). The RFCs routinely generate Quantitative Precipitation Estimates (QPEs) in 6-hour increments. QPEs are calculated representations of the quantity of precipitation that has fallen over a defined time frame and geographical area. Additionally, the RFCs produce a Quantitative Precipitation Forecast (QPF) with a 6-hour time resolution, which projects the volume of precipitation expected to fall over a specific duration with a lead time of 6 hours to up to 10 days. The CNRFC and NWRFC employ a mix of hydrologic models, QPF methods, and additional meteorological information to forecast forthcoming streamflow conditions at specific sites, known as forecast points, along water bodies (CNRFC, https://www.cnrfc.noaa.gov/qpf.php).

The NWS primarily employs the Sacramento Soil Moisture Accounting (SAC-SMA) model, which originated in the 1960s and 1970s (Burnash et al. 1973). This spatially lumped model produces streamflow simulations based on specified antecedent soil moisture conditions. While the model has proven to be proficient in creating accurate forecasts (Franz et al. 2003), it doesn't represent vegetation directly, nor does it have a basis for connecting the lower- and upper soil moisture zones and the processes that control exchanges within or between them on a physical basis. (Agnihotri et al. 2020; Burnash et al. 1973). Furthermore, neither its evapotranspiration nor its snow accumulation algorithms are physically based, thus it is challenged in capturing key hydrologic processes at the scale of individual catchments (e.g., Salas et al. 2018). Nevertheless,

the distributed model intercomparison project (DMIP) results showed that the SAC-SMA model generally was competitive with more modern distributed models (Reed et al. 2004).

Since its initiation in 2016, NOAA has actively supported development of the operational National Water Model (NWM), a comprehensive hydrological modelling system (NOAA 2016). WRF-Hydro, the Hydrological modelling framework built around the Weather Research and Forecasting Model (Gochis et al. 2020), forms the basis of the NWM. This extension merges the Noah Land Surface Model with multiple parameterization options (Noah-MP; Niu et al. 2011) which forms the hydrologic core of the NWM. One distinguishing attribute of Noah-MP is that it has multiple physics options for various hydrological processes including runoff generation. Noah-MP has not previously been compared with RFC operational forecasts, which are based on SAC-SMA. A comparison of Noah-MP and SAC-SMA for runoff simulation in the coastal Western U.S. would be beneficial for evaluation of the ultimate potential of Noah-MP in operational flood forecast applications.

The simulation performance of the NWM, with Noah-MP being the hydrologic core, has recently begun to be evaluated. Salas et al. (2018) studied three-month Noah-MP nowcasts and demonstrated the capacity to seamlessly predict reach scale streamflow at the continental scale. Their validation of the uncalibrated model using observed hourly streamflow at 5,701 U.S. Geological Survey (USGS) gages shows that about one-quarter demonstrate PBias $\leq |25\%|$, 11% demonstrate Nash-Sutcliffe Efficiency (NSE) $\geq 0.25$; among which they found better performance in the Pacific Northwest, Rocky Mountains, Central U.S., and Eastern U.S., and weak performance in the arid Southwest and Northern Plains. Lahmers et al. (2021) found that the runoff simulation performance of Noah-MP for 56 catchments within the southwest CONUS was improved by adding channel infiltration and performing calibration. Lin et al. (2018) evaluated the Noah-MP

streamflow simulation skill at 271 USGS gauges over Texas. They found that daily streamflow was better predicted in wet regions with the highest NSE ~0.7 and was most poorly predicted in dry regions with a large positive bias. These studies generally evaluated simulations instead of forecasts and no comparisons with operational archived forecasts have been made.

Little previous work has investigated Noah-MP flood forecasts, and the work that has been done has been limited to singular flood events in one basin. For example, Viterbo et al. (2020) investigate the utility of the Noah-MP configured in NWM to predict catastrophic flooding associated with an extreme rainfall event that occurred in Ellicott City, Maryland, on 27–28 May 2018. Their results suggested potential forecast utility in using NWM to predict high impact, local scale flood events, while also underscoring the need to comprehensively evaluate model performance at local scales and for high-impact, rare events.

Previous studies have emphasized the importance of calibration in conceptual models (e.g., Gupta et al. 2008 and references therein), and work by Duan et al. (2006) and Lahmers et al. (2021) (among others) has highlighted the positive impact of calibration on the performance of spatially distributed hydrological models as well. Mascaro et al. (2023) manually calibrated the parameter values of Noah-MP through a stepwise approach in Oak Creek Basin in central Arizona and found the flood simulation performance improved notably after streamflow calibration for years 2008–2011.This work suggests the potential for enhancing Noah-MP performance through calibration, especially compared with Noah-MP applications using default parameter estimates.

Notably absent from the literature is the exploration of multiple floods across a variety of watersheds, as well as an evaluation of physics parameterization options as we do here. In the same context, we perform a comprehensive analysis of the value of automatic calibration for Noah-MP -based flood forecasts. Our intention is to help refine the selection of physics options for Noah-

MP-based flood forecasts and improve the calibration of Noah-MP parameters, especially in regions where the model's performance is currently suboptimal. We also explore whether Noah-MP flood forecast performance can be competitive with the existing SAC-based forecasts utilized in the western U.S.

Our evaluation of Noah-MP-based flood forecasts is performed across multiple river basins that form a transect along the coastal western U.S. We evaluate the skill of the model in a reforecasting framework, in comparison with archived flood reforecasts in the CNRFC and NWRFC domains. Our investigation delves into and elaborates on three primary stages required for the application of Noah-MP to flood forecasting: (i) the selection of physical parameterizations, (ii) the calibration of Noah-MP parameters, and (iii) the generation and evaluation of flood reforecasts in comparison with archived operational forecasts from CNRFC and NWRFC. Section 2 introduces our study domains and methods. Section 3 explores the design of our experiments, followed by section 4 which assesses both flood reconstructions (simulations) and reforecasts. Section 5 discusses our findings and future pathways. Concluding remarks are presented in the final section, section 6.

### 3.2 Study region and model overview

#### 3.2.1 Study region

We focused on seven watersheds that form a transect along the U.S. Pacific Coast: the Green and the Upper Chehalis River basins in Washington State, the McKenzie River basin in Oregon, and the Smith River basin, the Van Duzen River basin, the Russian River basin, and the Carmel River basin, all in California (see Figure 3.1 and Table 3.1). The California rivers are all rain-dominated basins; the Upper Chehalis, Green and Mckenzie Rivers have modest contributions of snowmelt to flood runoff. Our choice of these river basins was informed by the availability of

long-term sub-daily streamflow observations and the presence of forecast points at which the two RFCs (NWRFC and CNRFC) routinely issue flood forecasts. Table 3.1 gives the drainage areas, stream gauge identifiers, and RFC forecast point designators for the river basins we study.

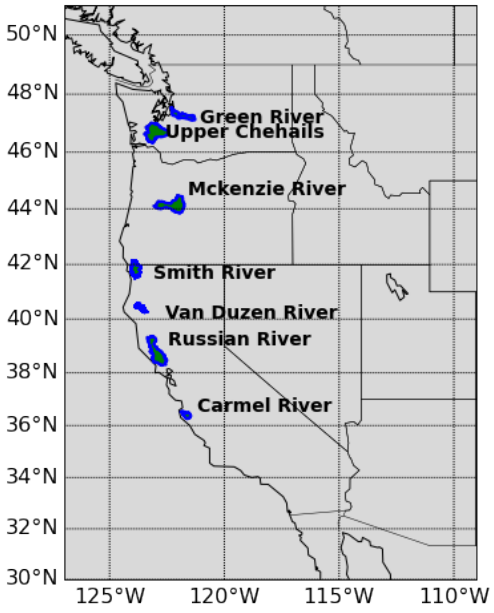



Figure 3.1 Map of study region including seven river basins: the Green and Upper Chehalis Rivers in Washington State, the Mckenzie River in Oregon, and the Smith, Van Duzen, Russian, and Carmel Rivers in California.

Table 3.1 Drainage area, observing sites and RFC sites for the study regions.

| River | Drainage Area ($km^2$) | Site Name | OBS Sites | RFC sites | States | QPE/QPF period |
|---|---|---|---|---|---|---|
| **Green River** | 570 | Howard Hanson Dam inflow | HHDW1 USACE | HHDW1 | WA | Oct 2017-Jan 2023 |
| **Upper Chehalis River** | 2320 | NR Grand Mound | 12027500 | CGMW1 | WA | Oct 2017-Jan 2023 |
| **McKenzie River** | 1528 | NR Walterville | 14163900 | MCZO3 | OR | Oct 2017-Jan 2023 |
| **Smith River** | 1590 | NR Crescent City | 11532500 | CREC1 | CA | Jun 2003-Sep 2020 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Van Duzen River** | 928 | NR Bridgeville | 11478500 | BRGC1 | CA | Jun 2003-Sep 2020 |
| **Russian River** | 3465 | Hacienda Bridge NR Guerneville | 11467000 | GUEC1 | CA | Jun 2003-Sep 2020 |
| **Carmel River** | 696 | A Robles Del Rio | 11143200 | RDRC1 | CA | Jun 2003-Sep 2020 |

### 3.2.2 Hydrological model and forcings overview

#### 3.2.2.1 Hydrological model

We employed Noah-MP in the WRF-Hydro framework (version 5.2.0) in an uncoupled mode that utilizes externally provided atmospheric forcings (essentially Noah-MP externally forced). In our off-line implementation, Noah-MP represents water and energy fluxes vertically within 6-km grid cells including surface runoff, soil water storage and drainage, evapotranspiration, snow melt and accumulation, and aquifer recharge (Niu et al. 2011). Surface runoff and subsurface drainage are routed laterally through surface and subsurface runoff modules as represented by a 300-m terrain grid, which captures changes in elevation and their effects on gravitational redistribution at the surface and in the subsurface, which eventually are directed into the channel network (Gochis et al. 2020). In our Noah-MP implementation, we opted to utilize a comparatively coarse Noah-MP grid and a finer routing grid. This approach balances the need to keep computational costs in check while still ensuring a sufficient level of detail in the stream network aspects of the simulation.

The model's land use data were aggregated from USGS 30-arc-s 24-land-use categories (USGS 2018), and the soil type similarly was aggregated from 30-arc-s hybrid State Soil Geographic Database (STATSGO) soil texture datasets (NCAR 2022). Noah-MP's initial land surface parameters, such as the vegetation conditions that control transpiration rates and soil properties that affect streamflow, were derived from these datasets using the WRF Preprocessing System (WPS) (WRF Preprocessing System Version 4.1: https://github.com/wrf-

model/WPS/archive/v4.1.tar.gz, accessed July 31, 2022). The hydrologic routing input files were generated via the WRF-Hydro GIS pre-processing tools(https://ral.ucar.edu/projects/wrf_hydro/pre-processing-tools, accessed July 31, 2022). Detailed instructions on preparing the input files can be found in the guidelines (WRF-Hydro Development Team, 2020)

Initially, we employed all options implemented in the NWM, including the Musk-Cunge-reach channel routing method as detailed in Gochis et al. (2020). Given our focus on flood streamflow, we paid particular attention to the runoff option in Noah-MP, which will be elaborated on further in the subsequent section.

#### 3.2.2.2 Meteorological forcings

Executing Noah-MP simulations requires spatially distributed meteorological forcings, including surface air temperature, specific humidity, surface pressure, wind speed, longwave and shortwave radiation, and precipitation. Notably, among these factors, precipitation is the most crucial in reforecasting flood flows. We used the same precipitation forcing as the RFCs so that we could better isolate the forecast skill of Noah-MP in comparison with RFC methods. Therefore, the precipitation we used for our baseline simulations, as well as our reforecasts, was QPE and QPF from CNRFC and NWRFC. The available time period for QPE and QPF is Jun 2003-Sep 2020 for CNRFC; and Oct 2017-Jan 2023 for NWRFC. The available time periods for each river basin are shown in Table 3.1. The lead time for CNRFC QPF is 6-72 hours before Sep 2012 and 144 hours thereafter. The initialization interval is once per day (12:00) before Oct 2010 for CNRFC and twice per day (12:00 and 18:00) after Oct 2010. For NWRFC, the lead time is 6-240 hours, and its initialization interval is once per day (12:00). Details are summarized in Table 3.2.

We interpolated the 4-km and 6-hourly QPE/QPF into 6-km and 3-hourly to run Noah-MP. All the forcings other than precipitation came from Pan et al. (2023), which is 1-km, hourly near real time over the conterminous U.S. based on North American Land Data Assimilation System (NLDAS) (Xia et al. 2012) and rescaled by Parameter-Elevation Regressions on Independent Slopes Model (PRISM; Daly et al. 2008). Here, we aggregated the forcings to 6-km and 3-hourly to save computational time. Prior to performing the baseline simulations, we conducted a spin-up simulation of 3 years using the Pan et al. (2023) forcings (including their precipitation) to remove initialization effects (largely in soil moisture and SWE) to the greatest extent possible.

Table 3.2 Lead time and initialization time of QPF from CNRFC and NWRFC

| **CNRFC QPF** | | **NWRFC QPF** | |
|---|---|---|---|
| **Lead time** | | | |
| Jun 2003-Sep 2012 | 6-72 hours | Oct 2017-Jan 2023 | 6-240 hour |
| Oct 2012-Sep 2020 | 6-144 hours | | |
| **Initialization time** | | | |
| Jun 2003-Oct 2010 | 12:00 | Oct 2017-Jan 2023 | 12:00 |
| Nov 2010 -Sep 2020 | 12:00&18:00 | | |

## 3.3 Experimental design

### 3.3.1 Noah-MP parameterization

The first step in configuring Noah-MP is to choose a suitable set of LSM parameterizations. Because we are only interested in floods, our focus was largely on parameterizations that impact high flows, specifically runoff options. Noah-MP has four options for runoff parameterizations (rnf) (Table 3.3); we assessed their impact on high flows as a first step. For physical parameterization other than runoff, we selected the same parameterizations as in the NWM configuration of Noah-MP (see Gochis et al. 2020 for details) including the channel routing scheme.

We ran Noah-MP simulations with the four different runoff options using the default (NWM) land surface parameters and compared the Kling-Gupta efficiency (KGE) (Gupta et al. 2009) of flood events. KGE is a widely used performance measure because of its advantages in orthogonally considering bias, correlation and variability (Knoben et al. 2019). KGE = 1 indicates perfect agreement between simulations and observations; according to Knoben et al. (2019) KGE values greater than -0.41 indicate that the model improves upon the mean flow benchmark (i.e., forecasting the mean).

Table 3.3 Noah-MP Runoff options selected for this study. The ID numbers refer to the values that can be specified in the model input namelist file.

| **Runoff and groundwater options** | **Descriptions** |
|---|---|
| **rnf 1** | TOPMODEL-based runoff scheme with the simple groundwater (hereafter SIMGM) (Niu et al. 2007). |
| **rnf2** | Simple TOPMODEL-based runoff scheme with an equilibrium water table (Niu et al. 2005) (hereafter SIMTOP). |
| **rnf3** | Infiltration-excess-based surface runoff scheme with gravitational free-drainage subsurface runoff scheme (Schaake et al. 1996) |
| **rnf4** | BATS runoff scheme, which parameterized surface runoff as a 4th power function of the top 2 m soil wetness (degree of saturation) and subsurface runoff as gravitational free drainage (Yang and Dickinson 1996). |

We used the peaks-over-threshold (POT) method (e.g., Lang et al. 1999) to identify floods. We set thresholds at each stream gauge that resulted in three extreme events per year on average, which we denote as POT3. Our results (Figure 3.2) show that when using the default free drainage option (option 3), the baseline flood simulation achieved the best performance across the study basins (and in fact was substantially better than the other options). This runoff option is also used in NWM and Mascaro et al. (2021).

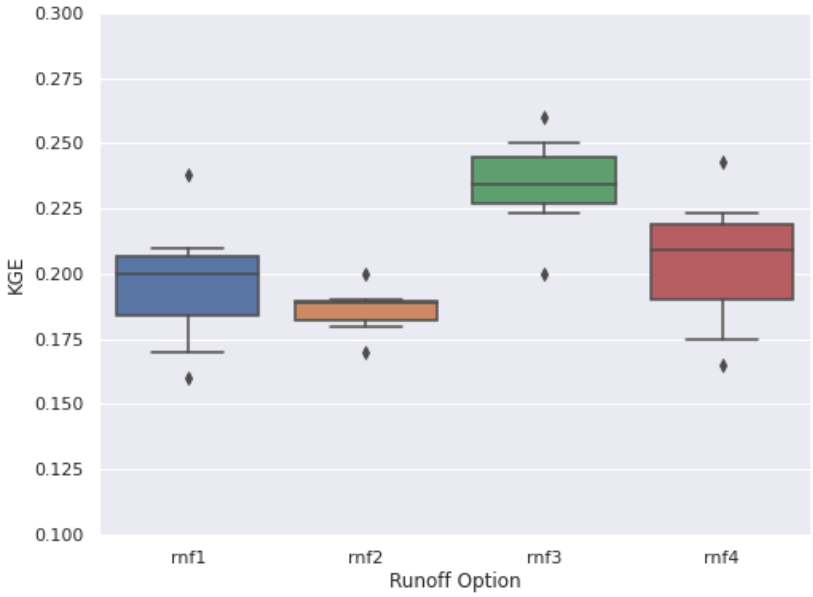



Figure 3.2 Boxplots of POT3 floods' KGE of the seven study basins when using default soil and runoff parameters and different Noah-MP runoff options.

### 3.3.2 Land surface parameter calibration

#### 3.3.2.1 Calibrated parameters

After selecting the best domain-wide parameterizations for runoff generation, we calibrated the spatially varying LSM parameters to further improve flood performance. Based on available computational resources and the findings from previous studies (Holtzman et al. 2020; Sun et al. 2020; Sofokleous et al. 2022; Quenum et al. 2022; Mascaro et al. 2023; Lahmers et al. 2019& 2021; Bass et al. 2023), we focused on five parameters. They are : saturated hydraulic conductivity (DKSAT), Saturated soil moisture (SMCMAX) (i.e. porosity), the pore-size distribution index (BEXP), the coefficient governing deep drainage (SLOPE) and surface runoff parameter REFKDT. Details of the processes controlled by these parameters and feasible ranges are given in Table 3.4. While the default values for DKSAT, SMCMAX, and BEXP in Noah-MP are based on STATSGO soil types, the values assigned to these parameters typically are highly uncertain. We considered a

range of physically realistic adjustments to each of the parameters (Table 3.4, column 5) based on values used in previous studies (Cai et al. 2014; Mendoza et al. 2015; Gochis et al. 2019; Hussein 2020).

We adjusted parameters values by scalar multipliers following the methods used by NCAR-RAL to calibrate the NWM, as reported in Lahmers et al. (2019&2021). This ensured that the original model parameters are physically consistent with a priori catchment properties (e.g., Gupta et al. 2008, 2009).

Table 3.4 Noah-MP land surface parameters selected for calibration.

| Parameter | Description | Unit | Main control on hydrological response | Range (source) |
|---|---|---|---|---|
| **Soil parameters** | | | | |
| **DKSAT** | Saturated hydraulic conductivity | m/s | Infiltration | $2 \times 10^{-9} to\ 0.07$(Cai et al. 2014) |
| **SMCMAX** | Saturated soil moisture | $\frac{m^3}{m^3}$ | Infiltration and soil evaporation | 0.1 to 0.71 (Cai et al. 2014) |
| **BEXP** | Pore-size distribution index | unitless | Infiltration | 1.12 to 22 (Cai et al. 2014; Gochis et al. 2019) |
| **Runoff parameters** | | | | |
| **SLOPE** | Linear scaling of "openness" of bottom drainage boundary | unitless | Aquifer recharge | 0.1-1 (Lahmers et al. 2021) |
| **REFKDT** | Parameter in surface runoff | unitless | Partitioning of total runoff into surface and subsurface runoff | 0.1-10 (Lahmers et al. 2021) |

#### 3.3.2.2 Calibration method

Consistent with the calibration approaches used for the NWM as reported by Feng et al. (2019) and Gochis et al. (2019), we used the Dynamically Dimensioned Search (DDS) algorithm (Tolson and Shoemaker, 2007). Unlike the widely adopted Shuffled Complex Evolution (SCE) function (Duan et al. 1992), which typically requires order 1,000 iterations to achieve an optimal solution, the DDS algorithm is more efficient, achieving near-optimal parameter sets within a significantly reduced range of about 100-500 iterations (see Lespinas et al. 2017). Our tests indicated that the DDS objective function (we used KGE of POT3) improved substantially between 50 and 200 DDS iterations, but with diminishing returns beyond 200. Therefore, we opted to use 250 iterations of DDS in all our calibrations.

As indicated above, we used spin-up of three years, and used the Pan et al. (2023) forcings (including precipitation) prior to the earliest year of available QPE. The model's states (soil moisture and SWE) following this period served as the starting 'warm' state for calibration. We carried out calibrations for the Green, Upper Chehalis, and McKenzie Rivers for the period October 2017 to January 2023, and for the Russian, Smith, Van Duzen, and Carmel Rivers from October 2003 to September 2020. We tested a variety of alternative objective functions, including (a) KGE of 6-hour streamflow higher than 85 percentiles, (b) KGE of 6-hour streamflow higher than 90 percentiles, (c) KGE of 6-hour streamflow concentrated on a nine-day period for selected POT3 flood events (four days preceding the flood peak date, the day of the flood peak and four days following the flood peak date, 36 flow values for each flood event), (d) KGE of 6-hour streamflow concentrated on a five-day period for selected POT3 flood events (two days preceding the flood peak date, the day of the flood peak and two days following the flood peak date, 20 flow values for each flood event), (f) KGE of 6-hour streamflow concentrated on a three-day period for

selected POT3 flood events (one day preceding the flood peak date, the day of the flood peak and one day following the flood peak date, 12 flow values for each flood event). After comparison, we found KGE of 6-hour streamflow concentrated on a three-day period for POT3 flood events worked the best in capturing the floods (by comparing KGE and flood hydrographs). Thus, our objective function used in all calibrations is the KGE of 6-hour QPE-forced Noah-MP streamflow concentrated on a three-day period for POT3 flood events compared with the 6-hour observation. Our focus was solely on flood periods; hence we used all events in the POT3 data set for each river basin. Figure B1 gives the events used for calibration for each of the river basins.

### 3.3.3 Noah-MP reforecasts

After calibration, we ran Noah-MP with the calibrated land surface parameters and QPE forcings and archived the model states for each timestep for all the flood events. These states (soil moisture and SWE) served as the starting "warm" state. For each flood event, and for each available QPF that initiated from 6 to 120 hours before the observed flood peak time, we ran Noah-MP starting from the "warm state" of the QPF initiating point and ran for a duration of the QPF lead time (72 hours to 240 hours depending on the basin and time period, refer to Table 3.2 for lead time details). In these runs, all the other forcings were the same as in the calibration runs, except that the QPE precipitation is replaced to QPF. We archived the flood forecasts from these runs and evaluated the skills in the following section.

## 3.4 Results

### 3.4.1 Calibration

Scatterplots of POT3 flood events produced by calibrated Noah-MP simulations for all seven river basins are shown in Figure 3.3. After calibration, the KGE of POT3 flood streamflow increased from about 0.2~0.3 to about 0.7~0.9. The calibrated POT3 KGE values for the Green,

Chehalis, McKenzie, Smith, Van Duzen, Russian and Carmel Rivers are 0.85, 0.70, 0.76, 0.89, 0.76, 0.85 and 0.78 respectively. The flood flow simulation skills are high compared with previous studies (Mascaro et al (2023) report an NSE of 0.59 for the Noah-MP simulated and observed streamflow time series for the period 2008 to 2011; Lin et al. (2018) evaluated the Noah-MP streamflow simulation skill against 271 USGS gauges over Texas and their highest NSE was ~0.7).

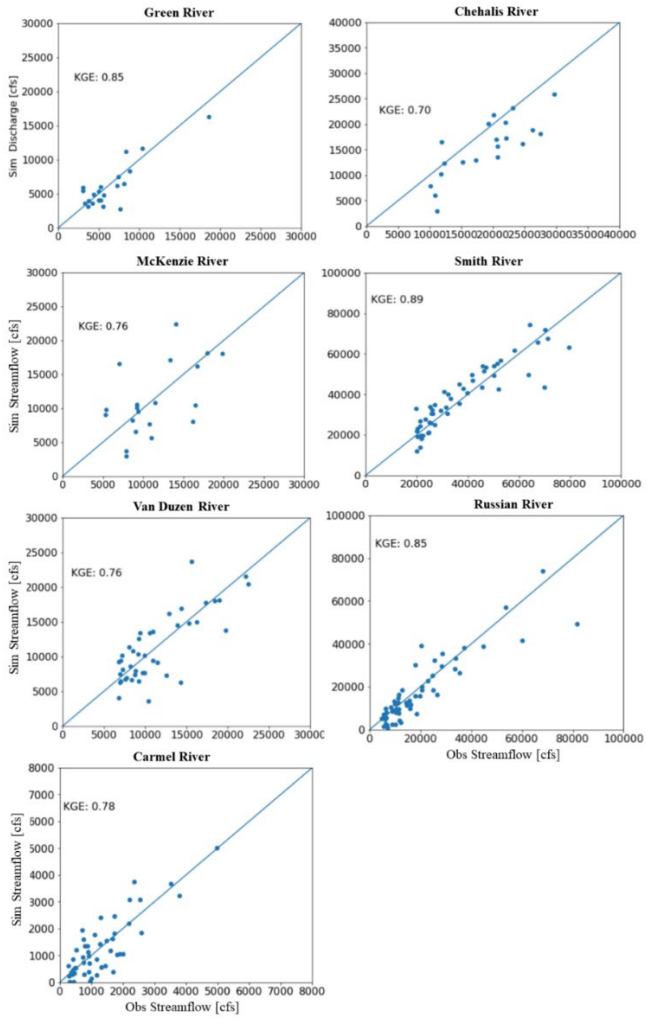



Figure 3.3 Scatter plot of simulated flood streamflow and observed flood streamflow in seven river basins.

### 3.4.2 Reforecasts

We compared our flood simulation and reforecast results generated by Noah-MP with the archived forecasts provided by RFC for all seven river basins.

**(a) POT3 flood events**

To get a better understanding of the overall performance of the Noah-MP reforecasts, we first summarized the skill for all the POT3 flood events within the study period for which there were both Noah-MP reforecasts and RFC forecasts. Figure B1 shows the 6-hour time series of these flood events in our study basins. The peak time and magnitude are also indicated in the figure. The floods we considered occurred between 2003 to 2020 and had daily peak magnitudes as small as ~750 cfs (Carmel River, Feb 26, 2010, with return period of less than a year) and as large as ~80,000 cfs (Smith River, Dec 28, 2008, with return period of ten years). Figure B2 shows the distribution (median and interquartiles) of the relative differences of floods peak streamflow in all the river basins. Here relative difference is defined as

$$\mathrm{Peak}_{\mathrm{Diff}_{\mathrm{Rela}}} = (\mathrm{Peak}_{\mathrm{forecast}} - \mathrm{Peak}_{\mathrm{Obs}})/(\mathrm{Peak}_{\mathrm{Obs}}) \quad (11)$$

Where $\mathrm{Peak}_{\mathrm{Diff}_{\mathrm{Rela}}}$ is relative flood streamflow difference, $\mathrm{Peak}_{\mathrm{forecast}}$ is forecasted peak streamflow, $\mathrm{Peak}_{\mathrm{Obs}}$ is observed peak streamflow. We limited our performance evaluation to lead times of up to 120 hours, due to our primary interest in forecast skill for up to five days lead (for longer leads, skill arguably is dominated by QPF, rather than hydrologic prediction skill). Moreover, we have fewer available reforecast beyond 120 hours lead time.

When we analyzed the performance of Noah-MP in comparison with RFC, it became evident that Noah-MP is roughly comparable to RFC in terms of forecast accuracy in the northern basins (Green, Chehalis, McKenzie), but falls short in the southern basins (see Figure B2). Consequently, we divided our domain into a northern region, which includes the Green, Chehalis, and McKenzie Rivers, and a southern region, which includes the Smith, Van Duzen, Russian, and Carmel Rivers.

