Title

Enhancing Hydrological Prediction through Physics-Informed Machine Learning Models and Leveraging Data Science for Predictions in Ungauged Basins

Author

Zhang, Liang

Publication Date

2024

Peer reviewed|Thesis/dissertation

Enhancing Hydrological Prediction through Physics-Informed Machine Learning Models and Leveraging Data Science for Predictions in Ungauged Basins

By

Liang Zhang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Civil and Environmental Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Laurel G. Larsen, Chair
Professor Bin Yu
Professor Fotini Chow

Spring 2024

Abstract

Enhancing Hydrological Prediction through Physics-Informed Machine Learning Models and Leveraging Data Science for Predictions in Ungauged Basins

by

Liang Zhang

Doctor of Philosophy in Engineering - Civil and Environmental Engineering

University of California, Berkeley

Professor Laurel G. Larsen, Chair

Data science is a fundamental tool in hydrology nowadays. The significance of data science lies in its ability to confront the multifaceted challenges posed by global warming, facilitating a deeper comprehension of hydrological processes, and enhancing the accuracy of runoff predictions. This dissertation embarks on a journey aimed at advancing our insights into hydrological processes, refining the physical-consistency of runoff predictions, and addressing the intricate task of forecasting hydrological behaviors in ungauged basins through the application of data science techniques. Comprising three main bodies of work, this dissertation unfolds a comprehensive exploration of these objectives. The first contribution (Chapter 2) centers on the synthesis of extensive hydrological datasets and subsequent analysis of hydrological trends under recent warming. Chapter 3 explores a physics-informed machine learning model designed for predicting streamflow tested across different scenarios. Lastly, the fourth chapter evaluates the potency of various watershed clustering mechanisms for predicting within ungauged basins (PUB).

Chapter 2 addresses a long-standing limitation in comparative hydrology: the scarcity of geographically extensive, inter-compatible monitoring data on comprehensive water balance stores and fluxes. These limitations have, for example, restricted comprehensive assessment of multiple dimensions of wetting and drying related to climate change and hampered understanding of why widespread changes in precipitation extremes are uncorrelated with changes in streamflow extremes. In this chapter, both the requirements of developing a new data synthesis product and using this data product to detect trends in the frequencies and magnitudes of a comprehensive set of hydroclimatic and hydrologic extremes are addressed. The Comprehensive Hydrologic Observatory Sensor Network (CHOSEN), a database encompassing

hydroclimatic and hydrologic variables from 30 diverse study areas across the United States is introduced. And a reproducible data pipeline that ensures data quality and accessibility is developed. Analyzing the CHOSEN dataset, the hotspots of hydroclimatic extremes in regions like the Pacific Northwest, New England, Florida, and Alaska are uncovered. The analysis reveals regional coherence in extreme streamflow wetting and drying trends, shedding light on the complex interplay between climate-induced changes and hydrologic processes.

Chapter 3 is built upon the development of the CHOSEN dataset to create subsequent analyses and a new runoff prediction model. The challenge of a lack of interpretability and physical consistency in machine learning models used for streamflow prediction is confronted. To address this issue, a physics-informed long short-term memory (PILSTM) model is proposed, incorporating water balance restrictions for runoff prediction. A physical rainfall-runoff model is combined with the long short-term memory (LSTM) model, and it is applied to eight intensively-monitored watersheds in the United States, selected based on data length and hydroclimatic diversity. LSTM, physical, and PILSTM models are used under non-stationary scenarios and data-scarce situations. Results show that the PILSTM exhibits similar or better performance to the LSTM counterpart in terms of multiple metrics and under various scenarios. Additionally, based on the analysis of feature importance, it is shown that adding physical constraints could potentially guide machine learning models to generate predictions that are more consistent with known physical processes.