In a comparison of skills between the northern and southern basins, it's clear that Noah-MP exhibits a more competitive performance in the northern basins (as shown in Figure 3.4). Here, the

Noah-MP relative median peak differences are approximately within ±0.1 and the Inter-Quartile Range (IQR) are within the range of 0.15-0.45 for forecasts less than 60 hours prior to the observed peak. The skills of Noah-MP are comparative to NWRFC with 60 hours lead time where the NWRFC relative median peak differences are less than -0.15 (in absolute value) and the IQR are within 0.18 - 0.43. However, the effectiveness of both Noah-MP and NWRFC rapidly declines beyond this point, with the worst Noah-MP median skill reaching ~-0.3 at a lead time of 120 hours and the biggest IQR reaching 0.82 at lead time of 114 hours. For NWRFC, the worst median skill is -0.46 at lead time of 108 hours and the biggest IQR is 0.54 at lead time of 114 hours. Noah-MP's skill is similar to NWRFC for forecasts with lead less than 48 hours, and are more accurate for longer leads in terms of median relative peak difference. While the first quartile (Q1) for Noah-MP and NWRFC are similar, the third quartile (Q3) is higher in Noah-MP than NWRFC, resulting in larger variability (greater IQR) in Noah-MP.

Conversely, in the southern basins, Noah-MP's performance is distinctly inferior to that of the RFC forecasts. The Noah-MP median relative peak difference here typically ranges from -0.1 to -0.23, with an IQR between 0.28 and 0.53 when the forecasted hours are less than 60 hours before the observed peak. This accuracy is consistently outperformed by CNRFC, which has a steadier median skill around ±0.1 and an IQR ranging from 0.18 to 0.38. For lead times longer than 60 hours, a sharper decline in skills is observed for both Noah-MP and CNRFC. Yet, throughout these durations, Noah-MP's median skill remains lower than CNRFC's. Overall, while both models often underestimate flood peaks, Noah-MP does so more substantially, and this tendency to underestimate grows with increased lead time.

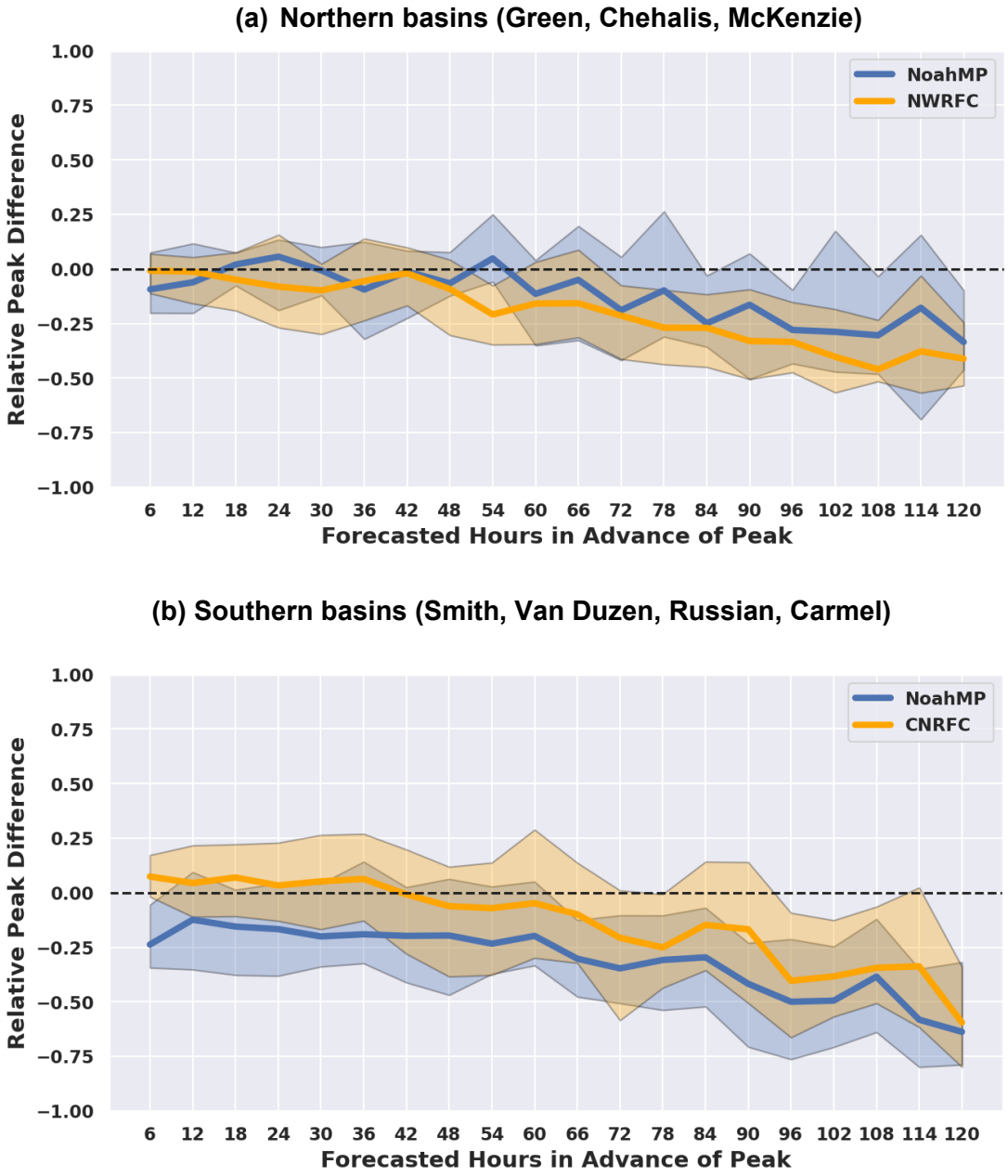



Figure 3.4 Median and interquartile range of the relative differences of POT3 floods peak streamflow of Noah-MP reforecasts and RFC forecasts against as a function of forecasted hours in advance of the observed peak (or lead time) in (a) Northern basins (Green, Chehalis, McKenzie); (b) Southern basins (Smith, Van Duzen, Russian, Carmel).

We should note that the events at each lead time used in assessing the skill of Noah-MP and RFC may differ. This disparity arises because the available QPF utilized to generate Noah-MP reforecasts and the flood forecast issuance numbers accessed from RFC are not always equivalent. Furthermore, the issuance timings and frequencies of flood forecasts from the RFCs fluctuate across different flood events. The QPFs obtained from the RFCs also exhibit variable issuance timings and frequencies for some river basins, generally adhering to a fixed schedule of either 12:00 or 18:00 daily. Given that flood peaks can occur at any time between 00:00 to 24:00, the floods analyzed for each lead time for both Noah-MP and RFC may vary. We present the number of floods used for skill calculations in Figure B4.

In addition to flood peak magnitudes, we also examined the peak time forecast skill. Here time difference is defined as

$$\mathrm{Time}_{\mathrm{Diff}} = \mathrm{Time}_{\mathrm{Peak}_{\mathrm{forecast}}} - \mathrm{Time}_{\mathrm{Peak}_{\mathrm{Obs}}} \tag{12}$$

Where $\mathrm{Time}_{\mathrm{Diff}}$ is the peak time difference in simulation and observation, $\mathrm{Time}_{\mathrm{Peak}_{\mathrm{forecast}}}$ is the forecasted peak time, $\mathrm{Time}_{\mathrm{Max}_{\mathrm{Obs}}}$ is the observed peak time.

Figure 3.5 indicates that Noah-MP performs competitively in predicting the flood peak time in the northern basins, maintaining a median peak time difference of approximately 0 to -6 hours and the IQR spans from 1 to 24 hours for most lead times. For NWRFC, the median difference in peak times varies between 6 to 12 hours, and the IQR lies between 12 to 36 hours. As the lead time extends, the variability of the peak time difference correspondingly increases for both Noah-MP and NWRFC. It is notable that while Noah-MP tends to forecast earlier peaks than observed, NWRFC generally predicts later. In the southern basins, both Noah-MP and CNRFC display high competence with a median skill around 6 hours and an IQR ranging from 6 to 12 hours when the lead time is under 48 hours. However, Noah-MP's accuracy noticeably diminishes beyond this

point, trending toward a substantial early bias in the flood peak time at long leads, notably with a lower median skill and higher IQR (a median peak time difference of 57 hours and an IQR of 69

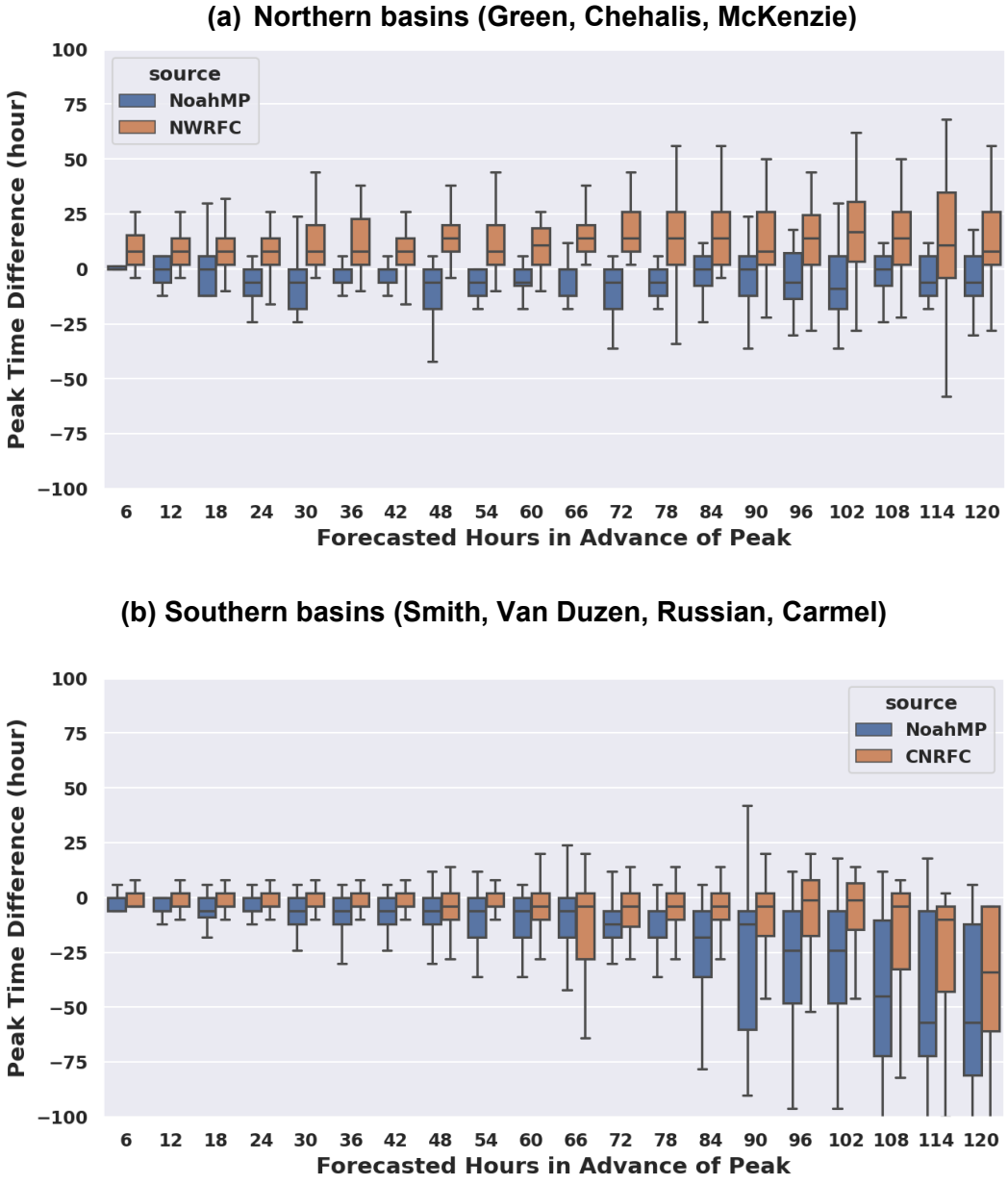



Figure 3.5 Boxplots of POT3 flood peak time differences for Noah-MP reforecasts and RFC forecasts vs the forecasted hours in advance of observed peak (or lead time) in (a) Northern basins (Green, Chehalis, McKenzie); (b) Southern basins (Smith, Van Duzen, Russian, Carmel).

hours for a 120-hour lead time). In contrast, CNRFC maintains relatively low bias in peak times, albeit with a slight tendency towards forecasting an earlier peak when the lead time exceeds 66

hours and its IQR also increases (a median peak time difference of 34 hours and an IQR of 57 hours for a 120-hour lead time).

**(b) Major floods**

Larger floods typically result in more significant damage and widespread consequences. Consequently, there is often heightened interest in understanding and forecasting the most severe flood events to reduce their impact. In line with this, we analyzed the peak intensity and timing, similar to our previous evaluations, but specifically for the top three flood incidents during our research span in each river basin. Upon assessing the forecasting skills for magnitudes, the results reflected the trends seen with POT3 floods. The forecasting capability of Noah-MP was competitive in the northern basins and fell short in the southern basins when compared with RFC concerning the median relative peak disparity (Figure 3.6). Additionally, the relative peak difference under Noah-MP exhibited greater variability (represented by IQR) in both northern and southern basins compared to RFC.

Upon evaluating the peak timing prediction capability, it showed that Noah-MP lags behind RFC in both the northern and southern basins, in terms of both the median peak time discrepancy and IQR (as shown in Figure 3.7). Noah-MP tended to forecast earlier peak for both northern and southern basins, while NWRFC showed a slightly late bias and CNRFC showed an early bias.

**(c) Largest floods**

Our analysis further revolves around representative largest flood events in each of these basins, as depicted in Figure 3.8. For each river basin, we selected the largest flood event that occurred within the study period, and for which both RFC forecast and Noah-MP reforecast data were available. These flood events were identified based on their return periods, calculated using

the Generalized Extreme Value (GEV) distribution (Jenkinson 1955). The specific return period values have been annotated in the respective subtitles.

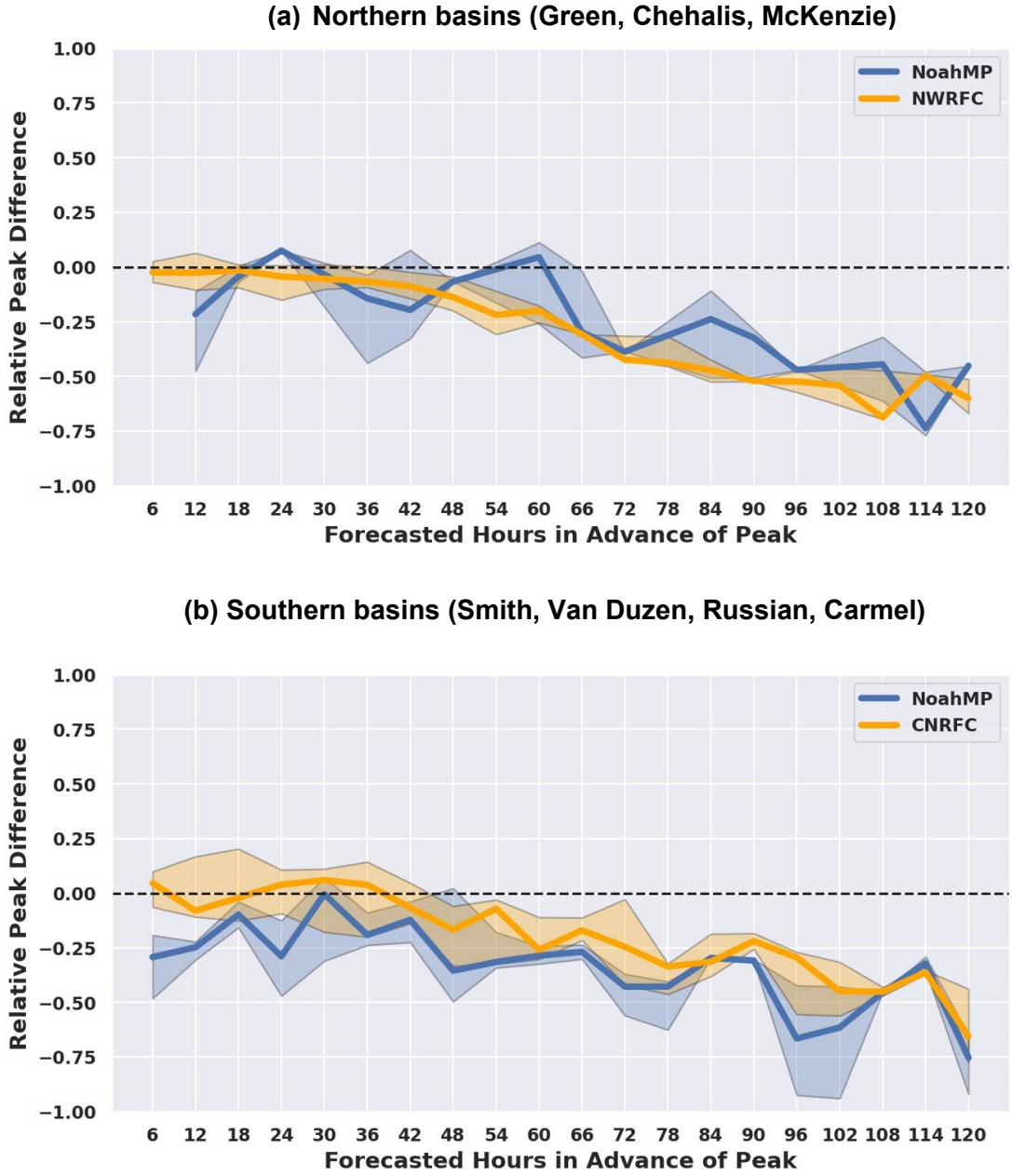



Figure 3.6 Median and interquartile range of the relative differences of largest three floods peak streamflow of Noah-MP reforecasts and RFC forecasts against as a function of forecasted hours in advance of the observed peak (or lead time) in (a) Northern basins (Green, Chehalis, McKenzie); (b) Southern basins (Smith, Van Duzen, Russian, Carmel). Please note that no Noah-MP forecasts are available for a 6-hour lead time.

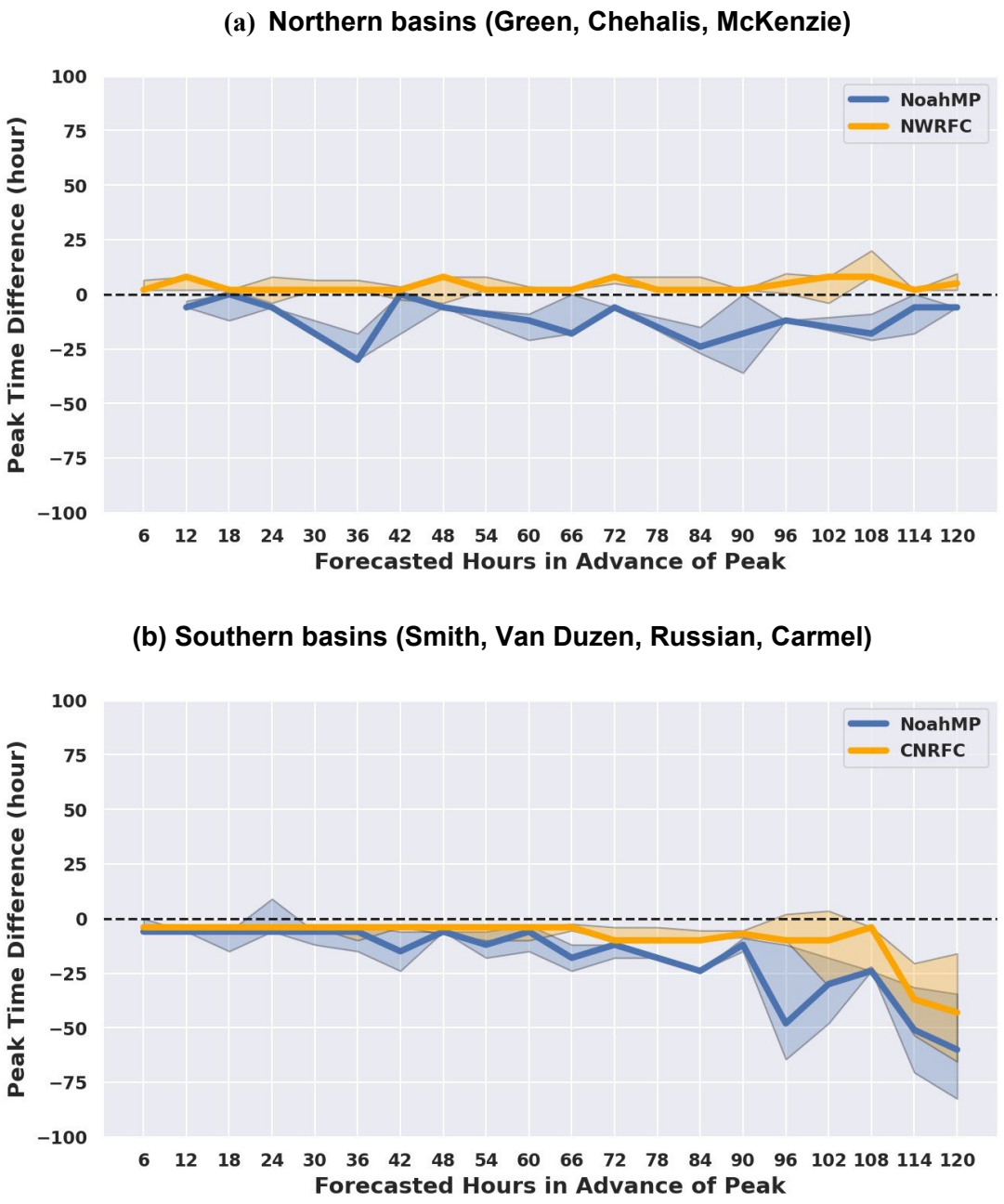



Figure 3.7 Median and interquartile range of the peak time difference of largest three floods peak streamflow of Noah-MP reforecasts and RFC forecasts against as a function of forecasted hours in advance of the observed peak (or lead time) in (a) Northern basins (Green, Chehalis, McKenzie); (b) Southern basins (Smith, Van Duzen, Russian, Carmel). Please note that no Noah-MP forecasts are available for 6-hour lead time.

A close examination of these comparisons revealed a few noteworthy patterns. For instance, Noah-MP was proficient in predicting both the magnitude and timing of the Feb 7, 2020 flood

(with a return period of 12 years) in the Green River, making it competitive with the NWRFC forecast. For the Jan 14, 2021 flood (with a return period of four years) in the Chehalis River, Noah-MP predicted a magnitude similar to observations but with an earlier peak, while NWRFC accurately predicted the timing of the peak, albeit with a slight underestimation of the flood's magnitude. In the flood that occurred on May 7, 2022, in the McKenzie River (with a return period of six years), Noah-MP forecasted the flood peak slightly earlier than observed, with a slight overestimation in flood magnitude. In contrast, NWRFC accurately predicted both the timing and magnitude of the flood. For the flood of Dec 18, 2008, in the Smith River (with a return period of ten years), both Noah-MP and CNRCF forecasts under-predicted the observed flood magnitude. Dual peaks were observed in this event, however, both Noah-MP and CNRCF only anticipated a single peak and their predicted peak time occurred between the two observed peaks. During the flood on Dec 28, 2005, in the Van Duzen River (with a return period of 12 years), Noah-MP predicted an earlier peak than observed, though the predicted magnitude was accurate. CNRFC, on the other hand, accurately predicted the timing, but the predicted magnitude was underestimated. In the flood event on Jan 1, 2006, in the Russian River (with a return period of 36 years), both Noah-MP and CNRFC accurately predicted the magnitude, but Noah-MP predicted an earlier peak than observed. For the flood on Jan 8, 2017, in the Carmel River (with a return period of 10 years), both Noah-MP and CNRFC predicted an earlier peak than observed. However, CNRFC accurately captured the magnitude of the flood, while Noah-MP underestimated it. To summarize, Noah-MP mostly produced comparable magnitudes with RFC forecasts for these largest floods, however had a tendency to predict earlier peaks than observed. The model's performance was comparable to RFC in certain events but fell short in others.

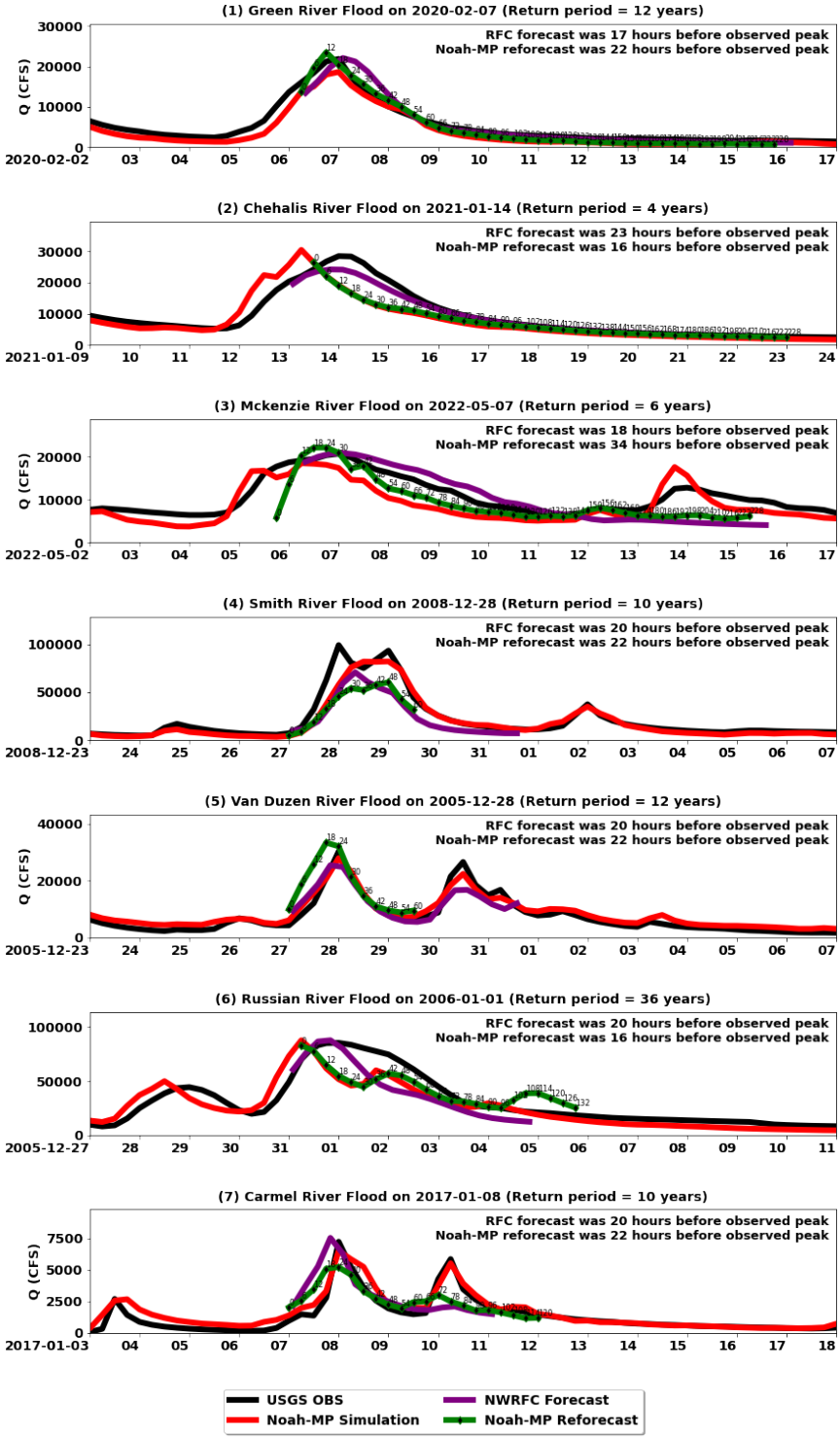



Figure 3.8 USGS observation, Noah-MP flood simulation and reforecasts and RFC archived forecasts for representative large floods in the study basins. The flood events and their return period are noted in the titles of each subplot. Observations are black, Noah-MP simulations are red, and the most recent available RFC forecast before the time of peak is purple; the Noah-MP reforecast initiated one-day before the peak

time are green. The initiation time of both RFC forecasts and Noah-MP reforecasts are indicated in the upper right corner of each subplot. The lead time steps of Noah-MP reforecasts are indicated with black dots and annotated with numbers beside.

## 3.5 Discussion

### 3.5.1 Model error

In the previous section, we demonstrated how the selection of the Noah-MP runoff option and calibration substantially enhanced the KGE of POT3 flood simulations (from approximately 0.2 ~ 0.3 to around 0.7 ~ 0.9) in the study basins. We found that the flood forecast skill of Noah-MP rivals that of RFC in northern basins, while it is inferior in southern basins. The northern basins typically receive most of their precipitation in the fall and winter, whereas the largest rainfall events in the southern basins tend to be in winter (see Figure B5). Among the southern basins, Noah-MP shows the best flood magnitude prediction skill in the Russian River basin, which has the largest drainage area of all of our study basins (refer to Figure B2 and Table 3.1). The Van Duzen (drainage area 928 $km^2$) and Carmel (drainage area 696 $km^2$) Rivers are the smallest study basins, and also showed the lowest flood magnitude prediction skill. These findings align with previous studies, which indicated that the NWM performs suboptimally in drier and smaller basins. (Hansen et al. 2019; Rojas et al. 2019; Lin et al. 2018).

Predicting floods in semi-arid regions like the southwestern U.S. poses multiple challenges rather to coastal rivers in more humid regions, like our northern basins. This is partly due to the complexities in accurately determining antecedent soil moisture levels and the sometimes brief and variable nature of severe precipitation events (Lahmers et al. 2019). Moreover, the south tends to experience unusually large variations in annual precipitation and streamflow totals relative to the north. Such variations can be attributed mainly to the notably few rainy days per year that account for most of its annual precipitation (Dettinger et al. 2011). These factors cumulatively

compound the difficulty in making accurate predictions in the southwestern U.S.. Future research could enhance Noah-MP's initial conditions by exploiting additional information, such as the space-time (and depth) variability of soil moisture.

In our current model configuration, the Noah-MP runoff parameterization scheme is infiltration-excess-based surface runoff scheme with a gravitational free-drainage subsurface runoff scheme (Niu et al. 2011). In arid climates, infiltration excess is more prominent, especially where the most intense storms usually have higher precipitation rates; while in forested watersheds, saturation excess mostly dominated. We utilized a uniform model configuration where we assumed the same model physics for all basins, and our choice of best model physics was under that assumption. This may not have been optimal since the most effective configuration may differ between wet and dry basins. Future work could use different optimal configurations for different basins. Additionally, our main emphasis centered on calibrating the magnitude of flood peaks, without considering timing during the calibration phase. This method can be enhanced by including the timing of peak flows in the calibration objective function. Moreover, we can explore deeper into the runoff production process in the future.

In some basins, like the Russian River, may have reservoir impacts. We tested a basic level pool reservoir method (Brunner and Ras 2008) for this basin, but it did not significantly improve either the forecast skill of flood peak or timing (results not shown here), so we opted not to incorporate reservoir adjustments in our study. Additionally, this study doesn't account for other human-induced influences, which could potentially result in our forecasts being less accurate than those from RFC. Future investigations could improve upon this by incorporating more pertinent data.

### 3.5.2 Precipitation forcing errors

A plausible contributor to forecast errors could be inaccuracies in the predicted precipitation (i.e., QPF). To examine this, we calculated the relative difference between QPF (precipitation forecast) and QPE (gridded observations) over an aggregation time period, defined as:

$$PREC_{\mathrm{T}_{\mathrm{diff}}} = (QPF_T - \mathrm{QPE}_T)/(\mathrm{QPE}_{\mathrm{T}}) \quad (13)$$

Here the time period T is the time over which the precipitation is aggregated (e.g., the number of hours in advance of the flood peak). We assessed T values up to 24 hours prior to a flood peak. In cases where the QPF was issued less than 24 hours before the flood peak, we only accounted for the hours available from the time the QPF was issued. If the QPF was issued more than 24 hours before the flood peak, we set T to 24 hours. This pattern is depicted in Figure 3.9(c). We identified general underestimation in the QPF, and this downward bias increased with longer lead times in both northern and southern basins.

When comparing the QPF skills between the northern and southern basins, we noted that the skill in the southern region was inferior to that in the northern basins (quite likely associated at least in part with higher precipitation variability in the southern basins). This pattern generally mirrored the performance of the Noah-MP flood predictions in these respective basins. This may potentially account for the bias observed in the Noah-MP flood reforecasts illustrated in Figure 3.4. It's also worth mentioning that the difference in precipitation timing can be greater at a 6-hour lead time and lower at 12- and 18-hour lead times, as shown in Figure 3.9(a). This is due to the 6-hour lead time only incorporating precipitation accumulated over a 6-hour period. The forecasted precipitation might be missed at the 6-hour lead time but is caught at the 12-hour or longer lead

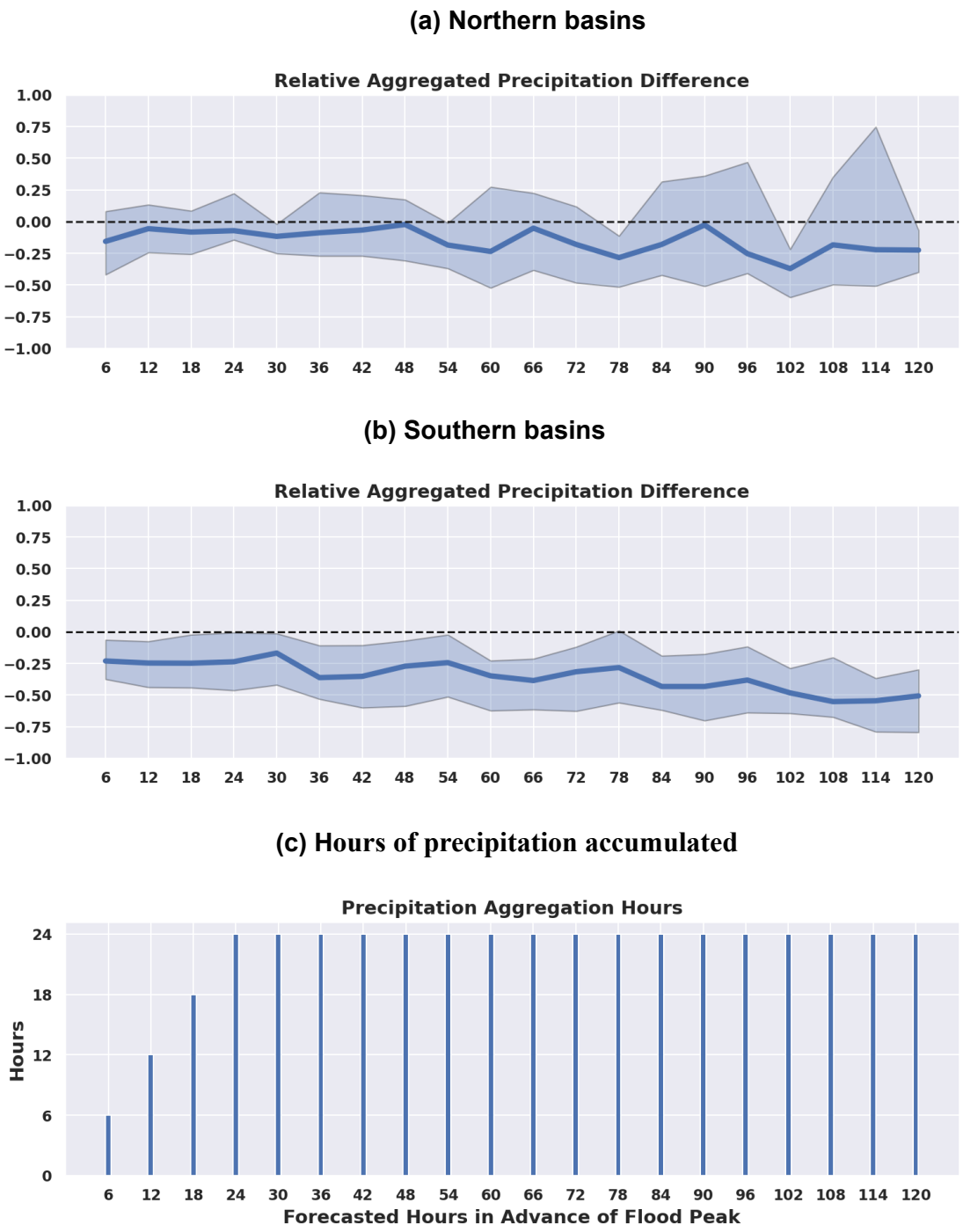



Figure 3.9 Relative difference of the aggregated precipitation reforecasts against the forecasted hours in advance of peak (or lead time) in (a) Northern basins (Green, Chehalis, McKenzie); (b) Southern basins (Smith, Van Duzen, Russian, Carmel). (c) The bottom subplot schematically shows the hours of precipitation aggregated when the forecasted hours in advance of peak differ.

times. However, this variability in QPF skills cannot account for why the CNRFC's performance (for southern basins) is not inferior to the NWRFC's (for northern ones). While we do not currently understand how RFCs handle QPF bias, there is potential for enhancing Noah-MP performance if we adopt similar strategies in the future.

### 3.6 Conclusion

The major contributions of this study are: a) we improved Noah-MP's flood simulation skill in coastal western U.S. river basins through the appropriate choice of model physics of free drainage and the implementation of a global optimal calibration approach. This process improved flood simulation skills from a range of KGE 0.2 ~ 0.3 to roughly 0.7 ~ 0.9 in the targeted basins; b) with the aid of calibrated parameters and QPF forcings, we assembled Noah-MP flood reforecasts and compared their forecast skills with those of RFC archived forecasts for a transect of seven river basins along the coastal Western U.S.

Our reforecast/forecast evaluation showed that:

1) For POT 3 flood magnitude forecasts, both Noah-MP and RFC show high skill within a 60-hour lead time, but their accuracies rapidly decline thereafter. In northern basins, Noah-MP exhibits comparative performance to NWRFC considering both bias and variability, for periods up to 60 hours before the observed peak. Conversely, in the southern basins, Noah-MP underperforms compared to RFC. While both models often underestimate flood peaks, Noah-MP does so more substantially, and this tendency to underestimate grows with increased lead time.
2) For POT3 floods peak time forecast in the northern basins, Noah-MP shows competitive predictions regarding both bias and variability. As lead time grows, the variability in peak timing predictions also increase for both Noah-MP and NWRFC, with Noah-MP

typically predicting earlier peaks and NWRFC later ones than observed. In the southern basins, while both Noah-MP and CNRFC perform well within a 48-hour lead time, Noah-MP's accuracy noticeably diminishes beyond that, showing a pronounced early bias with more variability than NWRFC.

3) For major floods (top three floods in each basin) magnitude forecast, Noah-MP is competitive with RFC in the northern basins but falls short in the southern ones. Across all basins, Noah-MP displayed higher variability than RFCs.
4) For major floods peak time forecast, Noah-MP consistently underperformed compared to RFCs across all study basins, both in timing accuracy and variability. Specifically, Noah-MP typically predicted earlier peaks, while NWRFC had a minor early bias and CNRFC leaned towards a later bias.
5) For the largest floods, Noah-MP mostly produces comparable magnitudes with RFC forecasts, however has a tendency to predict earlier peaks than observed. Noah-MP's performance is comparable to RFC in certain events but falls short in others.

In our study, we noted that Noah-MP tends to underperform in drier basins, which aligns with prior research findings. This underperformance is particularly evident in the southern region, where there are notable fluctuations in annual precipitation and streamflow compared to the northern areas. These significant variations in the south compound the challenges of making predictions. Additionally, we observed that forecast errors could be attributed to inaccuracies in the predicted precipitation (QPF). QPF predictions are less precise in the southern region than in the north, which mirrors the trends we noted with Noah-MP. Currently, the model operates on a uniform configuration, assuming consistent model physics for all basins. The chosen physics might not be ideal since the most effective setup could vary between wet and dry basins. Future studies

might benefit from employing basin-specific configurations and enhancing calibration by factoring in both peak flow timings and magnitudes.

Overall, our research highlights the potential of Noah-MP for flood forecasting, particularly in the northern basins, provided that suitable parameter selection and calibration are employed while more improvements should be done in southern drier basins. This could make a significant contribution to flood forecasting in the U.S., considering Noah-MP forms the hydrological core of the NWM.

**Acknowledgements**

We would like to thank Peter Fickenscher at NOAA for his sharing of the CNRFC flood forecasts. We thank Stephen King at NOAA for his sharing of the NWRFC flood forecasts. This work used the COMET supercomputer at UCSD.

**References**

Agnihotri, J. and P. Coulibaly, 2020: Evaluation of Snowmelt Estimation Techniques for Enhanced Spring Peak Flow Prediction. Water, 12(5), p.1290.

Bass, B., S. Rahimi, N. Goldenson, A. Hall, J. Norris, and Z.J. Lebow, 2023: Achieving Realistic Runoff in the Western United States with a Land Surface Model Forced by Dynamically Downscaled Meteorology. Journal of Hydrometeorology 24: 269-283.

Brunner, G. and H. RAS, 2008: River analysis system hydraulic reference manual. Do Defense, Davis.

Burnash, R., and R. Ferral, 1973: A Generalized Streamflow Simulation System. U.S. Department of Commerce, National Weather Service, and State of California.

Cai, X., Z.-L. Yang, C. H. David, G.-Y. Niu, and M. Rodell, 2014: Hydrological evaluation of the Noah-MP land surface model for the Mississippi River Basin. J. Geophys. Res. Atmos., 119, 23–38, https://doi.org/10.1002/2013JD020792.

Dettinger, M.D., F.M. Ralph, T. Das, P.J. Neiman, and D.R. Cayan, 2011: Atmospheric rivers, floods and the water resources of California. Water, 3(2), 445-478.

Dickinson, R. E., A. Henderson-Sellers, & P. J. Kennedy, 1993: Biosphere–Atmosphere Transfer Scheme (BATS) version 1e as coupled to the NCAR Community Climate Model. NCAR Tech. Note TN383+STR, NCAR.

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

Duan, Q., and Coauthors, 2006: Model parameter estimation experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. J. Hydrol., 320, 3–17, https://doi.org/10.1016/j.jhydrol.2005. 07.031.

Duan, Q., S. Sorooshian, and V. Gupta, 1992: Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resour. Res., 28, 1015–1031, https://doi.org/ 10.1029/91WR02985.

Franz, K. J., H. C. Hartmann, S. Sorooshian, and R. Bales, 2003: Verification of National Weather Service Ensemble Streamflow Predictions for Water Supply Forecasting in the Colorado River Basin. J. Hydrol., 4(6):1105-18, https://doi.org/10.1175/1525-7541(2003)004<1105:VONWSE>2.0.CO;2

Feng, X., A. Rafieeinasab, L. Karsten, W. Wu, D. Kitzmiller, Y. Liu, B. Cosgrove, L. Read, A. L. Dugger, Y. Zhang and K. FitzGerald, 2019: December. Calibrating the National Water Model V2. 1 over the Contiguous United States. In AGU Fall Meeting Abstracts (Vol. 2019, H43I-2134).

Jenkinson, A. F., 1955: The frequency distribution of the annual maximum (or minimum) of meteorological elements. Quart. J. Roy. Meteor. Soc., 81, 158–171, doi:10.1002/qj.49708134804.

Gochis, D., and Coauthors, 2019: Overview of National Water Model Calibration: General strategy and optimization. National Center for Atmospheric Research, accessed 1 January 2023, 30 pp.,https://ral.ucar.edu/sites/default/files/public/9_RafieeiNasab_CalibOverview_CUAHSI_Fall019_0.pdf

Gochis, D.J., M. Barlage, R. Cabell, M. Casali, A. Dugger, K. FitzGerald, M. McAllister, J. McCreight, A. RafieeiNasab, L. Read, K. Sampson, D. Yates, Y. Zhang, 2020: The WRF-

Hydro® modeling system technical description, (Version 5.2.0). NCAR Technical Note. 108 pages. Available online at: https://ral.ucar.edu/sites/default/files/public/projects/wrf-hydro/technical-description-user-guide/wrf-hydrov5.2technicaldescription.pdf

Gupta, H. V., T. Wagener, and Y. Liu, 2008: Reconciling theory with observations: Elements of a diagnostic approach to model evaluation. Hydrol. Processes, 22, 3802–3813, https://doi.org/10.1002/hyp.6989.

Gupta, H. V. et al.,2009: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of Hydrology, 377, 80-91.

Holtzman, N.M., T. M. Pavelsky, J. S. Cohen, M. L. Wrzesien and J. D. Herman, 2020: Tailoring WRF and Noah-MP to improve process representation of Sierra Nevada runoff: Diagnostic evaluation and applications. Journal of Advances in Modeling Earth Systems, 12(3), p.e2019MS001832.

Institute for Business & Home Safety, 2001: Subcommittee on Natural Disaster Reduction. Lessons from Living with Earth's Extremes. Washington, D.C.

Knoben, W.J., J. E. Freer and R. A. Woods, 2019: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. Hydrology and Earth System Sciences, 23(10), 4323-4331.

Kundzewicz, Z.W. and Z. Kaczmarek, 2000: Coping with hydrological extremes. Water International, 25(1), 66-75.

Lang, M., T. Ouarda, and B. Bobee, 1999: Towards operational guidelines for over-threshold modeling. J. Hydrol., 225, 103117, https://doi.org/10.1016/S0022-1694(99)00167-5.

Lahmers, T. M., H. Gupta, C. L. Castro, D. J. Gochis, D. Yates, A. Dugger, D. Goodrich and P. Hazenberg, 2019: Enhancing the structure of the WRF-hydro hydrologic model for semiarid environments. Journal of Hydrometeorology, 20(4):691-714.

Lahmers, T.M., et al., 2021: Evaluation of NOAA national water model parameter calibration in semiarid environments prone to channel infiltration. Journal of Hydrometeorology, 22(11), 2939-2969.

Lespinas, F., A. Dastoor, and V. Fortin, 2017: Performance of the dynamically dimensioned search algorithm: influence of parameter initialization strategy when calibrating a physically based hydrological model. Hydrol. Res., 49 (4), 971–988.

Lin, P., M. A. Rajib, Z.-L. Yang, M. Somos-Valenzuela, V. Merwade, D. R. Maidment, Y. Wang, and L. Chen, 2018: Spatiotemporal Evaluation of Simulated Evapotranspiration and Streamflow over Texas Using the WRF-Hydro-RAPID Modeling Framework. J. Amer. Water Resour. Assoc., 54, 40–54.

Mascaro, G., A. Hussein, A. Dugger, and D. J. Gochis, 2023: Process-based calibration of WRF-Hydro in a mountainous basin in southwestern US. JAWRA Journal of the American Water Resources Association, 59(1), 49-70.

NCAR, 2022: Noah-Multiparameterization Land Surface Model (Noah-MP LSM). Accessed 1 July, 2022, https://ral.ucar.edu/solutions/products/noah-multiparameterization-landsurface-model-noah-mp-lsm.

National Weather Service, 2014: United States Flood Loss Report - Water Year 2014. https://www.nws.noaa.gov/os/ water/Flood%20Loss%20Reports/WY14%20Flood%20Loss%20Summary.pdf.

National Weather Service, 2017: United States Flood Loss Report - Water Year 2017. https://www.weather.gov/media/water/WY17%20Flood%20Deaths%20and%20Direct%20Damagesv2.pdf

National Hydrologic Warning Council, 2002: Use and Benefits of the National Weather Service River and Flood Forecasts. https://www.weather.gov/media/water/AHPS_Benefits.pdf

Niu, G. Y., Z. L. Yang, K. E. Mitchell, F. Chen, M. B. Ek, M. Barlage, A. Kumar, et al., 2011: The Community Noah Land Surface Model with Multiparameterization Options (Noah-MP): 1. Model Description and Evaluation with Local-Scale Measurements. J. Geophys. Res. Atmos., 116, 1–19.

Niu, G. Y., Z. L. Yang, R. E. Dickinson, & L. E. Gulden, 2005: A simple TOPMODEL-based runoff parameterization (SIMTOP) for use in global climate models. Journal of Geophysical Research: Atmospheres, 110(D21).

Niu, G.-Y., Z.L. Yang, R. E. Dickinson, L. E. Gulden, and H. Su, 2007: Development of a simple groundwater model for use in climate models and evaluation with Gravity Recovery and Climate Experiment data. Journal of Geophysical Research, 112, D07103, https://doi.org/10.1029/2006JD007522.

NOAA. 2016: “National Water Model.” Improving NOAA's Water Prediction Service. https://water.noaa.gov/docum ents/wrn-national-water-model.pdf.

NOAA National Centers for Environmental Information (NCEI) U.S. Billion-Dollar Weather and Climate Disasters (2023). https://www.ncei.noaa.gov/access/billions/, DOI: 10.25921/stkw-7w73

Quenum, G.M.L.D., J. Arnault, N.A.B. Klutse, Z. Zhang, H. Kunstmann, and P.G. Oguntunde, 2022: Potential of the coupled WRF/WRF-hydro modeling system for flood forecasting in the Ouémé River (West Africa). Water, 14(8):1192.

Reed, S., V. Koren, M. Smith, Z. Zhang, F. Moreda, D.J. Seo, and D.M.I.P. Participants, 2004: Overall distributed model intercomparison project results. Journal of Hydrology, 298(1-4), 27-60.

Rogers, D. and V. Tsirkunov, 2011: Costs and benefits of early warning systems. *Global assessment rep*.

Rojas, M., F. Quintero, and W. F. Krajewski, 2019: Performance of the National Water Model in Iowa Using Independent Observations. J. Amer. Water Resour. Assoc., 56, 568–585.

Salas, F. R., M. A. Somos-Valenzuela, A. Dugger, D. R. Maidment, D. J. Gochis, C. H. David, W. Yu, D. Ding, E. P. Clark, and N. Noman, 2018: Towards Real-Time Continental Scale Streamflow Simulation in Continuous and Discrete Space. J. Amer. Water Resour. Assoc., 54, 7–27.

Schaake, J. C., V. I. Koren, Q.-Y. Duan, K. Mitchell, & F. Chen, 1996: Simple water balance model for estimating runoff at different spatial and temporal scales. Journal of Geophysical Research, 101(D3), 7461–7475. https://doi.org/10.1029/95JD02892

Sofokleous, I., A. Bruggeman, C. Camera, and M. Eliades, 2023: Grid-based calibration of the WRF-Hydro with Noah-MP model with improved groundwater and transpiration process equations. Journal of Hydrology, 617:128991.

Sun, M., Z. Li, C. Yao, Z. Liu, J. Wang, A. Hou, K. Zhang, W. Huo, and M. Liu, 2020: Evaluation of flood prediction capability of the WRF-hydro model based on multiple forcing scenarios. Water, 12(3):874.

Tolson, B. A. and C. A. Shoemaker, 2007: Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. Water Resour. Res., 43, W01413, https://doi.org/10.1029/2005WR004723.

USGS, 2018: USGS EROS archive - Land Cover Products – Global Land Cover Characterization (GLCC). Accessed 1 July 2022, https://doi.org/10.5066/F7GB230D.

Viterbo, F., K. Mahoney, L. Read, F. Salas, B. Bates, J. Elliott, B. Cosgrove, A. Dugger, D. Gochis, and R. Cifelli, 2020: A multiscale, hydrometeorological forecast evaluation of national water model forecasts of the May 2018 Ellicott City, Maryland, Flood. Journal of Hydrometeorology, 21(3), 475-499.

WRF-Hydro Development Team, 2020: How to Build & Run WRF-Hydro V5.1.1 in Standalone Mode. https://ral.ucar.edu/sites/default/files/docs/water/howtobuildrunwrfhydrov511instandalonemode.pdf, accessed July 31,2022

Xia Y., and coauthors, 2012: Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. Journal of Geophysical Research: Atmospheres, 117(D3).

# Chapter 4 Improving Runoff Simulation in the Western United States with Noah-MP and VIC

This chapter is submitted to Hydrology and Earth System Sciences as

Lu Su, Dennis P. Lettenmaier, Ming Pan, Benjamin Bass, 2023: Improving Runoff Simulation in the Western United States with Noah-MP and VIC., Hydrology and Earth System Sciences (under review).

The supplemental material for this chapter is provided in Appendix C.

## Abstract

Streamflow forecasts are critical for water and environmental management, especially in the water-short Western U.S.. Land Surface Models (LSMs), such as the Variable Infiltration Capacity (VIC) model and the Noah-Multiparameterization (Noah-MP) play an essential role in providing comprehensive runoff forecasts across the region. Virtually all LSMs require parameter estimation to optimize their predictive capabilities. We describe a systematic calibration of parameters for VIC and Noah-MP over 263 river basins in the Western U.S., and distribution of the calibrated parameters over the entire region. Post-calibration results showed a notable improvement in model accuracy in the calibration basins: the median daily streamflow Kling-Gupta Efficiency (KGE) for VIC rose from 0.37 to 0.70, and for Noah-MP, from 0.22 to 0.54. Employing the donor-basin regionalization method, we developed transfer relationships to hydrologically similar basins and extended the calibrated parameters to ungauged basins and the entire region. We assessed factors that influence calibration efficiency and model performance using regional parameter estimates. We evaluated high and low flow simulation capabilities of the two models and observed marked improvements after calibration and regionalization. We also generated gridded parameter sets for

both models across all 4816 HUC-10 basins in the Western U.S., a data set that is intended to support regional hydrologic studies and hydrologic climate change assessments.

## 4.1 Introduction

Streamflow forecasts play a key role in various aspects of water and environmental management, especially in the Western U.S. (WUS). In the short term, these forecasts provide early warnings for impending flood events, thereby enabling timely preparation and response to mitigate immediate flood risk and damages (Maidment, 2017). They also serve as crucial input for managing reservoirs effectively for water supply (Raff et al., 2013), hydroelectric power generation (Boucher & Ramos, 2018), and river navigation (by providing a basis for predicting water levels) (Federal Institute of Hydrology, 2020). In the longer term, streamflow forecasts enable water utilities and agencies to plan water distribution within and across multiple uses—urban, agricultural, and industrial—which is especially vital during drought conditions when efficient water use becomes a necessity (Anghileri et al., 2016). Streamflow forecasts also aid in understanding and predicting the impacts of climate change on water systems, thereby informing adaptive strategies for water resource management. Thus, in both short and longer-term contexts, streamflow forecasts are an important tool for promoting sustainable water practices and resilience to water-related challenges.

Streamflow forecasts are derived via a synthesis of hydrometeorological data, statistical methodologies, and computational modeling. Direct measurement of runoff is an important element of streamflow forecasts, however it is only possible in river basins with well-developed observational infrastructure (Sharma and Machiwal, 2021). This limitation leaves vast areas, often critical to water resource management and climatology, without direct runoff observations on which to base streamflow forecasts. As an alternative, Land Surface Models (LSMs) can be used

to simulate streamflow. LSMs typically are forced with air temperature, precipitation and other meteorological forcings. By integrating climatic, topographic, and land-use information, they can fill streamflow observation gaps and provide comprehensive, spatially distributed runoff forecasts (Fisher and Koven, 2020). The capabilities of LSMs equip us with the necessary tools to produce streamflow forecasts that can be used to prepare for severe weather conditions, form the basis for water resource management, and inform water management associated with our evolving climate. These benefits hold true irrespective of the limitations associated with direct streamflow observations. Through off-line simulations and reconstructions, LSMs enable us to gain insights into land surface hydrology at various scales - regional, continental, and global.

The parameterization of the underlying hydrological processes varies across different LSMs, but virtually all models require some level of parameter estimation based on historical observed streamflow data at forecast point, to ensure trustworthy predictions throughout the region (Beven, 1989; Troy et al., 2008; Gong et al., 2015). In cases where observations don't exist, parameters can be transferred from river basins where they do (Arsenault and Brissette, 2014). In cases where observations do exist but aren't current, we can use a shorter span of historical streamflow data for model calibration and subsequently produce streamflow forecasts using meteorological forcings when observed streamflow data aren't available.

The process of calibration can be computationally demanding, and prior research typically has focused on obtaining parameters appropriate to facilitating model simulations that match observations as closely as possible at the observation point (Duan et al., 1992; Tolson and Shoemaker, 2007). Most previous studies have concentrated on a limited number of basins and a single model (e.g. Mascaro et al., 2023; Sofokleous et al., 2023; and Gou et al., 2020). Here, we aim to establish parameterizations for two LSMs -- the Variable Infiltration Capacity (VIC) model

and the Noah-Multiparameterization (Noah-MP) LSM across the WUS. Both models have found extensive application both within the U.S. and internationally (Mendoza et al., 2015; Tangdamrongsub, 2023). The approach we use involves the application of globally optimized calibration methods and regionalization, with the objective of facilitating these models to provide reliable runoff simulations.

In particular, we explore and elucidate (i) the choice of physical parameterizations and calibration of land surface parameters, (ii) extension of these calibrated parameters to areas without gauges, and (iii) factors that influence calibration efficiency and LSM performance using regional parameter estimates. In the case of Noah-MP, which offers multiple runoff generation (physics) options, our initial step involves choosing the most effective runoff parameterization option. Following this, we perform the calibration of land surface parameters. In the case of the VIC model, the runoff parameterization scheme is predetermined, so we commence immediately with calibration. We implemented calibration in 263 basins across the WUS where streamflow observations were available (see section 4.2.1 for details) and compared simulated and observed streamflow as the model predictions were affected by soil and other land surface properties. Our second step extended the initial calibrated land surface parameters to ungauged basins. We then explored the variables that most impact the calibration proficiency of Noah-MP and VIC across the WUS. In section 4.4, we employ a regionalization technique known as the donor basin method, as implemented by Bass et al. (2023). Finally, we evaluate both flood and low flow simulation skills for the baseline, and after calibration and regionalization.