Chapter 4 explores the effectiveness of watershed clustering, a conventional practice in watershed regionalization, in combination with neural networks for predicting in ungauged basins. Traditionally, watershed clustering involves grouping basins with similar characteristics to facilitate knowledge transfer from monitored to ungauged basins within the same cluster. Recent advancements in data science, however, suggest that clustering may not be necessary. This study aims to investigate this matter and presents a comparative analysis of various watershed clustering methodologies. The concept is explored by directly integrating static watershed attributes into predictive models for streamflow (entity-aware LSTM). The analysis covers 415 sites from the CAMELS (Catchment Attributes and Meteorology for Large-sample Studies) dataset. Results indicate that pre-clustering generally does not enhance the performance of entity-aware LSTM models for predicting in ungauged basins. Models incorporating clustering results either match or perform worse overall compared to global models that directly integrate clustering features as static inputs. Notably, among the different features used for clustering, hydrological signatures prove most effective in extracting information for use in the LSTM model.

Chapter 2 addresses crucial gaps in data availability, while the subsequent chapters explore novel approaches for forecasting streamflow across diverse scenarios and ungauged basins, leveraging the power of data science. In Chapter 3, the integration of physical and machine learning models

is pursued, while Chapter 4 focuses on harnessing data science methodologies for predicting in ungauged basins. Collectively, these chapters offer an exploration of the intersection between data science and hydrology. This dissertation emphasizes the transformative potential of interdisciplinary strategies, which bridge data-driven insights with the dynamics of hydrological systems.

Table of Contents

Acknowledgements

First and foremost, I am deeply grateful to have you, Laurel, as my supervisor throughout my PhD journey. Your unwavering support and guidance have been invaluable to my academic research. I still recall the late nights we spent working on paper revisions. I was impressed by your ability to distill conclusions and patterns from numerous results and crafting the discussion section. And also, you serve as a true role model in both academia and life. Despite your busy schedule as a supervisor and also as a mom, you have always made time to provide guidance for my research whenever I needed it. Thank you very much for your support along the way.

I extend my heartfelt thanks to all the members of my lab. Edom, who joined the lab almost at the same time with me, has provided a lot of help during my early PhD journey. From balancing coursework and research to navigating citation practices, your experience as a senior is very helpful to me. Dino, I still remember the occasion when, shortly after we first met, you dedicated an hour in the lab to help me understand a mathematical problem. Your generosity in sharing your time and offering suggestions and advice on my research has been so helpful. To Sam, Rosanna, Hongxu, Saleem, Jordan, Sheila, Omar, and Galen, I am grateful to work with you and spend my time with you all in the lab. I cherish the moments when we come together to prepare for AGU meetings as a team and provide feedback on each others' posters or presentations. Together you form my favorite lab team! I really enjoyed the five years I spent with you.

My sincere gratitude also extends to my committee members: Professor Bin Yu, Tina Chow, Cynthia Gerlein-Safdi, and Yoram Rubin. Bin, your expertise in statistical knowledge and your dedication to fostering good scientific practices have a great influence on my research. I also admire your meticulous spirit in science and research. Tina, your fluid mechanics class not only captivated me but also led me to the opportunity to work with Laurel on her project. Thank you for facilitating that connection, too! Cynthia, your insights and suggestions during my qualifying exam were instrumental in shaping my research direction. I am so happy to have you as one of my committee members. Professor Rubin, your introductory class in hydrology during my first semester at Berkeley sparked my passion for the field, and I am grateful for your guidance.

I am also grateful to my funding sources, the Gordon and Betty Moore Foundation, and the US Geological Survey Powell Center, for their support of my research endeavors.

Lastly, I wish to express my profound gratitude to my parents for their unwavering support throughout my PhD journey, especially to my mom, who not only cares about my study but also/always reminds me to prioritize my well-being every time we have a video chat. Thank you, Haobin, for always being by my side. Make delicious meals for me on my busiest days. And all my PhD friends from UC village, Zhe, Qianhua, Mingyuan, Kaijing, Sili, Yunduan. I appreciate that we can get together from time to time, celebrating holidays or just talking about the

difficulties we recently have in research. With the companions of you all, I feel that I am never alone in my battle! Thank you.

Chapter 1

Introduction

Hydrological research has reached a juncture where both physical models and machine learning models coexist and flourish. Physical models are rooted in the governing equations of hydrological processes and leverage expert knowledge of catchments, while machine learning models directly discern patterns from input data, minimizing a predefined loss function to make predictions. Model training in machine learning is accomplished by minimizing the loss function, which is usually defined as the difference between the outputs of the model and the target values. In contrast, physical models attain performance optimization by calibrating model parameters, a process involving the exploration of various values or curve-fitting.