## 4.2 Study basins, land surface models and forcing dataset overview

### 4.2.1 Study Basins

We selected 263 river basins distributed across the WUS. Most of the basins were from USGS Gages II reference basins (Falcone, 2011) which have minimum upstream anthropogenic effects such as dams and diversions. Among these basins, our selection criteria included having at least 20 years of record, and a minimum drainage area of 144 square kilometers, which is the size of four model grid cells. In addition to 250 Gages II reference stations, we included 13 basins located in California's Sierra Nevada for which natural flows are available from the California Department of Water Resources (2021). The geographical distribution of the 263 basins is shown in Figure 4.1. We focused on the hydrological models' calibration to full natural flow (the same as observed streamflow for GAGES II stations; estimated by DWR for the 13 Sierra Nevada sites), which indicates water flow conditions devoid of human interventions like reservoirs or diversions. Each basin was calibrated using the most recent 20-year period when the observation is available.

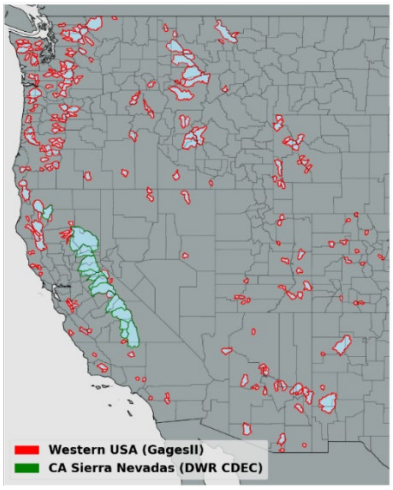



Figure 4.1 263 river basins for which calibration was performed. The Gages II reference basins are delineated with red boundaries and the CA Sierra Nevade basins with green boundaries.

### 4.2.2 Land Surface Models

We included two widely used hydrological models, VIC (Liang et al., 1994) and Noah-MP (Niu et al., 2011). This decision was informed by the varying levels of complexity these two models offer in conceptualizing the effects of vegetation, soil, and seasonal snowpack on the land surface energy and water balances (refer to Table 4.1 for more details). The two models also use different parameterizations for certain hydrological processes, including unique model equations for canopy water storage, base flow, and other processes. Both of these hydrological model structures have found extensive application both within the U.S. and internationally, as indicated by Mendoza et al. (2015) and Tangdamrongsub (2023).

To generate streamflow, the gridded runoff from Noah-MP and VIC was accumulated over each watershed. We didn't implement routing since its impact on daily streamflow simulations was small given the relatively small size of most of the basins. This aligns with earlier research (e.g., Li et al., 2019). However, in both the case of VIC and Noah-MP, the output of our simulations (runoff) could be used as input to routing models, such as those that are options in the implementation of both models.

#### 4.2.2.1 VIC

VIC is a macroscale, semi-distributed hydrologic model (described in detail by Liang et al 1994) that determines land surface moisture and energy states and fluxes by solving the surface water and energy balances. VIC is a research model and in its various forms it has been employed to study many major river basins worldwide (e.g. Adam et al., 2003 & 2006; Livneh et al., 2013; Schaperow et al., 2021). This model enjoys a broad user community — as per the citation index Web of Science, the initial VIC paper has been referenced more than 2600 times, with contributing authors spanning at least 56 different countries (Schaperow et al., 2021). We obtained initial VIC

model parameters from Livneh et al., 2013, who validated model discharges over major CONUS river basins. The origins of the soil and land cover data are outlined in Table 4.1. The version of the VIC model implemented here is 4.1.2, and it operates in energy balance mode.

Table 4.1 Overview of hydrologic model components and parameter data sources.

| Model | Snow accumulation and melt | Moisture in the soil and column/surface runoff | Base flow | Canopy storage | Vegetation data | Soil data |
|---|---|---|---|---|---|---|
| **VIC (V4.1.2)** | Two-layer energy–mass balance model | Infiltration capacity function. Vertical movement of moisture through soil follows 1D Richards equation. | A function of the soil moisture in the third layer. Linear below a soil moisture threshold and becomes nonlinear above that threshold. (Liang et al., 1994) | Mosaic representation of different vegetation coverages at each cell. | University of Maryland 1-km Global Land Cover Classification (Hansen et al., 2000) | 1-km STATSGO database (Miller and White, 1998) |
| **NOAH-MP (WRF-HYDRO 5.2.0)** | Three-layer energy–mass balance model that represents percolation, retention, and refreezing of meltwater within the snowpack. | 1. TOPMODEL-based runoff scheme | Simple groundwater (hereafter SIMGM) (Niu et al., 2007) | Semi-tile approach for computing longwave, latent heat, sensible heat and ground heat fluxes | MODIS 30-second Modified IGBP 20-category land cover product | 1-km STATSGO database (Miller and White, 1998) |
| | | 2. Simple TOPMODEL-based runoff scheme with an equilibrium water table (hereafter SIMTOP) | Similar to SIMGM, but with a sealed bottom of the soil column (Niu et al., 2005) | | | |
| | | 3. Infiltration-excess-based surface runoff scheme | Gravitational free-drainage subsurface runoff scheme (Schaake et al., 1996) | | | |
| | | 4. BATS runoff scheme, which parameterized surface runoff as a | Gravitational free drainage (Dickinson et al., 1993) | | | |

| | | |
|---|---|---|
| | | 4th power function of the top 2 m soil wetness (degree of saturation) |

#### 4.2.2.2 Noah-MP

Noah-MP is a state-of-the-art LSM originally designed as the land surface scheme for numerical weather prediction (NWP) models like the Weather Research and Forecasting (WRF) regional atmospheric model. Currently, it's being utilized for physically based, spatially-distributed hydrological simulations as a component of the National Water Model (NWM) (NOAA, 2016). It enhances the functionalities of the Noah LSM (as per Chen et al., 1996 and Chen and Dudhia, 2001) previously used in NOAA's suite of numerical weather prediction models by offering multiple options for key processes that control land-atmosphere transfers of moisture and energy. These include surface water infiltration, runoff, evapotranspiration, groundwater movement, and channel routing (see Niu et al., 2007; 2011). The model has been widely used for forecasting seasonal climate, weather, droughts, and floods not only across the continental United States (CONUS) but also globally (Zheng et al., 2019).

### 4.2.3 Forcing Dataset

We ran both models at a 3-hour time step and at $1/16°$ latitude–longitude spatial resolution. The forcings were the gridded observation dataset developed by Livneh et al. (2013) and extended to 2018 by Su et al. (2021) (hereafter referred to as L13). This data set spans the period from 1915 to 2018. For the VIC model, the L13 dataset provided daily values of precipitation, maximum and minimum temperatures, and wind speed (additional variables used by VIC including downward solar and longwave radiation, and specific humidity, are computed internally using MTCLIM algorithms as described by Bohn et al., 2013). The Noah-MP model, on the other hand, necessitated additional meteorological data such as specific humidity, surface pressure, and downward solar

and longwave radiation, in addition to precipitation, wind speed, and air temperature. We used the MTCLIM algorithms, as detailed by Bohn et al. (2013), to calculate specific humidity and downward solar radiation. We employed the Prata (1996) algorithm to compute the downward longwave radiation. Additionally, we deduced surface air pressure by considering the grid cell elevation in conjunction with standard global pressure lapse rates. Following this, we transitioned the daily data to hourly metrics using a cubic spline to interpolate between Tmax and Tmin, and derived other variables using the methods explained by Bohn et al. (2013). Lastly, we distributed the daily precipitation evenly across three hourly intervals.

## 4.3 Model calibration

### 4.3.1 Calibration methods

The initial step in our calibration effort was to optimize the land surface parameters of the two models for the 263 WUS basins. These parameters, primarily soil properties which can exhibit a substantial degree of uncertainty, were iteratively updated via hundreds of simulations to accurately reflect streamflow conditions in each basin. We calibrated six parameters for VIC and five for Noah-MP. This selection was guided by past research and the computational resources we had at our disposal (Mendoza et al., 2015; Hussein, 2020; Shi et al., 2008; Holtzman et al., 2020; Bass et al., 2023; Schaperow et al., 2023). Each parameter underwent consideration across a physically viable range (refer to Table 4.2), drawing from values utilized in prior studies (Cai et al., 2014; Mendoza et al., 2015; Hussein, 2020; Shi et al., 2008; Gochis et al., 2019; Holtzman et al., 2020; Lahmers et al., 2021; Bass et al., 2023; Schaperow et al., 2023). Through our iterative calibration method, each subsequent simulation learns from the previous ones using algorithms designed to reduce the discrepancy between the simulated and observed streamflow.

For VIC parameter estimation, we employed the Shuffled Complex Evolution algorithm developed at the University of Arizona (SCE-UA, Duan et al., 1992). This method is a global optimization method widely used in hydrology and environmental modeling, owing to its robustness and efficiency when addressing complex, non-linear, and multi-modal objective functions (Naeini et al., 2015).

For the Noah-MP model, which requires more computational core-hours per simulation, we used the Dynamically Dimensioned Search (DDS) algorithm of Tolson and Shoemaker (2007). This algorithm, specifically crafted for high-dimensional and computationally intensive problems, offers generally greater efficiency than SCE-UA. NOAA employs the DDS algorithm for their CONUS implementation of NWM, which is grounded in Noah-MP (Gochis et al., 2019). We evaluated both calibration methods (DDS and SCE-UA) for VIC for 20 randomly chosen basins and obtained similar results. For VIC, we chose SCE-UA due to its inherent compatibility with the model and because the additional computation (relative to DDS) was less important given that the inherent computation required for VIC is considerably less than for Noah-MP.

In our application of SCE-UA, we performed a maximum of 3000 iterations for each basin, while the DDS method employed 250 iterations for each basin for Noah-MP. Each basin was calibrated using the most recent 20 years of streamflow data. For both models, our objective function was the Kling-Gupta Efficiency (KGE, Gupta et al., 2009) metric for daily streamflow. KGE is a widely used performance measure because of its advantages in orthogonally considering bias, correlation and variability (Knoben et al., 2019). KGE = 1 indicates perfect agreement between simulations and observations; KGE values greater than -0.41 indicate that a model improves upon the mean flow benchmark (Konben et al., 2019).

Table 4.2 Calibration methods, parameters and modifications to their initial default values evaluated in the calibration.

| Model | VIC | | Noah-MP | |
|---|---|---|---|---|
| **Calibration Method** | SCE-UA | | DDS | |
| **Iterations** | 3000 | | 250 | |
| **Calibrated Parameter** | Variable Infiltration Curve Parameter (INFILT) | 0.001 – 0.4 (Shi et al., 2008) | Saturated Hydraulic Conductivity (Ksat) | $2 \times 10^{-9} to$ 0.07(Cai et al., 2014) |
| | Baseflow parameter (Ds) | 0.001 – 1.0 (Shi et al., 2008) | Saturation soil moisture content (MAXSMC) | 0.1 to 0.71 (Cai et al., 2014) |
| | Thickness of Soil in Layer 1 (Depth_1) | 0.01 – 0.2 (Shi et al., 2008) | Pore size distribution index (Bexp) | 1.12 to 22 (Cai et al., 2014; Gochis et al., 2019) |
| | Total thickness of soil column (Depth_total) | 0.6 – 3.5 (Shi et al.,2008) | Linear scaling of "openness" of bottom drainage boundary (Slope) | 0.1-1 (Lahmers et al., 2021) |
| | Max velocity parameter of baseflow (Dsmax) | 0.001 – 30 (Schaperow et al., 2023) | Parameter in surface runoff (REFKDT) | 0.1-10 (Lahmers et al., 2021) |
| | Fraction of max soil moisture where nonlinear baseflow occurs (Ws) | 0.001 – 1 (Shi et al., 2008) | | |

### 4.3.2 Noah-MP parameterization

As specified in Table 4.1, Noah-MP has four runoff and groundwater physics options (rnf). Initially, we adopted the options that are incorporated in the NWM, as elaborated in Gochis et al. (2020). Before we could proceed with calibrating Noah-MP for all the WUS basins, it was necessary to determine suitable rnfs. To streamline computational time, we initially selected 50 basins randomly from the total of 263 from which we created four experimental groups. Each group employed a different rnf option. We applied the DDS method to these groups and compared

the cumulative distribution functions (CDF) of their baseline and calibrated KGEs (Figure 4.2). From this figure, it's apparent that the KGE improved post-calibration for all four rnfs. Notably, rnf3, also known as free drainage, exhibited the most substantial performance enhancement after calibration. As a result, we chose to continue using this option which is incorporatead in the NWM. Nonetheless, it's worth noting that the use of different options for different basins—a feature currently not utilized in Noah-MP or WRF-Hydro—could potentially result in improved overall model performance.

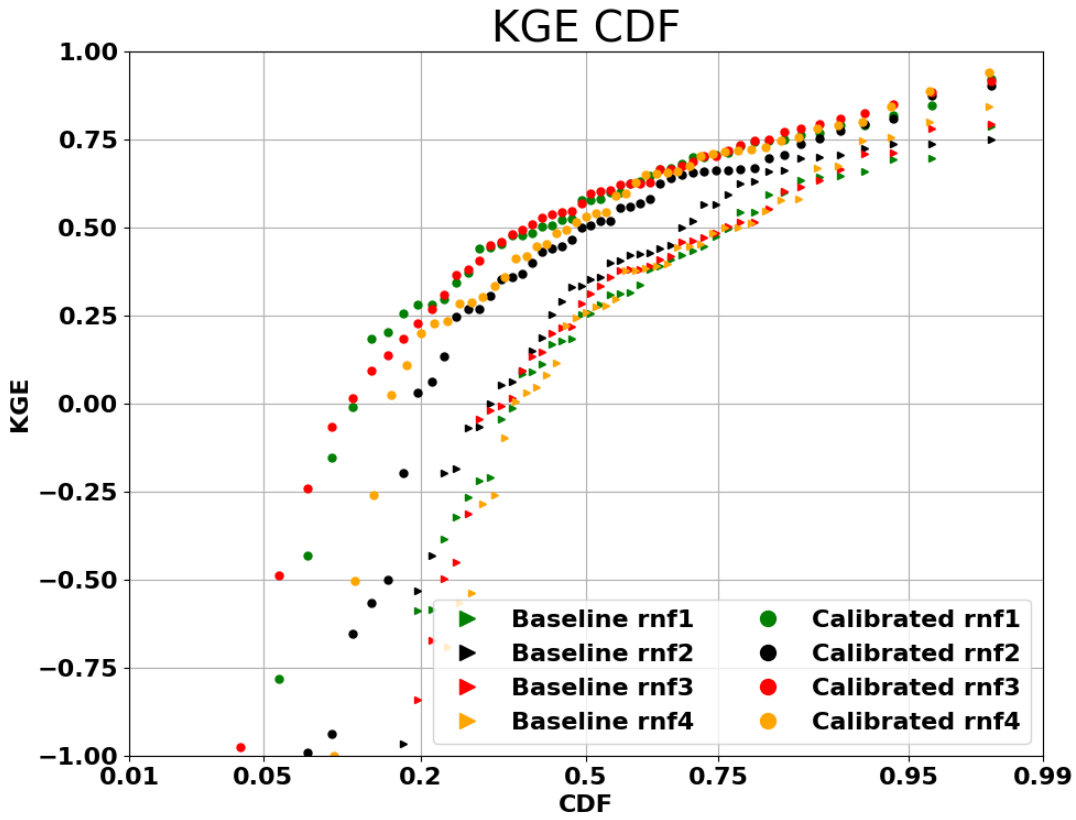



Figure 4.2 Streamflow performance (KGE of daily streamflow simulations) of different Noah-MP runoff parameterizations across 50 (of 263) randomly selected basins. The performances are shown for both baseline and calibrated simulations.

### 4.3.3 Calibration of gauged basins

Following the selection of the most effective set of runoff generation options across the domain, we estimated model parameters for all 263 basins. The comparative performance of the models, before and after calibration, is shown in Figure 4.3. It's apparent from the figure that both Noah-MP and VIC have significantly enhanced their daily streamflow simulation skills post-calibration. After calibration, the median KGE of Noah-MP improved from 0.22 to 0.54, and the VIC's median KGE increased from 0.37 to 0.70. When contrasting the two models, we observed that VIC outperformed Noah-MP both pre- and post-calibration. One possible explanation could be that the baseline VIC parameters were taken from Livneh et al. (2013), and these parameters had already been validated and adjusted for major U.S. basins (although not for our 263 basins specifically), while the Noah-MP parameters are default values from NWM. Another possibility is inherent differences in the physics of streamflow simulation between the two models (VIC primarily generates runoff via the saturation excess mechanism), although that isn't the main focus of our research.

Following the calibration with data from the past 20 years, we performed a test where we calibrated the streamflow using the first 10 years of data and validated with the subsequent 10 years of data. This test revealed that the KGE distribution from the 10-year calibration is similar to that from the 20-year data. The median KGE values for Noah-MP and VIC after calibration with 10 years of observations were 0.52 and 0.69, respectively. Correspondingly, the median KGEs during the validation period were 0.50 and 0.68, respectively, which are only slightly lower. These comparisons demonstrate general consistency over time in the performance of the calibrated parameters.

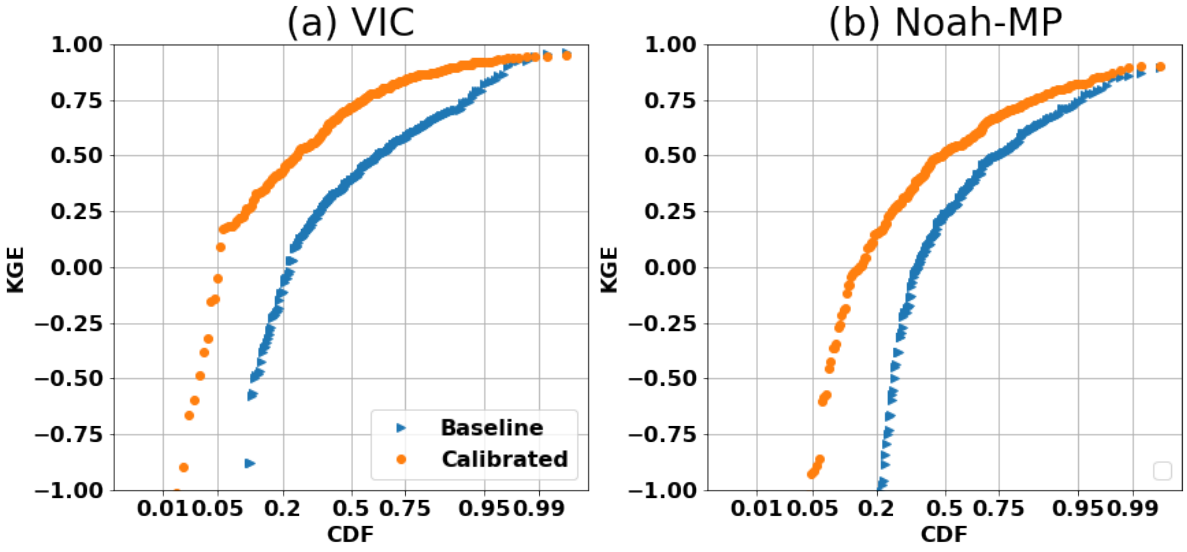



Figure 4.3 Cumulative Distribution Function (CDF) plot of the daily streamflow KGE for (a) VIC and (b) Noah-MP, comparing baseline and calibrated runs across all 263 basins.

We examined the spatial variability of daily streamflow KGE for Noah-MP and VIC, both before and after the calibration (see Figure 4.4). The highest baseline KGEs are along the Pacific Coast, in central to northern CA for both models. VIC's baseline KGE generally is high in the Pacific Northwest. Post-calibration improvements occurred for both models in most areas, especially in regions where the baseline KGE was low, such as southern CA and the southeastern part of the study region. Median improvements after calibration were 0.27 for Noah-MP and 0.30 for VIC.

We observed that basins displaying higher KGE values typically were more humid than those with lower KGE. To further delve into the relationship between KGE and basin characteristics, we explored correlations between KGE and 21 different characteristics, including drainage area, elevation, seasonal/annual average temperature and precipitation, annual maximum precipitation, and seasonal/annual runoff ratio. Of these, 12 characteristics were statistically significantly correlated with the VIC KGE, including four seasonal and annual runoff ratios; mean precipitation

in winter, spring, and fall; annual maximum precipitation; and minimum elevation. Figure 4.5 shows scatterplots of eight representative characteristics. Apart from minimum elevation and mean summer temperature, all other characteristics were positively correlated with KGE. Typically, spring runoff ratio, annual runoff ratio, mean annual max precipitation, and mean winter precipitation exhibited the highest correlations with KGE. This implies that basins with higher runoff ratios (particularly in spring), higher precipitation (especially maximum precipitation), lower summer temperature, and lower elevation are more likely to exhibit strong VIC performance. The same applies to Noah-MP, as indicated in Figure 4.6, although Noah-MP showed relatively weaker correlations. Correlations between mean summer temperature and mean fall precipitation and Noah-MP KGE weren't statistically significant.

The spatial distribution of the eight characteristics is qualitatively similar with the KGE spatial distribution, as shown in Figure 4.7. Generally, basins with higher KGE have higher characteristic values when the correlation is positive, and lower characteristic values when the correlation is negative.

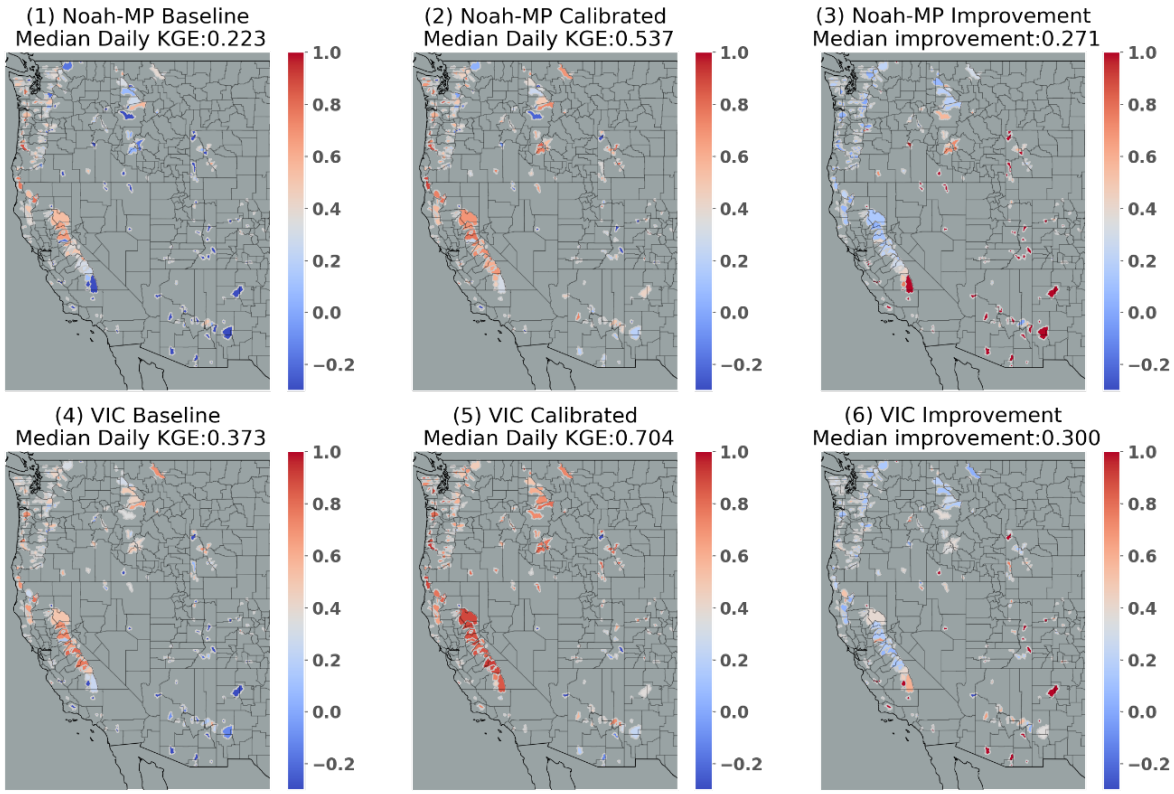



Figure 4.4 Spatial distribution of basins' daily streamflow KGE for Noah-MP baseline (1); calibrated Noah-MP (2); difference between calibrated and baseline Noah-MP; VIC baseline (4); calibrated VIC (5); difference between calibrated and baseline VIC.

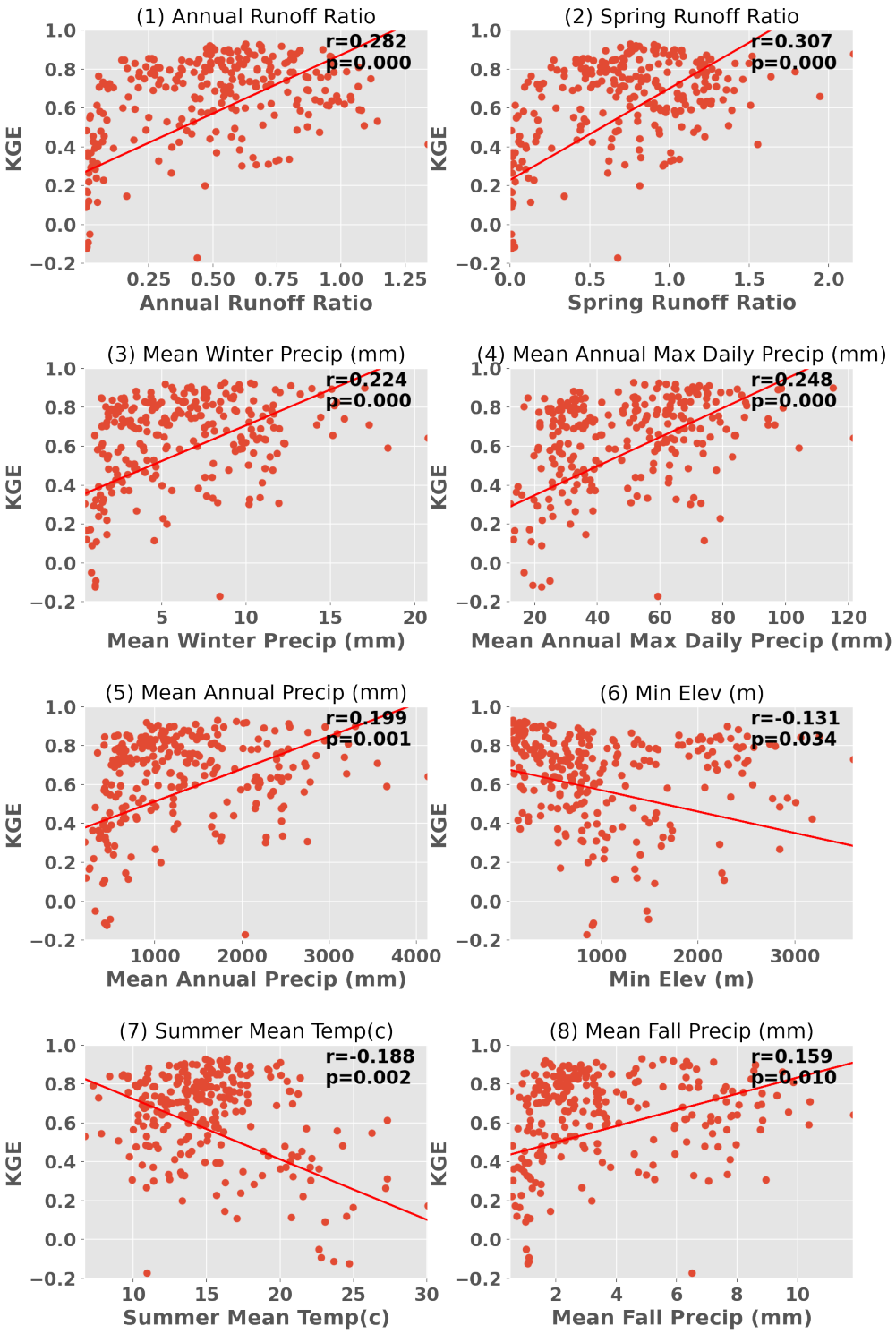



Figure 4.5 Scatterplots of VIC KGE in relation to significantly correlated characteristics. Each subplot indicates the corresponding Pearson correlation coefficients and the P-value.

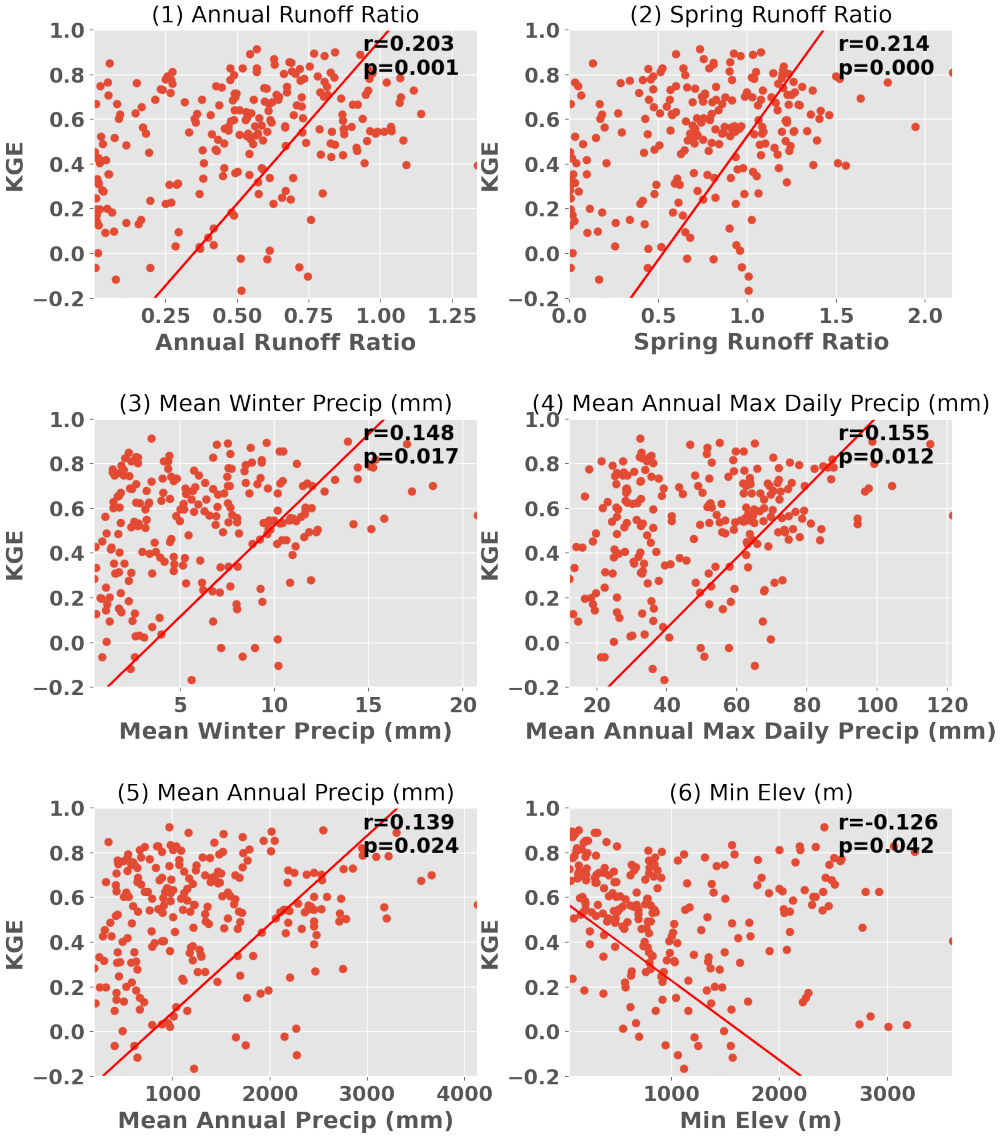



Figure 4.6 Scatterplot of Noah-MP KGE in relation to significantly correlated characteristics. Each subplot indicates the corresponding Pearson correlation coefficients and the P-value.

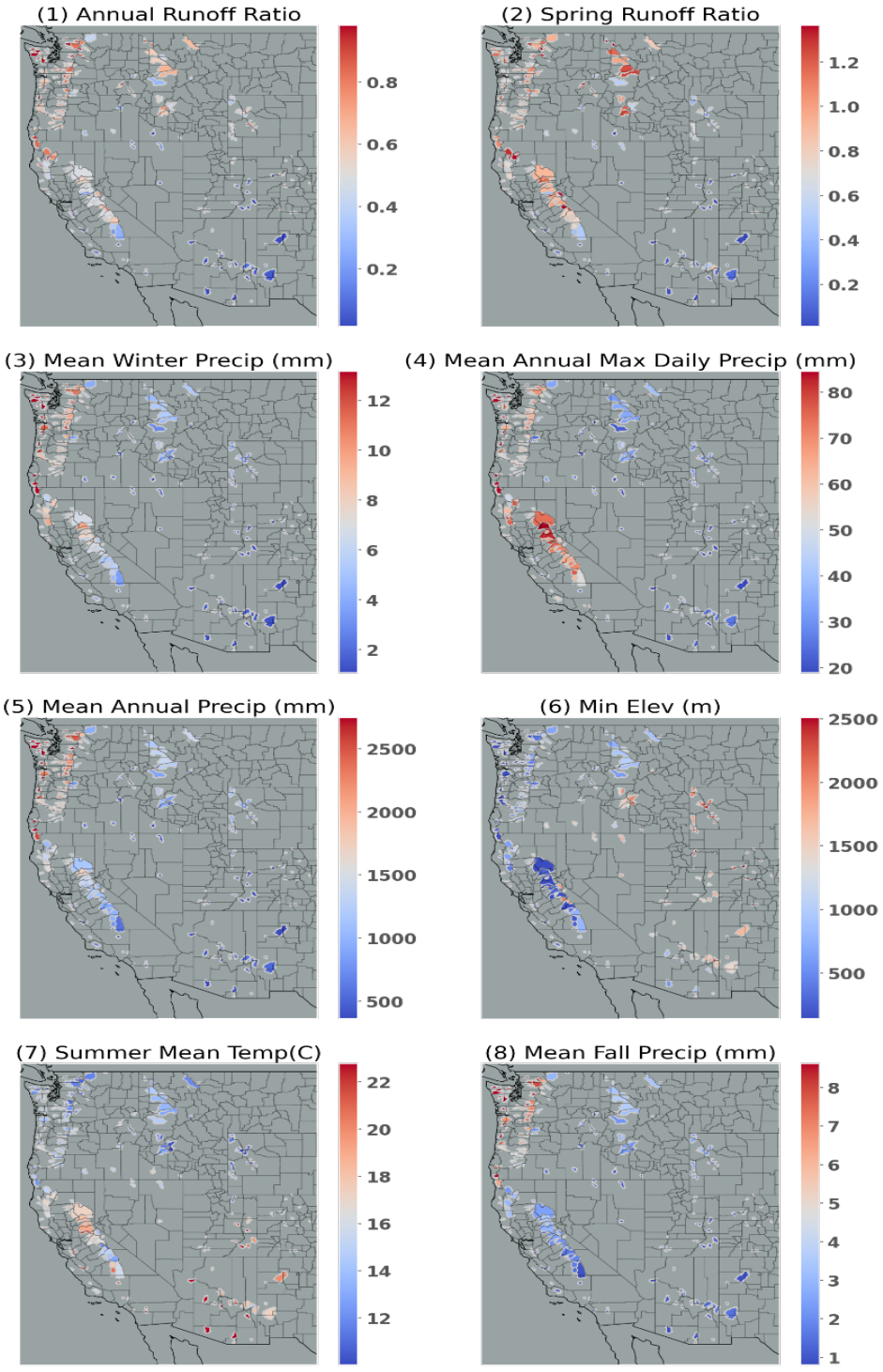



Figure 4.7 Spatial distribution of characteristics that are statistically significantly correlated with KGE. Note that all characteristics are significantly correlated with VIC KGE whereas only (1)-(6) are significantly correlated with Noah-MP KGE.

### 4.4 Regionalization

Following the calibration process, we regionalized the parameters from gauged to ungauged basins based on a mathematical assessment of the spatial and physical proximity between the gauged and ungauged basins, following previous studies by Arsenault and Brissette (2014), and Razavi and Coulibaly (2017). We opted for this method over an alternate approach that first regionalizes the streamflow attributes (such as runoff depth, high flow indicators) and then standardizes the model throughout (as proposed by Castiglioni et al., 2010; Oubeidillah et al., 2014; and Yang et al., 2017). The reason for our choice is our interest in actual streamflow time series rather than metrics. We carried out the regionalization after calibrating to specific streamflow gauges, ensuring high precision for these gauged basins and facilitating high-quality regionalization in ungauged basins. Specifically, we employed a donor-basin approach where an ungauged basin adopts calibrated parameters from its most similar gauged basin(s). This method has been applied in many studies including Arsenault and Brissette (2014); Poissant et al. (2017); Razavi and Coulibaly (2017); Gochis et al. (2019); Qi et al. (2021) and Bass et al. (2023).

In the donor-basin method, an ungauged basin inherits its land surface parameters from the most similar gauged basin(s) (or the top 'x' most similar gauged basins). Here, we evaluated the similarity or proximity between gauged and ungauged basins based on the similarity index SI as defined and used by Burn and Boorman (1993) and Poissant et al. (2017):

$$SI = \sum_{i=1}^{k} \frac{|X_i^G - X_i^U|}{\Delta X_i} \quad (13)$$

In this formula, k stands for the total number of features considered, $X_i^G$ represents the ith feature of the gauged basin G, $X_i^U$ is the ith feature of a specific ungauged basin, and $\Delta X_i$ is the range of potential values for the ith feature, grounded in the data from the gauged basins. This yields a unique value of SI for each gauged basin, contingent on the specific ungauged basin it is compared

with. Typically, gauged basins that exhibit greater resemblance to the ungauged basin will have a smaller SI.

We assessed the donor-basin method's efficacy using a cross-validation approach, where each gauged basin was treated as ungauged one at a time. The pseudo-ungauged basin inherits its hydrological parameters from its three most similar gauged basins, determined by SI. The parameters inherited are a weighted average from the three donor basins. After testing one to five donor basins, we found that using three donors yielded the best results. Thus, every basin inherits parameters from the three most similar gauged basins in each simulation, offering a concise evaluation of the donor-basin method's regionalization performance.

We used 18 basin-specific features in the donor basin method, detailed in Table C1, calculated based on the forcings and parameters used in the study. For feature selection in the donor-basin method, we adopted an iterative approach. Each iteration added a single feature to the index, with the most beneficial feature (based on median KGE improvement) retained. This process was repeated until the median KGE no longer improved. Only basins with a KGE exceeding 0.3 were considered, following previous studies suggesting that inclusion of poorly performing basins can lower regionalization performance. We found that a KGE threshold of 0.3 resulted in a median performance improvement of 0.08 larger than did a KGE threshold of 0, hence it was chosen. After screening, 223 basins were utilized in VIC regionalization and 194 in Noah-MP regionalization.

We found five features generated the best regionalization performance for VIC (longitude centroid, latitude centroid, maximum elevation, fall mean precipitation, and fall mean temperature) and three features were best for Noah-MP (latitude centroid, longitude centroid, and drainage area) (see Figure 4.8). Among them, latitude and longitude are the common features that contribute the most to regionalization when using the similarity index method. This suggests that geographical

similarities are the most important factor in parameter information transfer from gauged to ungauged basins.

Upon evaluating the performance of baseline, calibrated, and regionalized simulations, the respective median daily KGEs for the VIC model were found to be 0.41, 0.71, and 0.49. For the Noah-MP, these values were 0.38, 0.60, and 0.49 (refer to Figures 4.8 & 4.9). These metrics are for basins that have a calibrated KGE greater than 0.3 only, resulting in higher median KGEs than for all 263 basins (See Figure 4.3). The KGE distribution also improved overall. It's noteworthy that the regionalization improvement relative to baseline is higher for Noah-MP than for VIC. While VIC's baseline and calibrated KGE skill distribution outperforms Noah-MP's, the regionalized skills of Noah-MP and VIC are quite comparable. This observation might be attributable to the constraints of the regionalization setup and could warrant future investigation.

After optimizing the features and specific design of the donor-basin method, parameters were regionalized to 4816 ungauged USGS Hydrologic Unit Code (HUC) 10 basins across the WUS. HUCs are delineated and quality controlled by USGS using high-resolution DEMs. The final hydrologic parameters for both VIC and Noah-MP for all WUS HUC-10 basins are shown in Figures C1&2. The baseline HUC-10 parameters are shown in Figures C3&4.

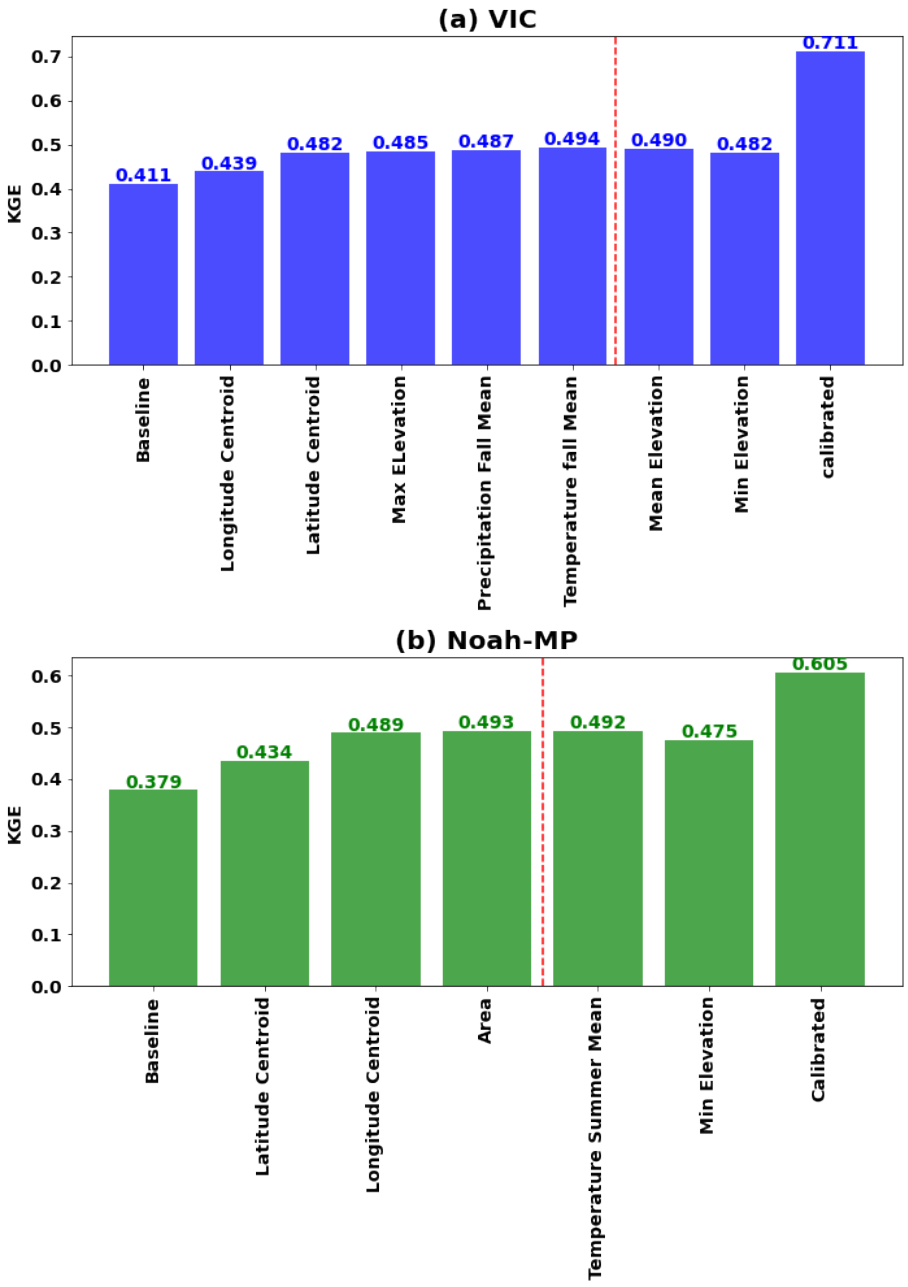



Figure 4.8 Best regionalization features for (a) VIC and (b) Noah-MP. The final regionalization to ungauged basins of the WUS incorporated all features up to the point marked by the red line since the addition of further features doesn't improve KGE.

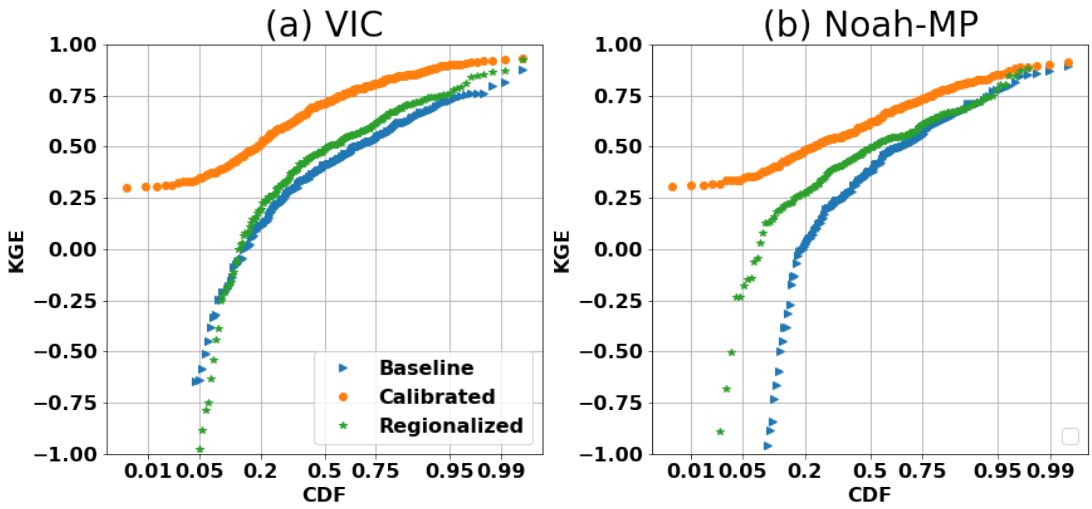



Figure 4.9 CDF of daily KGE for (a) VIC and (b) Noah-MP, comparing baseline and calibrated runs across selected basins within the WUS.

### 4.5 Evaluation of high and low flow simulation skills

To understand the capabilities of the two models in reconstructing high and low streamflow, we assessed their performance across baseline, calibrated, and regionalized settings.

(a) Evaluation of high flow performance

We used the peaks-over-threshold (POT) method (Lang et al., 1999) to identify extreme streamflow events as in Su et al. (2023) and Cao et al. (2019, 2020). We first applied the event independence criteria from USWRC (1982) to daily streamflow data to identify independent events. We set thresholds at each basin that resulted in 3 extreme events per year on average. After selecting the flood events over the study period based on the observation, we sorted the floods based on the return period and then calculated the KGE of baseline, calibrated and regionalized floods. Figure 4.10 displays the associated CDF plots. The median KGE for baseline floods in Noah-MP was 0.14, which rose to 0.37 post-calibration, and receded to 0.22 after regionalization. For VIC, the flood KGE started at 0.11, increased to 0.41 after calibration, and declined to 0.20 post-regionalization. As anticipated, these numbers are lower than (all) daily streamflow skill due

to our calibration target being daily streamflow. Still, flood competencies experienced considerable enhancement, surpassing the Noah-MP KGE benchmark of -0.41 found by Knoben et al. (2019).

(b) Evaluation of low flow performance

To assess low flow performance, we utilized the 7q10 metric. This hydrological statistic, commonly adopted in water resources management and environmental engineering, is the lowest 7-day average flow that occurs (on average) once every 10 years (EPA, 2018). Scatterplots of 7q10 (Figure 4.11) showed high correlation between our model's simulated low flows and the observed data. Post-calibration, this alignment intensified. The VIC model tended to underestimate the low flows. After calibration, the median bias improved from -23.6% to -9.9%, and with regionalization, it was -11.7%. In contrast, Noah-MP began with an 11.20% overestimation in the baseline, improved to 0.61% post-calibration, and was -9.5% after regionalization. The outcomes underline the proficiency of both models for low flow prediction, exhibiting enhanced competencies post-calibration and commendable performance after regionalization.

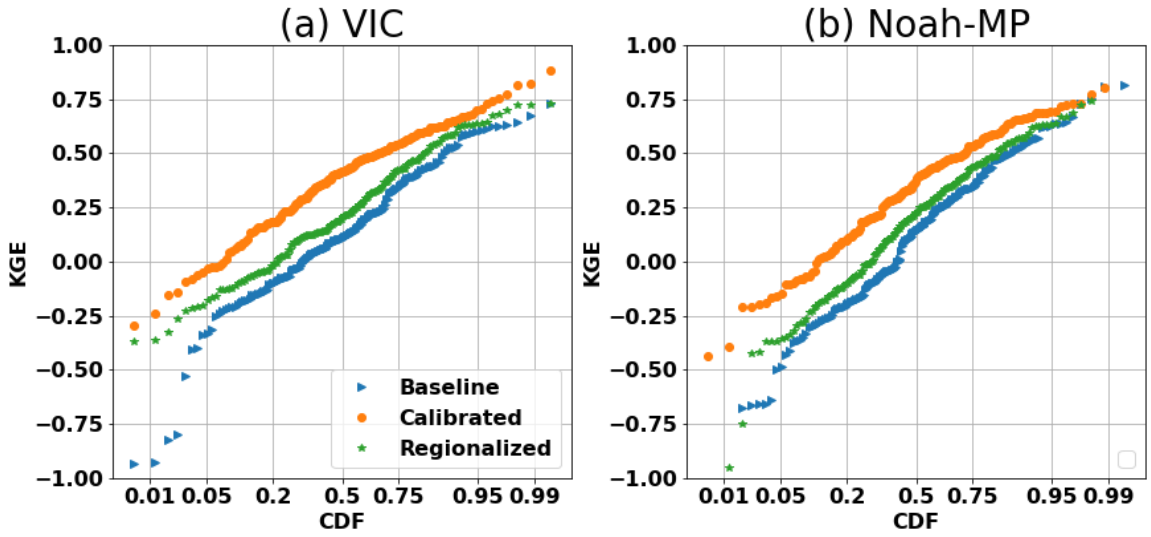



Figure 4.10 CDF of high flow KGE for (a) VIC and (b) Noah-MP, comparing baseline and calibrated runs across selected basins within the WUS.

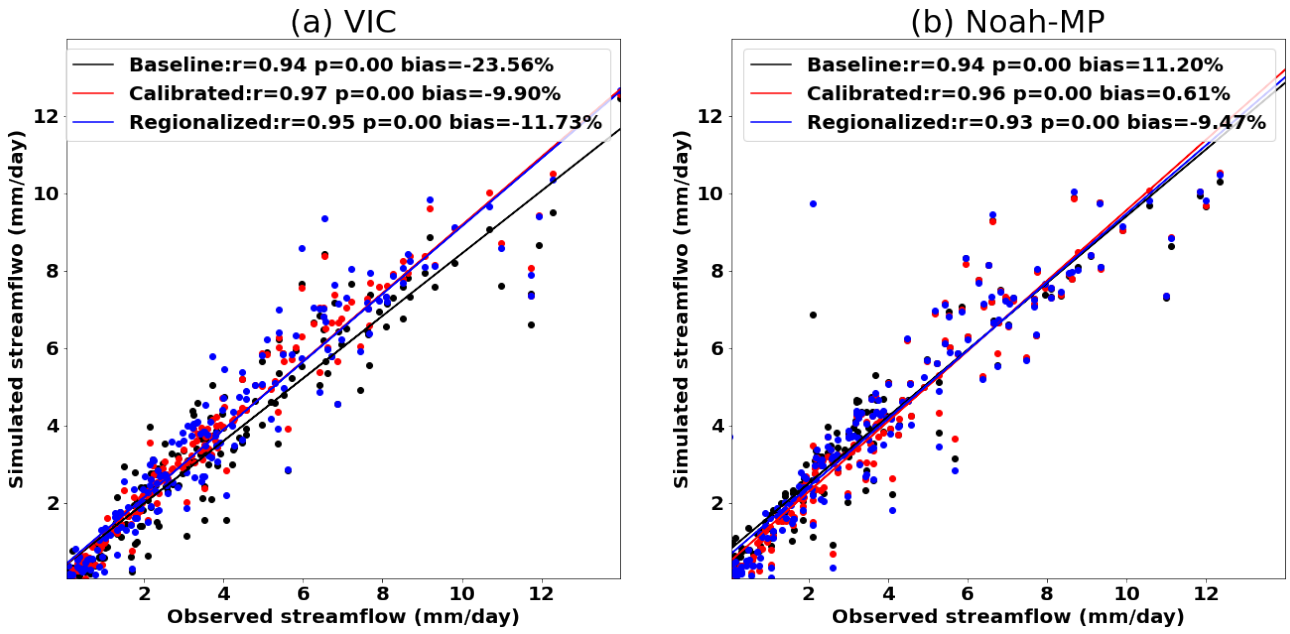



Figure 4.11 Scatterplot of 7q10 low flows (the lowest 7-day average flow that occurs (on average) once every 10 years) for the baseline and calibrated and regionalized runs for (a) VIC model and (b) Noah-MP. The correlation coefficients, P-values and percentage bias are denoted in the upper section of the figures. The x axis is observed low flow and the y axis is simulated low flow.

### 4.6 Summary

Our objective was to produce parameter sets for VIC and Noah-MP over WUS that could be used in regional studies, and would result in better model performance than default or other "off the shelf" parameters. We identified preferred runoff generation options for Noah-MP (physics options are fixed in VIC) using a subset of our WUS basins (50 in total) for which we evaluated all four Noah-MP runoff generation options. Once we identified the optimal runoff generation options for Noah-MP, we identified (calibrated) parameters for both Noah-MP and VIC for each of our 263 basins across WUS using the most recently available 20-years of streamflow observations. Following calibration, the Noah-MP median KGE increased from 0.22 to 0.54, while the median VIC KGE rose from 0.37 to 0.70. VIC KGEs were higher than Noah-MP's both before and after calibration across the 263 basins, possibly because the initial VIC parameters had the

benefit of some previous calibration, albeit for much larger river basins across WUS (in the case of post-calibration KGE, it's unclear whether and how they might have been affected by the choice of initial parameters). Other possible cause of the differences could be inherent differences in streamflow simulation physics between the two models. We also conducted a test using the initial 10 years of data for calibration and the following 10 years for validation, and found results that were consistent with those we obtained using the entire 20 years for calibration.

Upon the selection of suitable parameterizations for Noah-MP and calibration of gauged basins for both VIC and Noah-MP, we extended the use of the calibrated parameters to ungauged basins across the WUS for both models. This extension was achieved through the donor-basin regionalization method, which allows ungauged basins to inherit parameters from gauged basins with similar hydroclimatic properties. We discovered that using a weighted combination of three similar basins yielded better regionalization results (in terms of KGE) compared to using the single most similar donor basin, as determined by a similarity index. Following regionalization, the median KGE for VIC rose from 0.41 to 0.49, and for Noah-MP it increased from 0.38 to 0.49 over the selected basins. Interestingly, even though the pre-regionalization KGE for VIC was considerably higher than for Noah-MP, the post-regionalization values for the two models were nearly identical. Stated otherwise, the regionalization enhancement was considerably greater for Noah-MP than for VIC. We further evaluated high and low flow simulation skills and found the skill significantly improved after calibration for both VIC and Noah-MP and improvements remained after regionalization. Following calibration and regionalization, we developed gridded parameter sets for both models at 1/16° latitude-longitude resolution for all 4816 HUC-10 basins across the WUS. These parameter sets should be useful for regional hydrologic and river hydrodynamic modeling studies over all or parts of the WUS domain. Improving the accuracy of

the models' predictions should have benefits for water management across the region, and more and more generally for understanding the potential impacts of climate change across the region. Moreover, the methods and procedures we utilized are not restricted to our current research domain; they could be transferred readily to other geographic regions. In effect, our research contributes to both local and global efforts to understand and manage our critical hydrological systems better, demonstrating its broader relevance and utility.

**Acknowledgements**

We would like to thank Yuan Yang at UCSD for her sharing of the SCE-UA codes. This work used the COMET supercomputer at UCSD.

**References**

Adam, J.C. and Lettenmaier, D.P.: Adjustment of global gridded precipitation for systematic bias, J. Geophys. Res., 108(D9), 1-14, doi:10.1029/2002JD002499, 2003.

Adam, J.C., Clark, E.A. , Lettenmaier, D.P. and Wood, E.F.: Correction of Global Precipitation Products for Orographic Effects, J. Clim., 19(1), 15-38, doi: 10.1175/JCLI3604.1, 2006.

Anghileri, D., Voisin, N., Castelletti, A., Pianosi, F. , Nijssen, B. and Lettenmaier, D.P.: Value of Long-Term Streamflow Forecasts to Reservoir Operations for Water Supply in Snow-Dominated River Catchments. Water Resources Research 52: 4209–25, 2016.