The parameters derived from physical models hold additional value for watershed characterization, as they often correlate with catchment attributes such as soil properties, vegetation cover, watershed topography, soil moisture content, and characteristics of groundwater aquifers (Werkhoven et al., 2008). These parameters, once determined, contribute to an enhanced understanding of catchment dynamics. Physical models can be further categorized based on their spatial distribution, ranging from simple spatially lumped conceptual models to semi-distributed models and fully distributed models, the latter exhibiting the highest degree of spatial distribution (Fleming & Gupta, 2020). Distributed models function as mechanistic representations of catchments, simulating natural processes continuously in both time and space. Consequently, these models necessitate an extensive dataset comprising information about watershed topography and measurements of numerous variables across space. In contrast, lumped models distill hydrological processes into aggregated components at the catchment level, such as groundwater recharge and snowmelt, combining these elements to calculate discharge.

The use of statistical models in hydrology has evolved significantly, from the century-plus-old use of simple regression models for representing hydrological trends to the more recent use of machine learning models. Sophisticated machine learning models, particularly deep neural networks, have emerged for predictive purposes in hydrology as data science knowledge has advanced and a wealth of hydrological data for analysis has become available, owing to extended measurement networks and increased use of remote sensing products (Lange & Sippel, 2020). Deep learning and machine learning, in particular, have made significant contributions to modeling and predicting hydrological processes, climate change impacts, and earth systems. Recent studies show the critical importance of these strategies in improving hydrology model accuracy, resilience, efficiency, computational cost-effectiveness, and overall model performance (Ardabili et al., 2020).

Despite the demonstrated superior predictive performance of machine learning models compared to physical models in numerous hydrological studies and significant advancements in their application to address hydrological questions (Figure 1), the direct implementation of these models for practical purposes, such as guiding water resource management and flood prediction, lags behind the progress achieved in research (Fleming et al., 2021). A prominent concern revolves around the inherent omission of physical processes and components in data-driven models. Unlike physical models, where hydrologists can discern the physical processes and evaluate the plausibility of predictions based on their understanding, certain machine learning models sacrifice interpretability for predictive power. The challenge is particularly pronounced with deep neural networks, which exhibit remarkable predictive capabilities but prove challenging to interpret and explain. Hydrologists remain cautious about relying solely on these artificial neural networks for predictions that guide practical decision-making, especially when the prediction path and the underlying hydrological behaviors are not well understood. The trade-off between interpretability and predictive power introduces hesitancy in embracing machine learning models for real-world applications in hydrology.

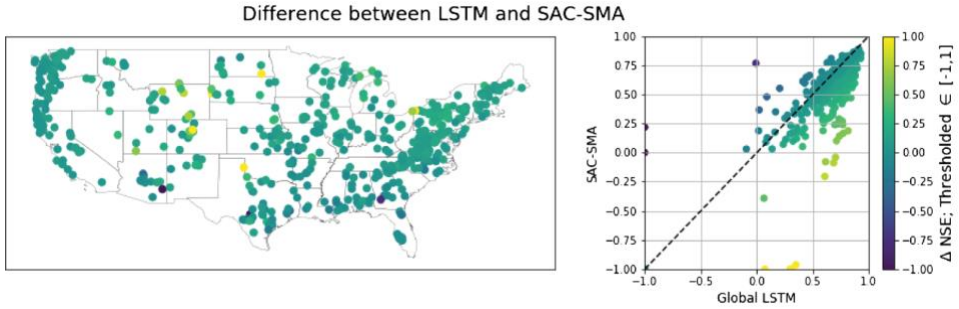


Figure 1. Comparison between the LSTM model and the SAC-SMA (Sacramento Soil Moisture Accounting Model) model for predicting streamflow in CAMELS sites (Kratzert, Klotz, Herrnegger, et al., 2019).