Arsenault, R., and Brissette, F. P.: Continuous streamflow prediction in ungauged basins: The effects of equifinality and parameter set selection on uncertainty in regionalization approaches. Water Resour. Res., 50, 6135–6153, https://doi.org/ 10.1002/2013WR014898, 2014.

Bass, B., Rahimi, S., Goldenson, N., Hall, A., Norris, J. and Lebow, Z.J.: Achieving Realistic Runoff in the Western United States with a Land Surface Model Forced by Dynamically Downscaled Meteorology. Journal of Hydrometeorology, 24(2), 269-283, 2023.

Bennett, A. R., Hamman, J. J. and Nijssen, B.: MetSim: A Python package for estimation and disaggregation of meteorological data. J. Open Source Software, 5, 2042, https://doi.org/10.21105/joss.02042, 2020.

Beven, K.: Changing ideas in hydrology-the case of physically-based models. Journal of Hydrology, 105(1-2), 157–172. https://doi. org/10.1016/0022-1694(89)90101-7, 1989.

Bohn, T. J., Livneh, B., Oyler, J. W., Running, S. W., Nijssen, B. and Lettenmaier, D. P.: Global evaluation of MTCLIM and related algorithms for forcing of ecological and hydrological

models. Agric. For. Meteor., 176, 38–49, https://doi.org/10.1016/j.agrformet.2013.03.003, 2013.

Boucher, M.-A., and Ramos, M.-H.: Ensemble Streamflow Forecasts for Hydropower Systems. In Handbook of Hydrometeorological Ensemble Forecasting, edited by Q. Duan, F. Pappenberger, J. Thielen, A. Wood, H.L. Cloke, and J.C. Schaake, 1–19. Berlin Heidelberg: Springer, 2018.

Burn, D. H., and Boorman, D. B.: Estimation of hydrological parameters at ungauged catchments. J. Hydrol., 143,429454, https://doi.org/10.1016/0022-1694(93)90203-L, 1993.

Cai, X., Yang, Z.-L. , David, C. H., Niu, G.-Y. and Rodell, M.: Hydrological evaluation of the Noah-MP land surface model for the Mississippi River Basin. J. Geophys. Res. Atmos., 119, 23–38, https://doi.org/10.1002/2013JD020792, 2014.

California Department of Water Resources: California data exchange center: Daily full natural flow for December 2022. California Department of Water Resources, accessed 1 October 2021, https://cdec.water.ca.gov/reportapp/javareports?name= FNF, 2021.

Cao, Q., Mehran, A. , Ralph, F. M. and Lettenmaier, D. P.: The role of hydrological initial conditions on atmospheric river floods in the Russian River basin. J. Hydrometeor., 20, 16671686, https://doi.org/10.1175/JHM-D-19-0030.1, 2019.

Cao, Q., Gershunov, A., Shulgina, T., Ralph, F. M. , Sun, N. and Lettenmaier, D. P.: Floods due to atmospheric rivers along the U.S. West Coast: The role of antecedent soil moisture in a warming climate. J. Hydrometeor., 21, 1827–1845, https:// doi.org/10.1175/JHM-D-19-0242.1, 2020.

Castiglioni, S., Lombardi, L., Toth, E. , Castellarin, A. and Montanari, A.: Calibration of rainfall-runoff models in ungauged basins: A regional maximum likelihood approach. Advances in

Water Resources, 33(10), 1235–1242. https://doi.org/10.1016/j.advwatres.2010.04.009, 2010.

Chen, F., and Dudhia, J.: Coupling an advanced land surface–hydrology model with the Penn State–NCAR MM5 modeling system. Part I: Model implementation and sensitivity. Mon. Wea. Rev., 129, 569–585, https://doi.org/10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2, 2001.

Chen, F., and Coauthors: Modeling of land-surface evaporation by four schemes and comparison with FIFE observations. J. Geophys. Res., 101, 7251–7268, https://doi.org/10.1029/95JD02165, 1996.

Cosby, B.J., Hornberger, G.M., Clapp, R.B. and Ginn, T.: A statistical exploration of the relationships of soil moisture characteristics to the physical properties of soils. Water resources research, 20(6), 682-690, 1984.

Dickinson, R. E., Henderson-Sellers, A. & Kennedy, P. J.: Biosphere–Atmosphere Transfer Scheme (BATS) version 1e as coupled to the NCAR Community Climate Model. NCAR Tech. Note TN383+STR, NCAR, 1993.

Duan, Q., Sorooshian, S. and Gupta, V. : Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resour. Res., 28, 1015–1031, https://doi.org/ 10.1029/91WR02985, 1992.

Environmental Protection Agency (EPA) Office of Water: Low Flow Statistics Tools: A How-To Handbook for NPDES Permit Writers. EPA-833-B-18-001, 2018.

Falcone, J.: GAGES-II: Geospatial attributes of gages for evaluating streamflow. U.S. Geological Survey, accessed 1 April 2021, https://water.usgs.gov/GIS/metadata/usgswrd/XML/ gagesII_Sept2011.xml, 2011.

Federal Institute of Hydrology: “SOSRHINE.” http://sosrhine.euporias.eu/en/sosrhine_overview, 2020.

Fisher, R.A. and Koven, C.D.: Perspectives on the future of land surface models and the challenges of representing complex terrestrial systems. Journal of Advances in Modeling Earth Systems, 12(4), p.e2018MS001453, 2020.

Gochis, D. and Coauthors: Overview of National Water Model Calibration: General strategy and optimization. National Center for Atmospheric Research, accessed 1 January 2023, 30 pp., https://ral.ucar.edu/sites/default/files/public/9_RafieeiNasab_CalibOverview_CUAHSI_Fall019_0.pdf, 2019.

Gong, W., Duan, Q., Li, J., Wang, C., Di, Z., Dai, Y., et al.: Multi-objective parameter optimization of common land model using adaptive surrogate modeling. Hydrology and Earth System Sciences, 19(5), 2409–2425. https://doi.org/10.5194/hess-19-2409-2015, 2015.

Gou, J., Miao, C., Duan, Q., Tang, Q., Di, Z., Liao, W., Wu, J. and Zhou, R.: Sensitivity analysis-based automatic parameter calibration of the VIC model for streamflow simulations over China. Water Resources Research, 56(1), e2019WR025968, 2020.

Gupta, H. V., et al.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of Hydrology, 377, 80-91,2009.

Holtzman, N.M., Pavelsky, T.M., Cohen, J.S., Wrzesien, M.L. and Herman, J.D.: Tailoring WRF and Noah-MP to improve process representation of Sierra Nevada runoff: Diagnostic evaluation and applications. Journal of Advances in Modeling Earth Systems, 12(3), p.e2019MS001832, , 2020.

Hussein, A.: Process-based calibration of WRF-hydro model in unregulated mountainous basin in Central Arizona. M.S. thesis, Ira A. Fulton Schools of Engineering, Arizona State University,

110 pp., https://keep.lib.asu.edu/_flysystem/fedora/c7/224690/Hussein_asu_0010N_19985.pdf, 2020.

Kimball, J. S., Running, S. W. and Nemani, R. R.: An improved method for estimating surface humidity from daily minimum temperature. Agric. For. Meteor., 85, 87–98, https://doi.org/10.1016/S0168-1923(96)02366-0, 1997.

Lahmers, T.M., et al.: Evaluation of NOAA national water model parameter calibration in semiarid environments prone to channel infiltration. Journal of Hydrometeorology, 22(11), 2939-2969, 2021.

Li, D., Lettenmaier, D. P., Margulis, S. A. and Andreadis, K.: Theroleofrain-on-snowinflooding over the conterminous United States. Water Resour. Res., 55, 8492–8513, https://doi.org/10.1029/2019WR024950, 2019.

Liang, X., Lettenmaier, D. P. , Wood, E. F. and Burges S. J. : A simple hydrologically based model of land surface water and energy fluxes for general circulation models, J. Geophys. Res., 99(D7), 14415–14428, doi:10.1029/94JD00483, 1994.

Livneh B, Rosenberg, E.A. , Lin, C. , Nijssen, B. , Mishra, V. , Andreadis, K. , Maurer, E.P. and Lettenmaier, D.P.: A long-term hydrologically based data set of land surface fluxes and states for the conterminous United States: Updates and extensions, Journal of Climate, doi:10.1175/JCLI-D-12-00508.1, 2013.

Maidment, D.R.: Conceptual Framework for the National Flood Interoperability Experiment. Journal of the American Water Resources Association 53: 245–57, 2017.

Mascaro, G., Hussein, A., Dugger, A. and Gochis, D.J.: Process-based calibration of WRF-Hydro in a mountainous basin in southwestern US. Journal of the American Water Resources Association, 59(1), 49-70, 2023.

Mendoza, P.A., Clark, M.P., Mizukami, N., Newman, A.J., Barlage, M., Gutmann, E.D., Rasmussen, R.M., Rajagopalan, B., Brekke, L.D. and Arnold, J.R.: Effects of hydrologic model choice and calibration on the portrayal of climate change impacts. Journal of Hydrometeorology, 16(2), 762-780, 2015.

Miller, D.A. and White, R.A.: A conterminous United States multilayer soil characteristics dataset for regional climate and hydrology modeling. Earth interactions, 2(2), pp.1-26, 1998.

Naeini, M.R., Analui, B., Gupta, H.V., Duan, Q. and Sorooshian, S.. Three decades of the Shuffled Complex Evolution (SCE-UA) optimization algorithm: Review and applications. Scientia Iranica, 26(4), pp.2015-2031, 2019.

Niu, G.-Y., Yang, Z.-L. , Dickinson, R. E. , Gulden, L. E. and Su, H.: Development of a simple groundwater model for use in climate models and evaluation with gravity recovery and climate experiment data. J. Geophys. Res., 112, D07103, https://doi.org/10.1029/2006JD007522, 2007.

Niu, G. Y., Yang, Z. L., Dickinson, R. E., & Gulden, L. E.: A simple TOPMODEL-based runoff parameterization (SIMTOP) for use in global climate models. Journal of Geophysical Research: Atmospheres, 110(D21), 2005.

Niu, G.-Y., and Coauthors: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. J. Geophys. Res., 116, D12109, https://doi.org/10.1029/2010JD015139, 2011.

NOAA (National Oceanic and Atmospheric Administration): National Water Model: Improving NOAA's Water Prediction Services, 2016.

Oubeidillah, A. A., Kao, S.-C., Ashfaq, M. , Naz, B. S. and Tootle, G.: A large-scale, high-resolution hydrological model parameter data set for climate change impact assessment for

the conterminous US. Hydrology and Earth System Sciences, 18(1), 67–84. https://doi.org/10.5194/hess-18-67-2014, 2014.

Prata, A.J.: A new long-wave formula for estimating downward clear-sky radiation at the surface. Quarterly Journal of the Royal Meteorological Society, 122(533), 1127-1151, 1996.

Poissant, D., Arsenault, A. and Brissette, F. : Impact of parameter set dimensionality and calibration procedures on streamflow prediction at ungauged catchments. J. Hydrol. Reg. Stud., 12,220–237, https://doi.org/10.1016/j.ejrh.2017.05.005, 2017.

Qi, W.Y., Chen, J. , Li, L. , Xu, C.-Y. , Xiang, Y.-h. , Zhang, S.-B. and Wang, H.-M.: Impact of the number of donor catchments and the efficiency threshold on regionalization performance of hydrological models. J. Hydrol., 601, 126680, https://doi.org/10. 1016/j.jhydrol.2021.126680, 2021.

Raff, D., Brekke, L. , Werner, K. , Wood, A. and White. K.: Short-Term Water Management Decisions: User Needs for Improved Climate, Weather, and Hydrologic Information. U.S. Bureau of Reclamation. https://www.usbr.gov/research/st/roadmaps/WaterSupply.pdf, 2013.

Razavi, T., and Coulibaly, P.: An evaluation of regionalization and watershed classification schemes for continuous daily streamflow prediction in ungauged watersheds. Can. Water Resour. J., 42,2–20, https://doi.org/10.1080/07011784.2016. 1184590, 2017.

Schaake, J. C., Koren, V. I., Duan, Q.-Y., Mitchell, K., & Chen, F.: Simple water balance model for estimating runoff at different spatial and temporal scales. Journal of Geophysical Research, 101(D3), 7461–7475. https://doi.org/10.1029/95JD02892, 1996.

Schaperow J.R, Li, D., Margulis, S.A., Lettenmaier D.P. :A near-global, high resolution land surface parameter dataset for the variable infiltration capacity model. Scientific Data. Aug 11;8(1):216, 2021.

Sharma, P. and Machiwal, D.: Streamflow forecasting: overview of advances in data-driven techniques. Advances in Streamflow Forecasting,1-50. https://doi.org/10.1016/B978-0-12-820673-7.00013-5, 2021

Shi, X., Wood, A.W. and Lettenmaier, D.P. : How essential is hydrologic model calibration to seasonal streamflow forecasting? Journal of Hydrometeorology, 9(6), 1350-1363, 2008.

Sofokleous, I., Bruggeman, A., Camera, C. and Eliades, M.: Grid-based calibration of the WRF-Hydro with Noah-MP model with improved groundwater and transpiration process equations. Journal of Hydrology, 617, 128991 , 2023

Su, L., Cao, Q. , Xiao, M., Mocko, D. M., Barlage, M. , Li, D. , Peters-Lidard, C. D. and Lettenmaier, D. P.: Drought variability over the conterminous United States for the past century. J. Hydrometeor., 22, 1153–1168, https://doi.org/10.1175/JHM-D-20-0158.1, 2021.

Su, L., Cao, Q. , Shukla, S., Pan, M. and Lettenmaier, D.P.: Evaluation of Subseasonal Drought Forecast Skill over the Coastal Western United States. Journal of Hydrometeorology, 24(4), 709-726, 2023.

Tangdamrongsub, N.: Comparative Analysis of Global Terrestrial Water Storage Simulations: Assessing CABLE, Noah-MP, PCR-GLOBWB, and GLDAS Performances during the GRACE and GRACE-FO Era. Water, 15(13), p.2456, 2023.

Thornton, P. E., and Running, S. W.: An improved algorithm for estimating incident daily solar radiation from measurements of temperature, humidity, and precipitation. Agric. For. Meteor., 93, 211–228, https://doi.org/10.1016/S0168-1923(98) 00126-9, 1999.

Tolson, B. A., and Shoemaker, C. A.: Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. Water Resour. Res., 43, W01413, https://doi.org/10.1029/2005WR004723, 2007.

Troy, T. J., Wood, E. F. and Sheffield, J.: An efficient calibration method for continental-scale land surface modeling. Water Resources Research, 44, W09411. https://doi.org/10.1029/2007WR006513, 2008

USWRC: Guidelines for determining flood flow frequency. Bulletin 17B of the Hydrology Subcommittee, 183 pp., https:// water.usgs.gov/osw/bulletin17b/dl_flow.pdf, 1982.

Yang, Y., Pan, M., Beck, H.E. , Fisher, C.K., Beighley, R.E. , Kao, S.C. , Hong, Y. and Wood, E.F.: In quest of calibration density and consistency in hydrologic modeling: Distributed parameter calibration against streamflow characteristics. Water Resources Research, 55(9), 7784-7803, 2019.

Yang, Z.-L., and Dickinson R. E. : Description of the BiosphereAtmosphere Transfer Scheme (BATS) for the soil moisture workshop and evaluation of its performance, Global Planet. Change, 13, 117–134, doi:10.1016/0921-8181(95)00041-0, 1996.

Zheng, H., Yang, Z.-L. , Lin, P. , Wei, J. , Wu, W.-Y., Li, L. , Zhao, L. and Wang, S.: On the sensitivity of the precipitation partitioning into evapotranspiration and runoff in land surface parameterizations. Water Resour. Res., 55, 95–111, https://doi.org/10.1029/2017WR022236, 2019.

## Chapter 5 Conclusion

This dissertation has aimed to provide a better understanding of the potential of subseasonal drought and National Water Model-based flood forecasting capabilities. Furthermore, it produced calibrated parameter sets for two widely used hydrological models across Western U.S., which should facilitate advances in regional hydrological modeling and prediction endeavors. In Chapter 1, I posed three objectives that guided my research:

(1) Investigate subseasonal forecast accuracy for drought onset and termination using SubX reforecasts.

(2) Assess the flood forecasting capabilities of the National Water Model (Noah-MP) and compare it with current NWS River Forecast Center forecasts.

(3) Develop high-resolution calibrated parameters across the Western U.S. for two widely used LSMs: the Variable Infiltration Capacity (VIC) model and Noah-MP.

To address these questions, I conducted experiments using the Noah-MP and VIC hydrological models over the western U.S. and analyzed drought and floods forecast skill as well as streamflow simulation performance across the Western U.S..

In chapter 2, I examined the performance of SubX-driven forecasts of droughts in the coastal Western U.S. with leads from 1 to 4 weeks. I began by assessing SubX reforecasts for precipitation and temperature in the cool season months October–March as precipitation is generally much lower over most of our domain in the warm season. My results indicated high accuracy for precipitation forecasts in the initial week, which decreased rapidly in subsequent weeks with little usable forecast skill by weeks 3 and 4. Temperature forecast accuracy, however, while declining with lead, remained high with an anomaly correlation coefficient of about 0.4 even at weeks 3 and 4 for most forecast models. When evaluating multi-model ensemble averages, both precipitation

and temperature demonstrated skill were enhanced relative to individual models. These observations align with previous research, e.g., Cao et al. (2021).

After applying statistical downscaling (disaggregating from one-degree to 1/16th-degree resolution) and bias correction of the forcings, I ran the Noah-MP LSM over the coastal Western U.S. for the period 1999–2016. I then assessed the skill of SubX-based drought forecasts with a focus on drought termination and onset using multiple metrics. My evaluation covered both major droughts and more modest drought events. From my analysis, the first two weeks showed notable forecast capabilities for major D0-D2 droughts; however, by the third week, only D0-D1 droughts had some discernible skill, and by the fourth week, the predictability was nearly non-existent. As the severity of droughts increased, the forecast accuracy decreased. I found the skill was consistently higher for drought termination compared to onset across all drought magnitudes. Additionally, there was a geographical trend: moving from the southern to the northern part of the domain, the drought forecast accuracy improved for all event categories. This was possibly due to a similar improving pattern of SubX precipitation forecast skill from south to north over the coastal western U.S. for most of the models and at all lead times.

In Chapter 3, I explored Noah-MP flood simulation skill for coastal western U.S. river basins by selecting optimal model parameterizations and employing a global calibration approach. This elevated the flood simulation skill metric (Kling-Gupta Efficiency or KGE) from approximately 0.2-0.3 to 0.7-0.9. Using calibrated parameters and NWS Quantitative Precipitation Forecast (QPF) model forcings, I compared retrospective Noah-MP flood forecast skill with that of archived forecasts produced by the two Pacific Coast RFCs (Pacific Northwest and California Nevada). Both Noah-MP and RFC showed decreasing ability to forecast flood peak magnitudes with increased lead time, probably reflecting trends in precipitation forecast accuracy. Noah-MP

exhibits a more competitive performance in the northern basins with relative peak differences within approximately +/- 0.1 when the forecast lead is less than 66 hours prior to the peak, which is comparable to RFC skills. Conversely, in the southern basins, Noah-MP's performance is inferior with relative peak differences typically between -0.1 and -0.25 with leads less than 60 hours before the peak, which is consistently below that of CNRFC skills typically in the +/- 0.1 range. Alongside the magnitude of the flood peak, we also examined the peak time forecast skill. Noah-MP demonstrated high accuracy in predicting flood peak timings in the northern region, with most errors restricted to within about a 6-hour window. However, its skill diminishes in the southern basins for lead times longer than 48 hours. This decline in accuracy in the southern basins might be linked to insufficient data on initial hydrologic conditions (soil moisture) in these drier basins. Furthermore, the free drainage runoff generation physics used in this research may not be best suited for all basins. In essence, this chapter suggests that Noah-MP could produce usable flood forecasts, especially for northern basins, given the right parameter and calibration choices. However, there is no indication that Noah-MP forecasts are inherently superior to those currently produced by NWS. Furthermore, enhancements would be necessary for the southern (drier) basins.

In chapter 4, I produced high-resolution calibrated parameters for the VIC and Noah-MP models over the Western U.S.. The general process I followed included multiple stages, specifically identification of model parameters to be calibrated, the calibration (optimization) process itself, and the extension to ungauged basins. By calibrating a selected subset of basins with varied Noah-MP runoff parameterizations, I determined the best model physics selections. Subsequently, I produced calibrated parameters for both Noah-MP and VIC LSMs using autocalibration methods. Post-calibration data showed substantial improvements resulting from calibration: the median KGE across the 263 basins increased from 0.22 to 0.54 for Noah-MP, and

from 0.37 to 0.70 for VIC. Notably, VIC outperformed Noah-MP, which could be attributed to its initial parameters being previously adjusted for major U.S. basins (although not for the specific calibration basins). In contrast, Noah-MP utilized default initial parameters. Differences in the calibration techniques and intrinsic streamflow generation physics could also play a role in post-calibration model performance, a topic I intend to delve into in future research.

Following calibration, I extended the calibrated parameters to ungauged basins across the western U.S. for both models using the donor-basin regionalization method (Poissant et al., 2017; Gochis et al., 2019), which allows ungauged basins to inherit parameters from gauged basins with similar hydroclimatic properties. Following regionalization, the median KGE for VIC rose from 0.41 to 0.49, and for Noah-MP it increased from 0.38 to 0.49 (these numbers are for the selected basins with KGE higher than 0.3 for both models). Intriguingly, Noah-MP's enhancement was more pronounced than VIC's, prompting questions about possible regionalization nuances — this is a topic I plan to investigate further in future studies.

After calibration and regionalization, I was able to develop hydrologic parameters for the models at a high precision 1/16° latitude-longitude resolution in every HUC10 basin across the western United States. These optimized parameters should enhance the accuracy of hydrological model application, aiding in accurate water resource predictions and flood risk assessments. While my work was focused on the Western U.S., the methods I employed are applicable globally, signifying their widespread value in understanding and managing hydrological systems.

This dissertation presents what I believe are important findings and improvements in drought and flood forecasting as well as streamflow simulation. I examined the proficiency of subseasonal drought forecasting, revealing credible skill within the initial two weeks. I found a pronounced accuracy in the northern regions compared to the south and a more discernible skill in

predicting drought termination than its onset. I assessed the potential of Noah-MP flood forecast skill and found it comparable to RFC in northern Pacific Coast basins while inferior in southern ones. I showed that Noah-MP has the potential to produce flood forecasts with accuracy comparable to current NWS methods for northern flood predictions, whereas refinements are necessary for its southern applications. I also developed calibrated hydrologic parameters for VIC and Noah-MP at a high precision 1/16° latitude-longitude resolution across the western U.S.. These optimized parameters improve hydrological modeling, paving the way for more precise predictions related to water resources and flood risks.

Moving forward, my objectives include:

(1) Delving into the disparities in Noah-MP flood forecast efficiencies between the northern and southern basins and striving to enhance its southern performance.

(2) Customizing runoff parameterizations in flood reforecasts tailored to the unique characteristics of individual basins.

(3) Investigating and refining the regionalization techniques employed for Noah-MP and VIC in streamflow predictions.

**References**

Cao, Q., S. Shukla, M. J. DeFlorio, F. M. Ralph, and D. P. Lettenmaier, 2021: Evaluation of the Subseasonal Forecast Skill of Floods Associated with Atmospheric Rivers in Coastal Western US Watersheds. Journal of Hydrometeorology, 22(6), 1535-1552.

Gochis, D., and Coauthors, 2019: Overview of National Water Model Calibration: General strategy and optimization. National Center for Atmospheric Research, accessed 1 January 2023, 30 pp.https://ral.ucar.edu/sites/default/files/public/9_RafieeiNasab_CalibOverview_CUAHSI_Fall019_0.pdf

Poissant, D., A. Arsenault and F. Brissette, 2017: Impact of parameter set dimensionality and calibration procedures on streamflow prediction at ungauged catchments. J. Hydrol. Reg. Stud., 12,220–237, https://doi.org/10.1016/j.ejrh.2017.05.005.

## Appendix A

### Introduction

This document includes the description of the Noah-MP options used in this study. It also includes the evaluation of the hydrological model dependance and calibration effects. In addition, we show an alternate of the evaluation of drought forecast skill at subregion scale.

### Text S1. Noah-MP options used in this study

We adopted the WRF-HYDRO recommended physical options as in https://ral.ucar.edu/sites/default/files/public/Noah-MPOptionsIndicatorsofusagewithWRFHydroNWM.pdf

We list the details of our options below.

*DYNAMIC_VEG_OPTION(options for dynamic vegetation)*

*4 -> off (use table LAI; use maximum vegetation fraction)*

*CANOPY_STOMATAL_RESISTANCE_OPTION (options for canopy stomatal resistance)*

*1 -> Ball-Berry*

*BTR_OPTION(options for soil moisture factor for stomatal resistance)*

*1 -> Noah (soil moisture)*

*RUNOFF_OPTION (options for runoff and groundwater)*

*3 -> original surface and subsurface runoff (free drainage)*

*SURFACE_DRAG_OPTION (options for surface layer drag coeff (CH & CM)*

*1 -> M-O*

*SUPERCOOLED_WATER_OPTION (options for supercooled liquid water (or ice fraction))*

*1 -> no iteration (Niu and Yang, 2006 JHM)*

*FROZEN_SOIL_OPTION (options for frozen soil permeability)*

*1 -> linear effects, more permeable (Niu and Yang, 2006, JHM)*

*RADIATIVE_TRANSFER_OPTION (options for radiation transfer)*

*3 -> two-stream applied to vegetated fraction (gap=1-FVEG)*

*SNOW_ALBEDO_OPTION(options for ground snow surface albedo)*

*1 -> BATS*

*PCP_PARTITION_OPTION (options for partitioning precipitation into rainfall & snowfall)*

*1 -> Jordan (1991)*

*TBOT_OPTION(options for lower boundary condition of soil temperature)*

*2 -> TBOT at ZBOT (8m) read from a file (original Noah)*

*TEMP_TIME_SCHEME_OPTION(options for snow/soil temperature time scheme (only layer 1) )*

*1 -> semi-implicit; flux top boundary condition*

*SURFACE_RESISTANCE_OPTION (options for surface resistent to vaporization/sublimation)*

*1 -> Sakaguchi and Zeng, 2009*

*GLACIER_OPTION(options for glacier treatment)*

*1 -> include phase change of ice*

**Text S2 Hydrological model dependance and calibration effects evaluation**

To evaluate the model dependency of our study, we ran Variable Infiltration Capacity (VIC) V4.1.2.d (Liang et al., 1994) with the Livneh et al (2013) forcings (using the same parameters as in Livneh et al (2013)) over the baseline period 1961-2016 with the same spin up period used in our Noah-MP experiments.

To address the calibration effects on drought forecast skill, we included calibrated VIC in comparison with uncalibrated VIC over CA. We don't have a calibrated version of Noah-MP results for now (this is a topic of future work). In all the simulations that we show below (VIC calibrated and uncalibrated and Noah-MP) we performed a long spin-up (1951-2016 twice before initiating the 1961-2016 analysis period) as in the runs reported for Noah-MP in the main text.

We took the calibrated VIC parameters from (currently unpublished) ongoing work. We applied the Shuffled Complex Evolution (SCE-UA) calibration method (Duan et al, 1994) for ~100 basins across CA which resulted in median daily KGE improvement from 0.48 to 0.75. After that,

we applied the donor-basin method to regionalize the calibrated parameters to all of CA (Arsenault and Brissette, 2014; Yang et al. 2018; Qi et al. 2021).

We identified historical droughts and compared the dry area over CA for Noah-MP, uncalibrated VIC and calibrated VIC baseline experiments (Figure A1). This figure shows that the dry area from Noah-MP, uncalibrated VIC and calibrated VIC all have similar patterns.

To further evaluate the relationship of droughts constructed from different hydrological models, we calculated the Spearman correlation coefficients and Nash–Sutcliffe model efficiency coefficients (NSE) between baseline drought area (1961-2016) based on Noah-MP and VIC (uncalibrated) (Figure A2). We see high correlations (>0.8) in all regions (OR, WA and CA). The NSE can be as high as 0-6-0.8 for D1 and it decreases as the drought severity increases but still is higher than 0.4.

To further evaluate the effects of calibration, we compared the correlation and NSE between Noah-MP, uncalibrated VIC [denoted as VIC in the figure] and calibrated VIC [denoted as calibVIC in the figure] dry areas and the USDM dry area in CA in Figure A3. The results show (a) uncalibrated VIC and calibrated VIC show very high correlation and NSE; (b) Noah-MP and VIC (we only show Noah-MP with calibrated VIC for easy reading) show high correlation and NSE; (c) Noah-MP has higher correlation and NSE with USDM than VIC does; (d) Calibrated VIC shows higher metrics with USDM than uncalibrated VIC but the improvement is limited.

To further evaluate the drought forecast dependency on hydrological model and calibration, we compared the drought onset forecast skill when using Noah-MP and VIC based on the EMC-GEFS model (11 SubX ensemble members). The computation procedure is expensive and time consuming, so here we limited ourselves to just the EMC-GEFS (which has the largest number of ensemble members among all six SubX models). The results show that Noah-MP and VIC

generated comparable drought onset skills for different drought levels at lead week1 (Figure A4). We repeated the procedure for calibrated VIC which showed no improvement in performance due to calibration. This is predictable since we calculate drought characteristics based on soil moisture percentiles, which reduced the effect of calibration. This is supported by Shi et al (2008) who found that the reduction in seasonal streamflow forecast error that is achieved by bias correction alone is nearly as great as that resulting from hydrologic model calibration. They concluded that calibration didn't make much difference to seasonal hydrologic forecast skill – the skill basically is inherent in the forcings to the model, and not the model construct (or parameters) itself.

Overall, we conclude that, while there are some differences between models and before and after calibration, our results are not strongly dependent on the specific model and calibration.

**Text S3. Drought forecast skill evaluation at subregion scale.**

To reduce noise spatially, we tried an alternative to the method described in section 4.3 to assess the drought forecast skill for different subregions. We first averaged the soil moisture by subregion to produce a single forecast at each subregion, then constructed droughts at the subregion level and evaluated the BSS prediction skills (Figure A5). (We chose BSS rather than POD here because the correction term in the BSS calculation considers the effects of small sample sizes). We found positive skill in most cases (except for some cases in drought termination in southern CA and drought onset in central and southern CA). Drought onset shows relatively lower skill than drought termination.

We found generally decreasing forecast skill as the lead time increases, consistent with similar behavior in precipitation (and temperature) prediction skill. There are a few exceptions, for example, D3 has higher drought termination skill at lead week 3 than the other lead weeks. When averaged spatially, the forecast skill is not necessarily higher for less severe droughts than more

severe ones in certain regions. For example, D3 shows higher drought termination skill than D0, D1 and D2 in WA at lead weeks 1, 2 and 3. This might be caused by the different level of noise cancellation after spatial averaging. Despite of these exceptions, the subregion BSS values are generally consistent with what we found in grid cell based BSS (Figures 2.8, 2.9).

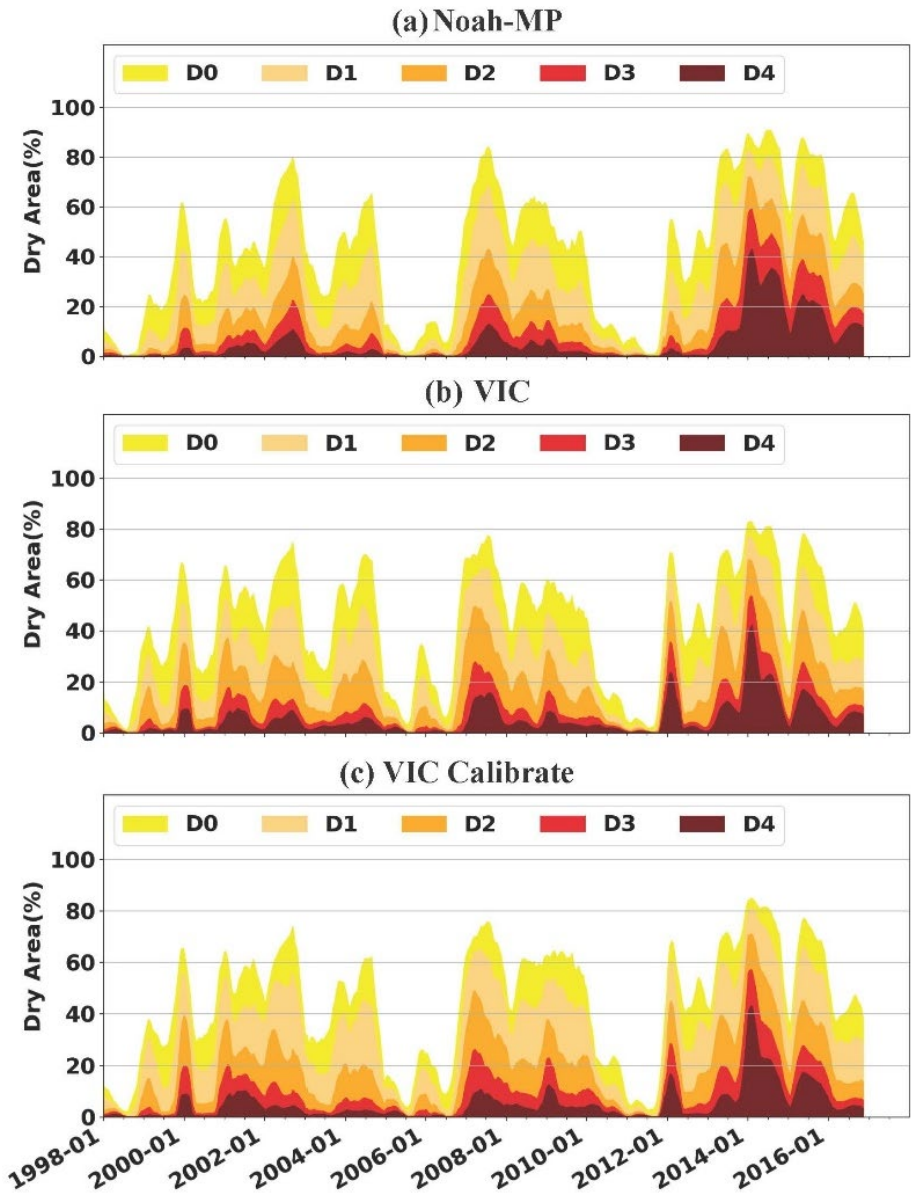
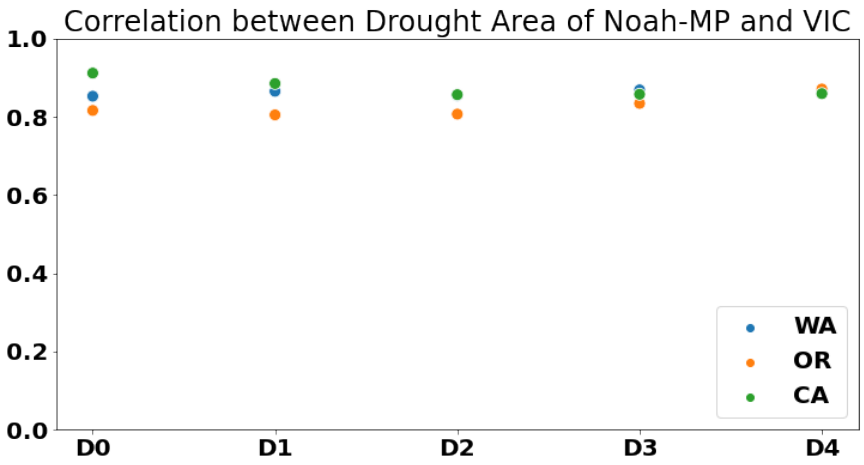



Figure A1 California dry area of Noah-MP, uncalibrated VIC and calibrated VIC.

(a)




(b)

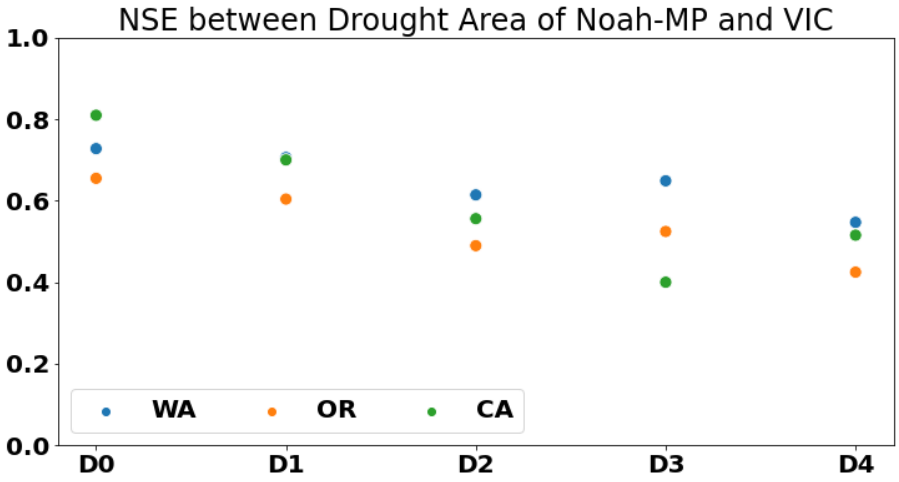
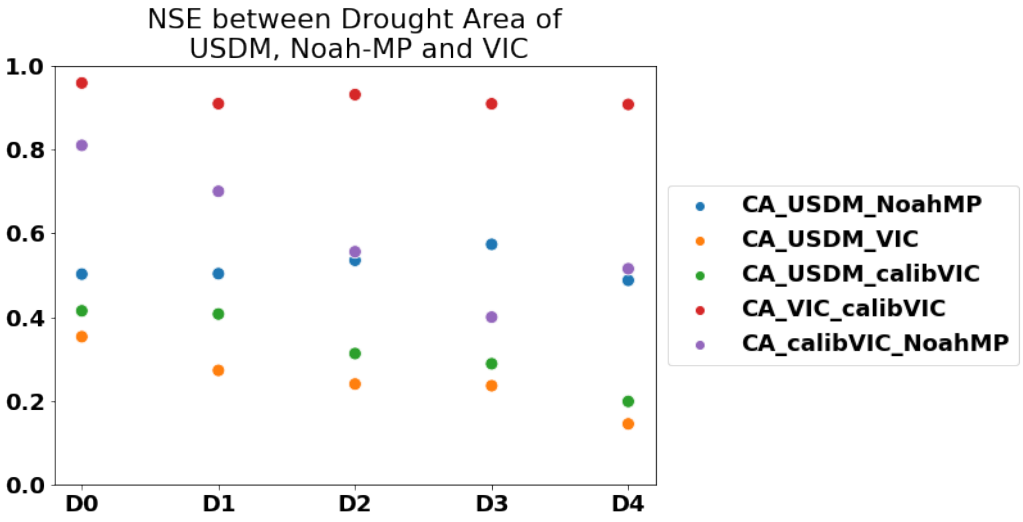



Figure A2 Spearman correlation coefficient (a) and NSE (b) between drought area (1961-2016) from Noah-MP and VIC for OR, WA and CA.

(a)




(b)

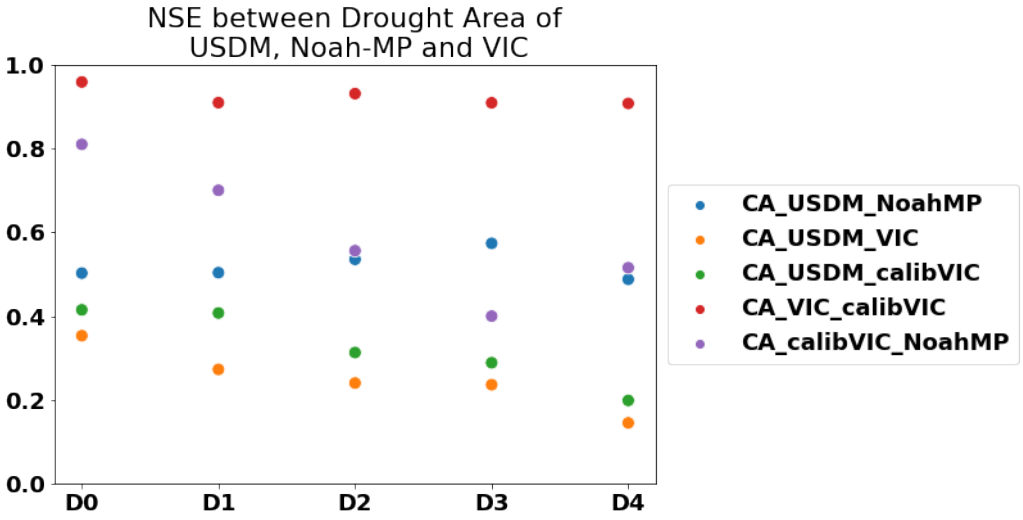



Figure A3 Spearman correlation coefficient and NSE between drought area (1961-2016) from USDM, Noah-MP and VIC for CA.

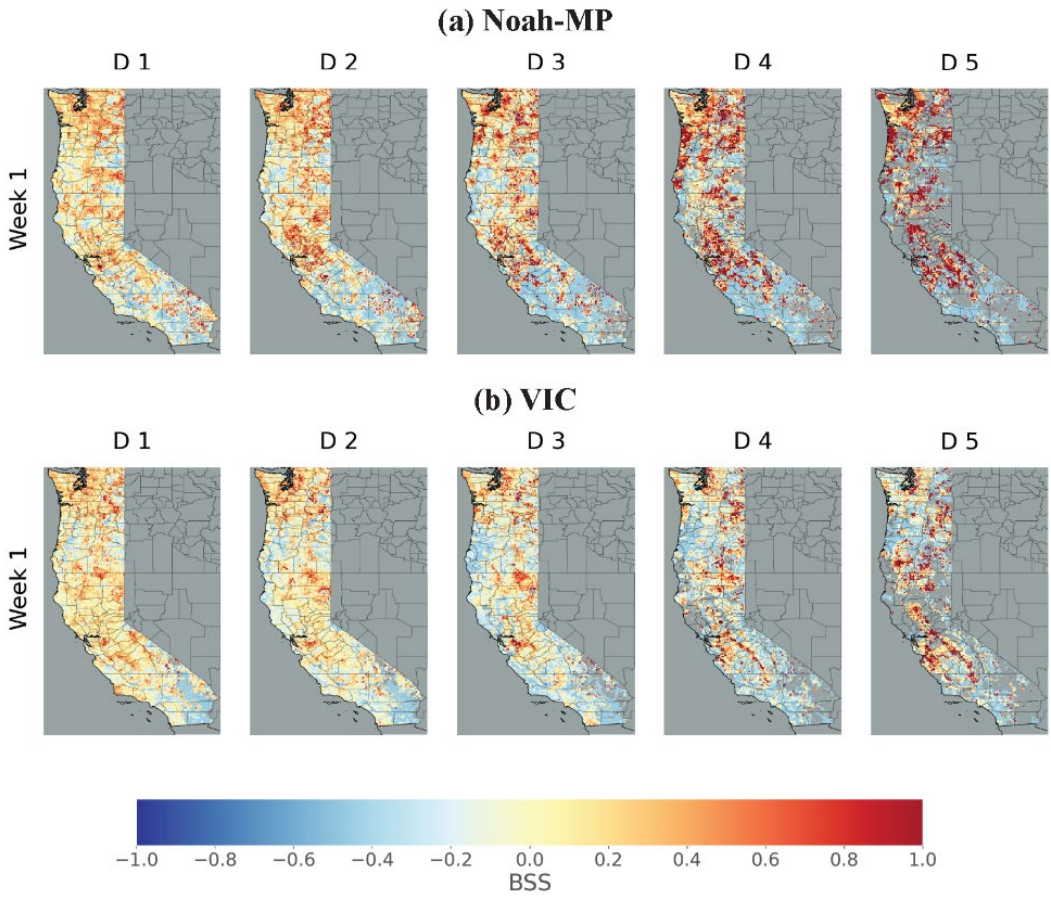



Figure A4 EMC-GEFS based debiased Brier skill score (BSS) for lead weeks 1-4 for drought onset at lead week 1 based on Noah-MP and VIC.

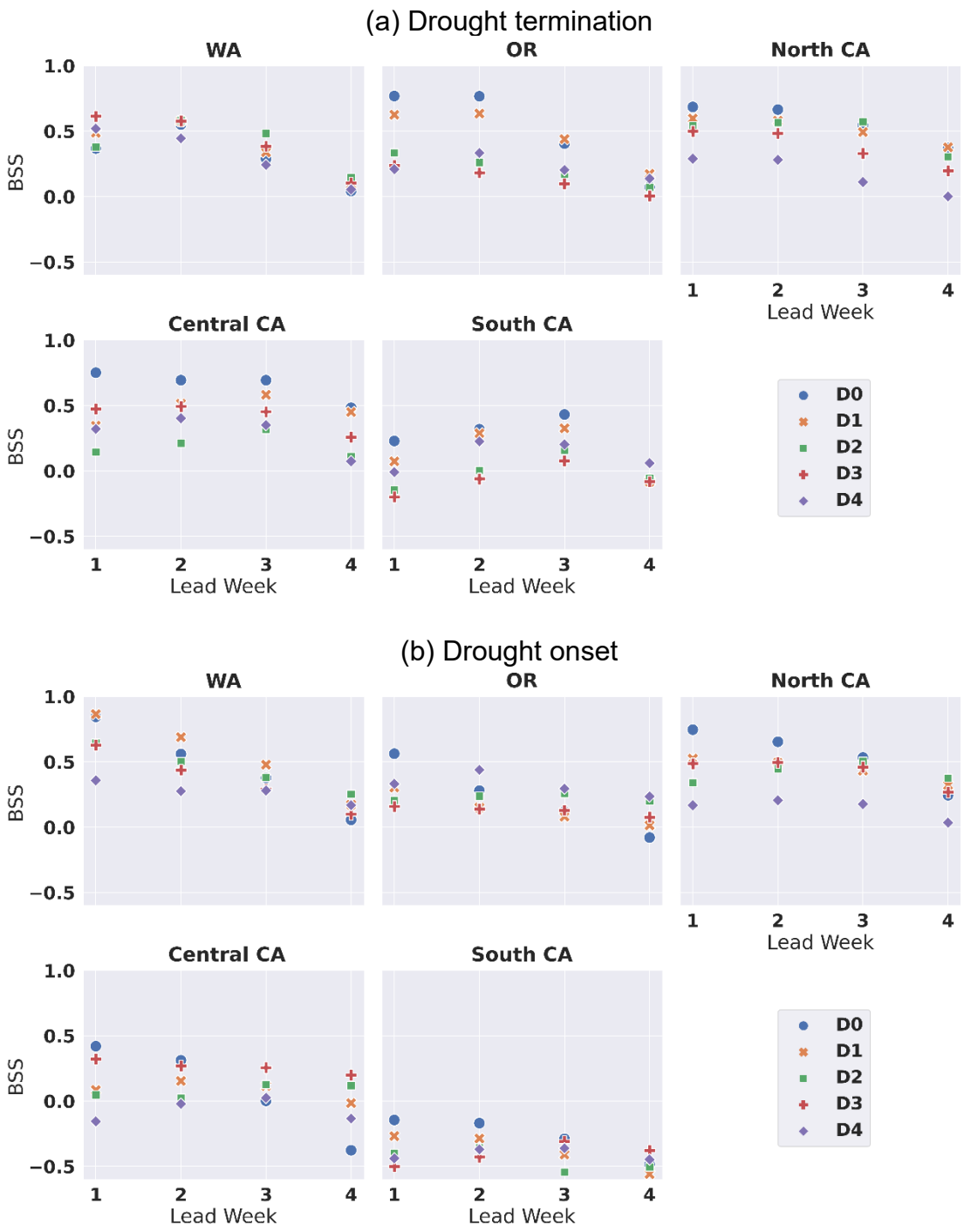



Figure A5 SubX-based debiased Brier skill score (BSS) for lead weeks 1-4 for (a) drought termination, (b) drought onset by drought levels and by subregions.

**References**

Arsenault, R. and F.P. Brissette, 2014: Continuous streamflow prediction in ungauged basins: The effects of equifinality and parameter set selection on uncertainty in regionalization approaches. Water Resour. Res., 50, 6135– 6153, https://doi.org/10.1002/2013WR014898.

Duan, Q., S. Sorooshian, and V. K. Gupta, 1994. Optimal use of the SCE-UA global optimization method for calibrating watershed models. Journal of hydrology, 158(3-4), 265-284.

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

Shi, X., A.W. Wood and D.P. Lettenmaier, 2008: How essential is hydrologic model calibration to seasonal streamflow forecasting? J. of Hydrometeorology 9(6), 1350-1363.

Qi, W.Y., J. Chen, L. Li, C.Y. Xu, Y.H. Xiang, S.B. Zhang, and H.M. Wang, 2021: Impact of the number of donor catchments and the efficiency threshold on regionalization performance of hydrological models. Journal of Hydrology, 601, https://doi.org/10.1016/j.jhydrol.2021.126680.

Yang, X., J. Magnusson, J. Rizzi, C.Y. Xu, C.Y., 2018: Runoff prediction in ungauged catchments in Norway: comparison of regionalization approaches. Hydrology Research 49, 2, 487–505. https://doi.org/10.2166/nh.2017.071.

## Appendix B

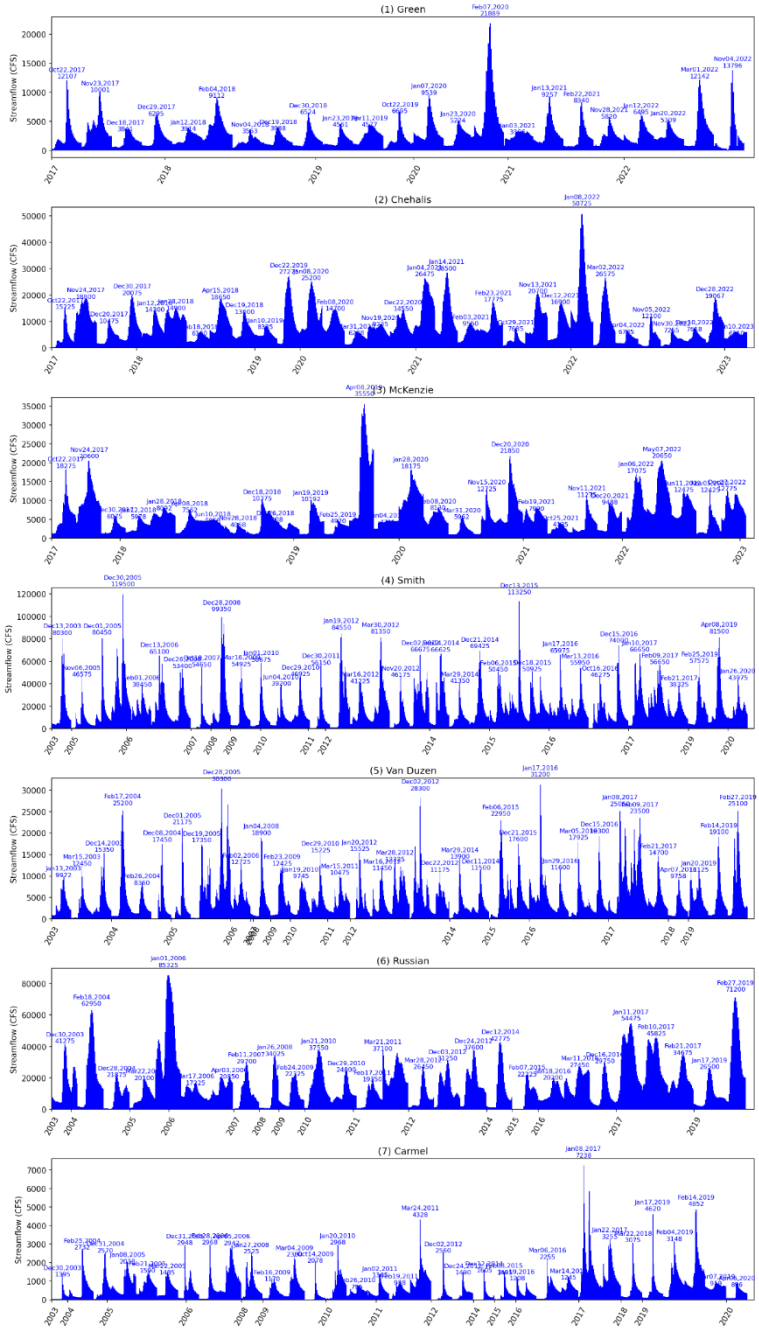



Figure B1 Time series of floods events that analyzed in the study basins. Please note that the time axis is not uniformly distributed. We only show the eight days of each flood event – four days preceding the flood peak time and four days following the flood peak time. The time resolution is 6-hour. The flood peak streamflow and peak date are annotated in the figure.

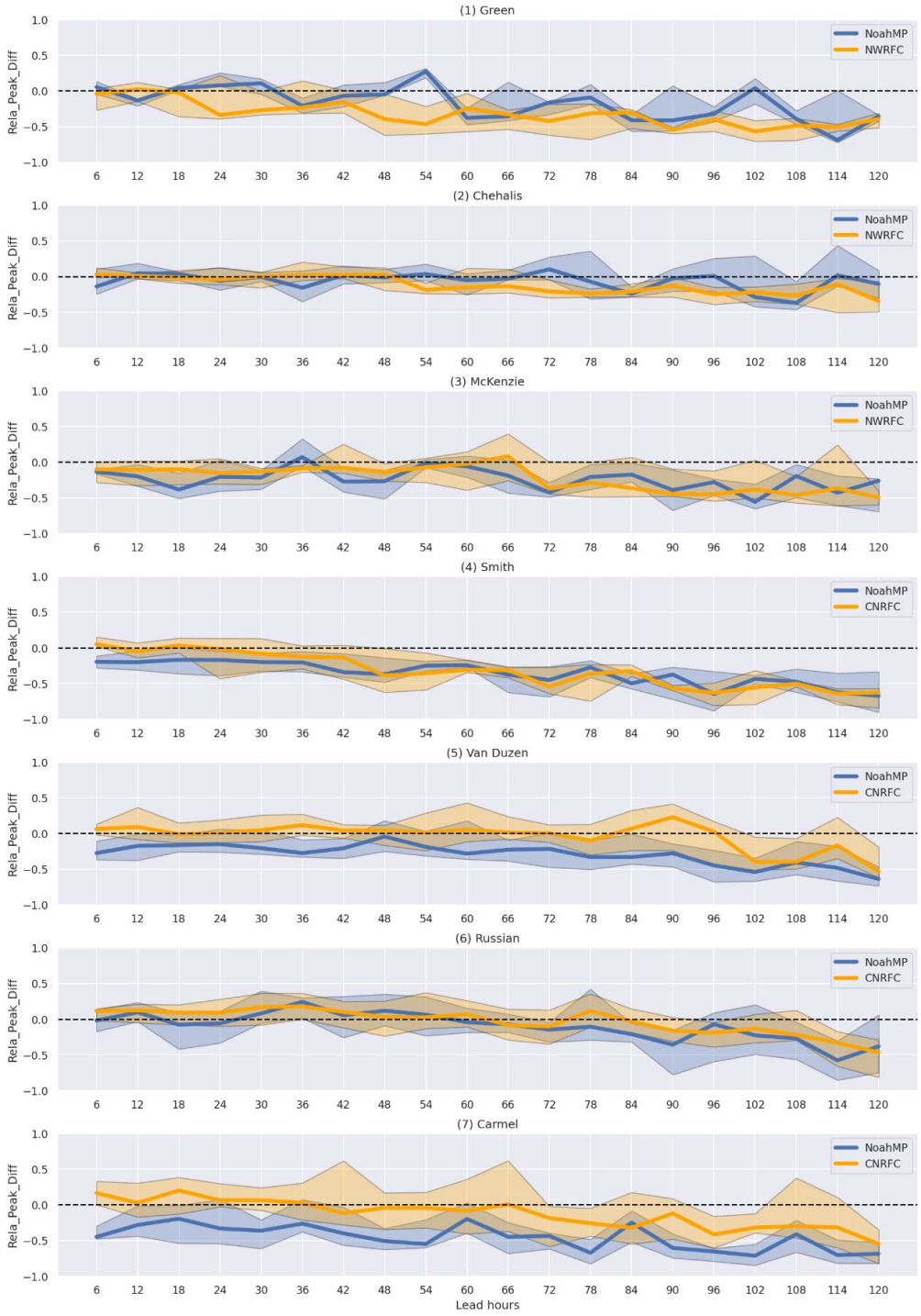



Figure B2 Median and interquartile range of the relative differences of floods peak streamflow of Noah-MP reforecasts and RFC forecasts against lead hours in (1) Green, (2) Chehalis, (3) McKenzie, (4) Smith, (5) Van Duzen, (6) Russian, (7) Carmel rivers.

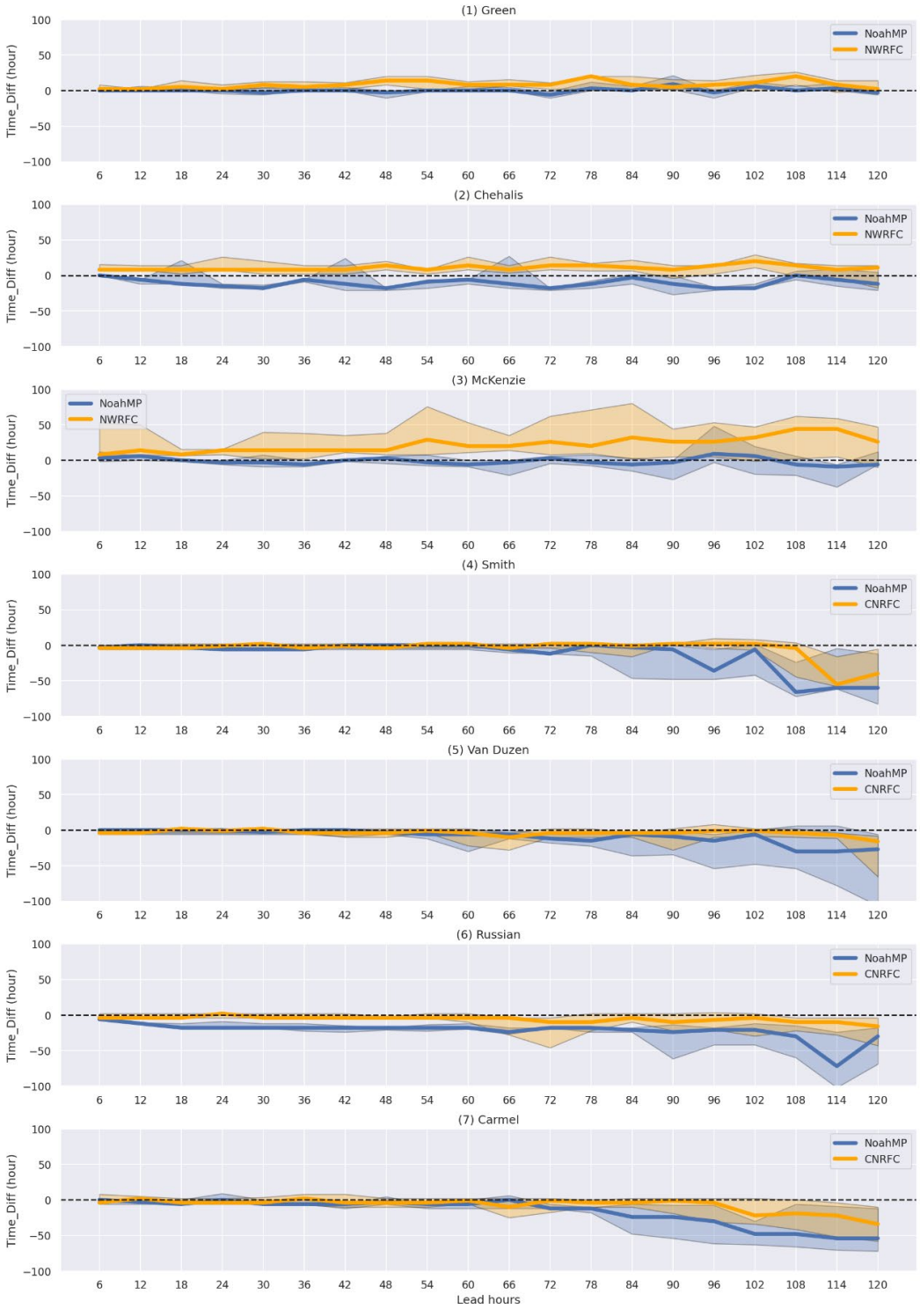



Figure B3 Median and interquartile range of the difference of floods peak time of Noah-MP reforecasts and RFC forecasts against lead hours in (1) Green, (2) Chehalis, (3) McKenzie, (4) Smith, (5) Van Duzen, (6) Russian, (7) Carmel rivers.

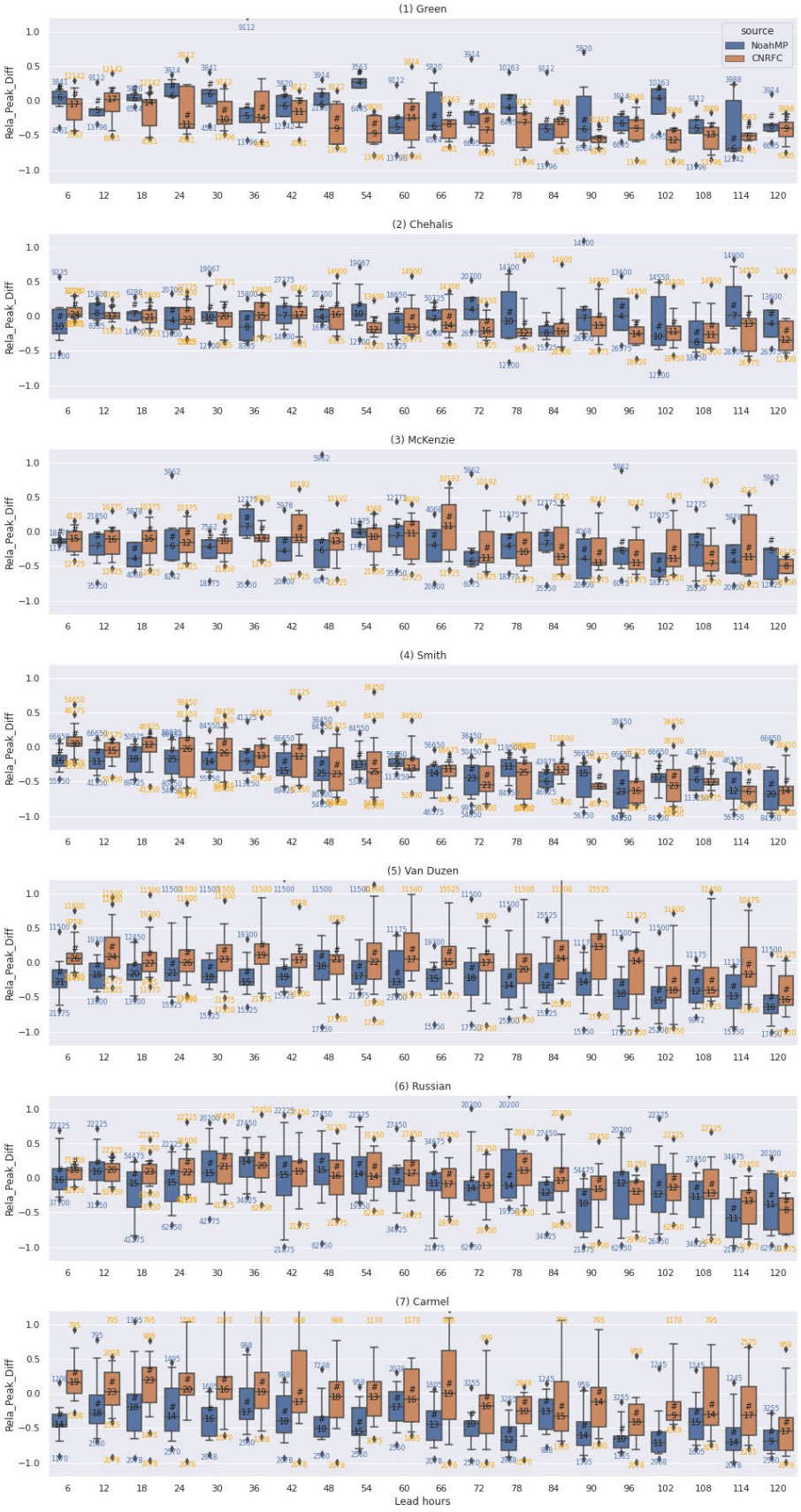

(1) Green
(2) Chehalis
(3) McKenzie
(4) Smith
(5) Van Duzen
(6) Russian
(7) Carmel
source
NoahMP
CNRFC
Rela_Peak_Diff
Lead hours

Figure B4 Boxplots of relative differences of floods peak streamflow of Noah-MP reforecasts and RFC forecasts against lead hours in Smith River basin. The numbers in the box that start with # indicate the number of events summarized in the box. Since the QPF forcing we have mostly initiated at 12:00 or sometimes also at 18:00 while the flood peak time can be anytime between 00:00-24:00, the numbers of flood events calculated at different lead time can vary for Noah-MP reforecasts. The RFC forecasts initialization interval changed for different basins and for different time periods, so the numbers of flood events calculated at different lead times also vary for RFC forecasts. The numbers near the outliers indicate he peak value (in cfs) of the flood that corresponds to the outliers. The blue color is for the Noah-MP and the orange color is for the RFC.

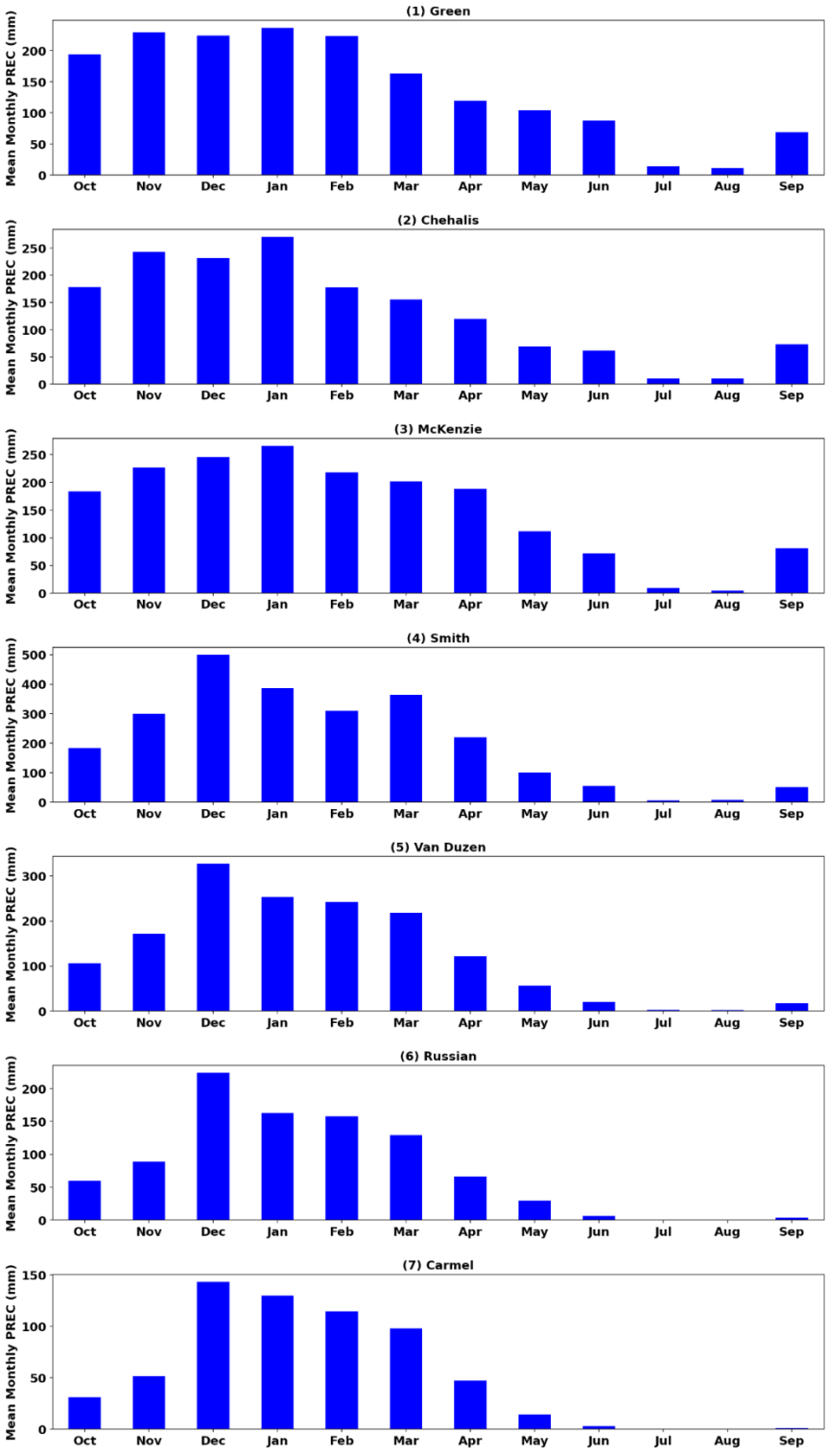



Figure B5 Mean monthly total precipitation (mm) (averaged over the study period) in the seven study basins.

## Appendix C

Table C1 Features considered for regionalization of calibrated parameters to ungauged basins in VIC and Noah-MP models.

| Features in order of rank for VIC | Features in order of rank for Noah-MP |
|---|---|
| Longitude Centroid | Latitude Centroid |
| Latitude Centroid | Longitude Centroid |
| Max Elevation | Area |
| Precipitation Fall Mean | **KGE stops increasing for below features** |
| Temperature Fall Mean | Temperature Summer Mean |
| **KGE stops increasing for below features** | Min Elevation |
| Mean Elevation | Precipitation Annual Mean |
| Min Elevation | Precipitation Fall Mean |
| Temperature Summer Mean | Perimeter |
| Precipitation Spring Mean | Temperature Spring Mean |
| Precipitation Winter Mean | Max Elevation |
| Precipitation Summer Mean | Precipitation Winter Mean |
| Precipitation Annual Mean | Temperature Winter Mean |
| Mean Annual Max 1-D Precipitation | Temperature Annual Mean |
| Temperature Winter Mean | Precipitation Spring Mean |
| Temperature Spring Mean | Precipitation Summer Mean |
| Temperature Annual Mean | Mean Elevation |
| Perimeter | Temperature Fall Mean |
| Area | Mean Annual Max 1-D Precipitation |

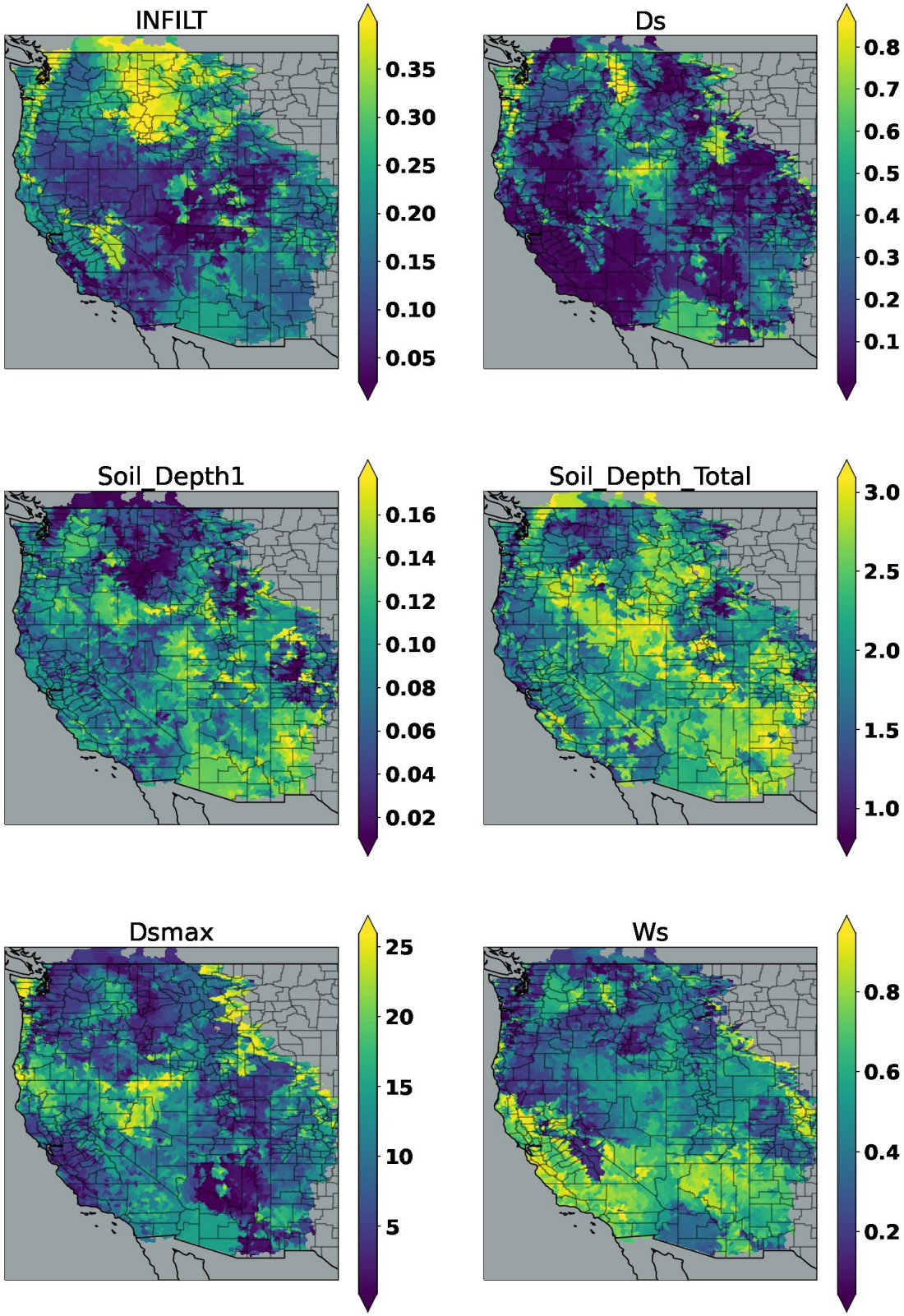



Figure C1 Regionalized VIC Land surface parameters over WUS.

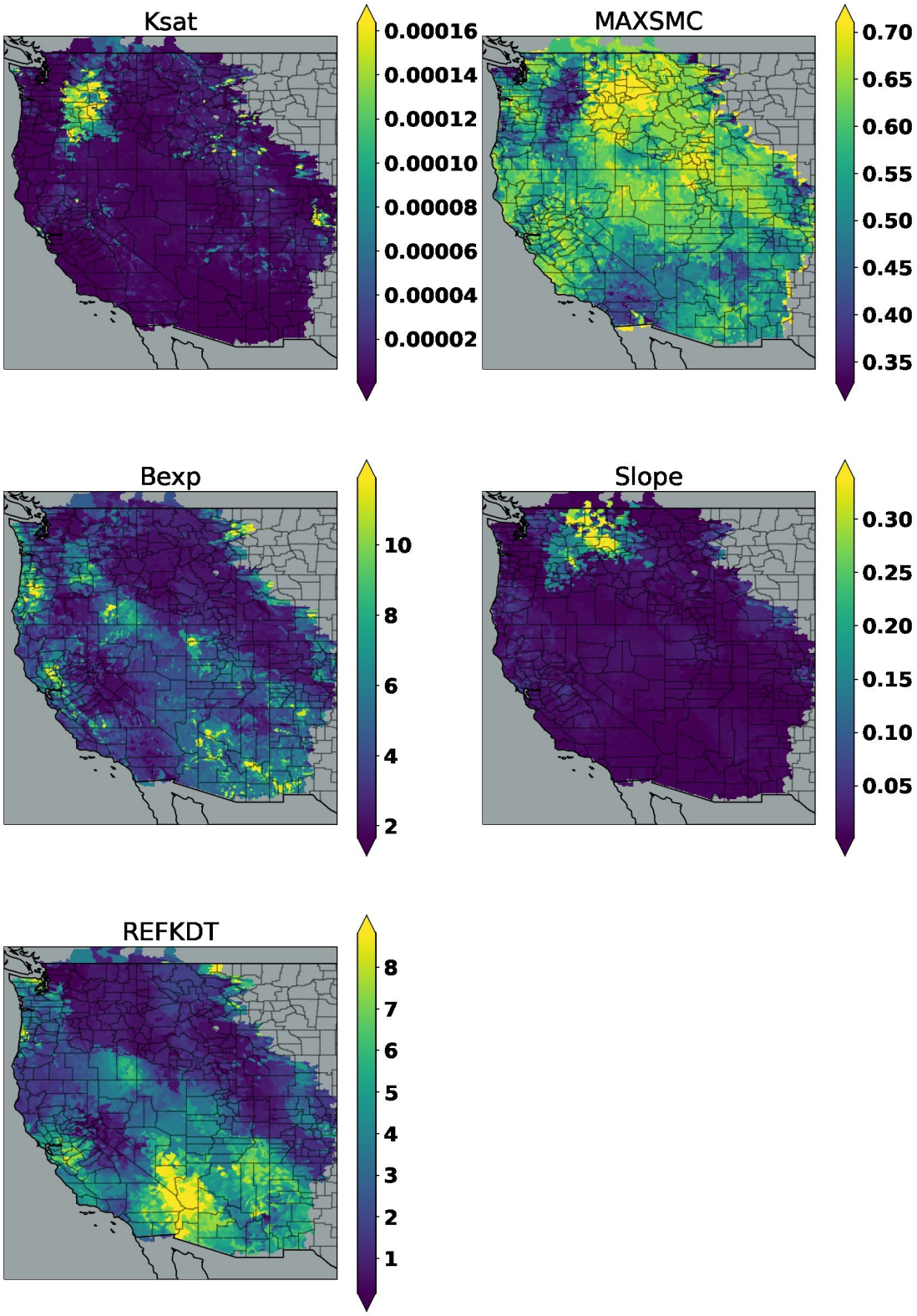



Figure C2 Regionalized Noah-MP Land surface parameters over WUS.

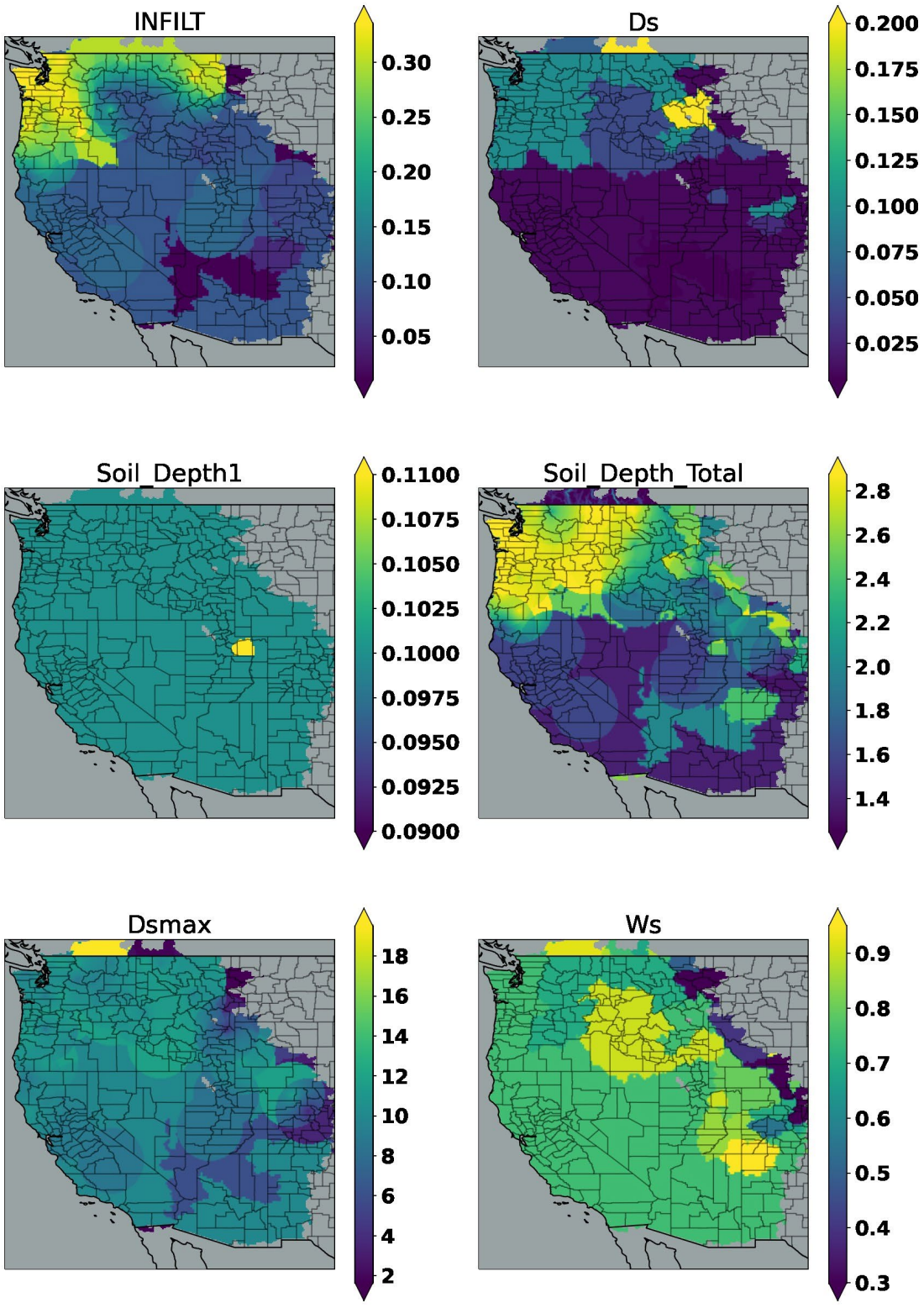



Figure C3 Baseline VIC Land surface parameters over WUS.

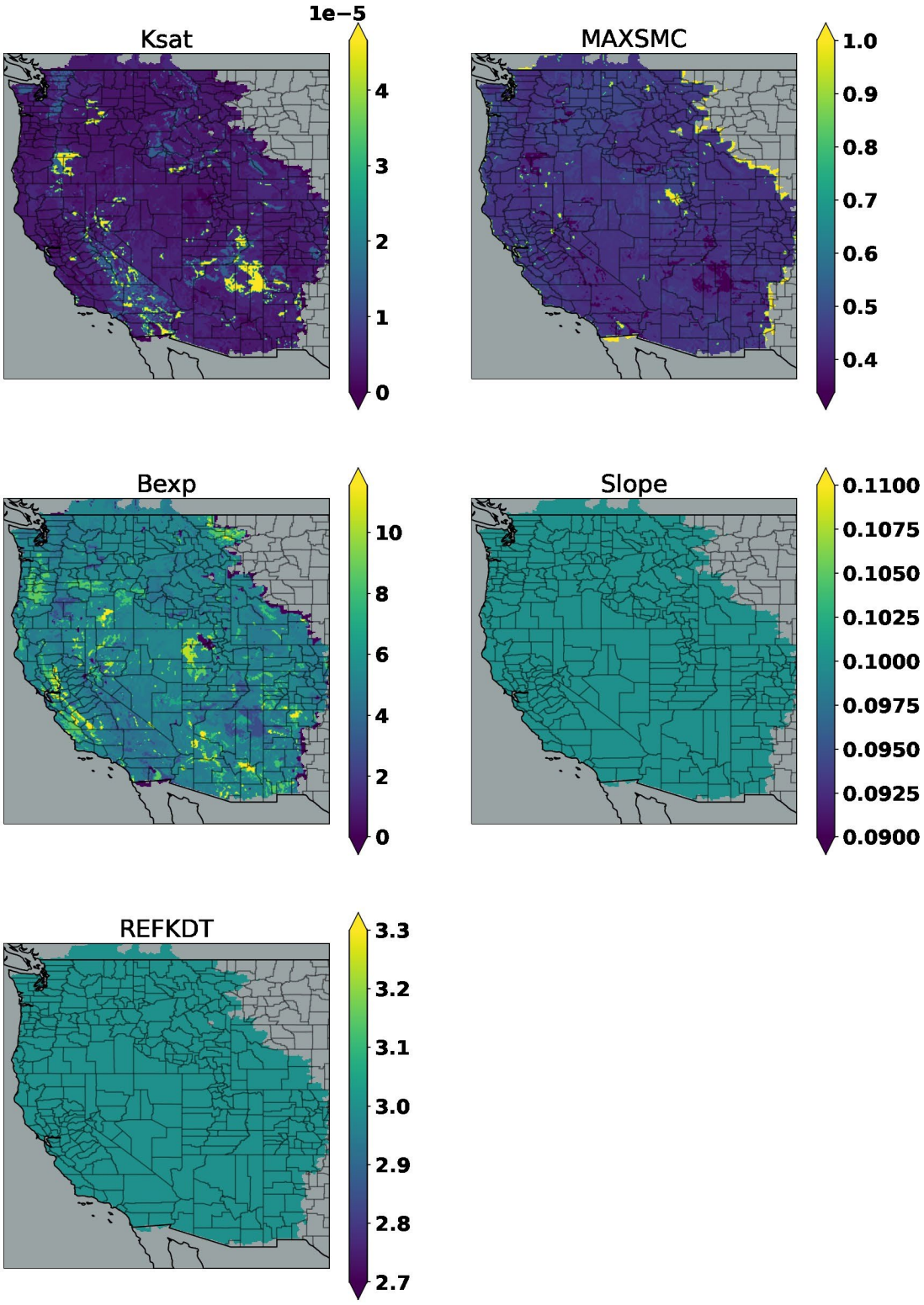



Figure C4 Baseline Noah-MP Land surface parameters over WUS.