Addressing this challenge and seeking to bridge the gap between machine learning and solving problems in hydrology, a novel research field has emerged known as physics-informed machine learning. Hybrid models, integrating physical knowledge into machine learning frameworks, have been explored across various physics-related domains, including material science, quantum chemistry, biomedical science, turbulence modeling, and earth science (Karpatne et al., 2017; Willard et al., 2020). The infusion of physical insights into machine learning models has resulted in enhancements in model accuracy, interpretability, and robustness for out-of-sample cases (Jia et al., 2019; Konapala et al., 2020; Lu et al., 2021; Xie et al., 2021). This progress opens avenues for the practical implementation of machine learning models. Notably, Frame et al., (2021) applied Long Short-Term Memory (LSTM) models to enhance the US National Water Model (NWM) using two distinct combinations of input variables, showcasing superior performance

compared to the stand-alone NWM in terms of Nash-Sutcliffe Efficiency (NSE) and other metrics.

Regardless of the rapid and advanced development of machine learning models in hydrology, the foundational support for their progress lies in the availability of well-documented and high-quality dataset. A persistent challenge in the hydrological community is the scarcity of high-quality, publicly available, and large-sample datasets. Hydrologists often must invest considerable time in screening and cleaning data before conducting analyses. Nevertheless, two large-sample datasets that have significantly contributed to advancing hydrological studies are the CAMELS dataset (Addor et al., 2017; Newman et al., 2015) and the MOPEX dataset (Duan et al., 2006). These resources are invaluable in supporting comparative hydrological studies and enhancing the understanding of hydrological processes. These hydrological datasets still have limitations, however, such as non-up-to-date data, meteorological inputs being modeled rather than field measurements, and unclear and non-reproducible data preprocessing procedures. For sites where an abundance of data variables are collected in the field, such as sites from Long-Term Ecological Research observatories (LTER) and Critical Zone Observatories (CZO), the data are collected and recorded separately and lack standardized formats for comparative analysis. While efforts such as CUAHSI HydroShare aim to curate these diverse data sources, hydrologists still face challenges in checking, harmonizing formats, and ensuring comparability for analysis when utilizing these data for comparative studies.

Another persistent challenge relevant to input data quality revolves around predicting hydrological phenomena in locations lacking long-term observations altogether, commonly known as PUB (Prediction in Ungauged Basins) (Hrachowitz et al., 2013). Two primary strategies are used to address this challenge. For simple models with a limited number of parameters, researchers have explored relationships between calibrated parameters and catchment attributes. This allows the prediction of model parameters for ungauged basins based on catchment characteristics (bottom-up approach). Regionalizations, grouping watersheds with similar hydrologic signatures and fitting forecasting models for each cluster, achieve a similar goal (top-down approach). Unlike physical models, where parameters are linked to static catchment features, applying the bottom-up method to ungauged basins with machine learning models is not intuitive and especially not practical for highly parameterized models such as deep neural networks. Researchers have successfully pre-trained machine learning models with extensive data from monitored sites and applied them to ungauged basins (Kratzert, Klotz, Herrnegger, et al., 2019), providing an opportunity to employ deep learning models for PUB problems. However, further analysis is required, such as assessing the applicability of the bottom-up approach and exploring how classification and regionalization can enhance machine learning models for PUB.

In light of the challenges present in hydrology, this dissertation aims to leverage data science techniques to improve hydrological predictions and establish a robust data foundation for future research. Chapter 2 involves synthesizing a hydrological dataset from intensively monitored watersheds and implementing a publicly available data preprocessing pipeline (Zhang et al., 2021). Additionally, trends pertaining to climate change across the United States are analyzed using this synthesized hydrological dataset. Chapter 3 introduces a new physics-informed machine learning model that incorporates a single equation hydrological model into LSTM to maintain mass balance. The chapter also explores the pretraining of machine learning models on a large sample dataset and tests the hybrid model across non-stationary and data-scarce scenarios. Integrated gradients are utilized to compare the difference in feature importance with and without physical information. Chapter 4 investigates the effectiveness of classification when employing neural networks for PUB. Three different sets of hydrological features, including static attributes and statistics reflecting dynamic interactions between meteorological time series and hydrological signatures, are tested for their efficacy in watershed clustering.

Chapter 2

CHOSEN: A synthesis of hydrometeorological data from intensively monitored catchments and comparative analysis of hydrologic extremes[1]

2.1 Abstract

Comparative hydrology has been hampered by limited availability of geographically extensive, intercompatible monitoring data on comprehensive water balance stores and fluxes. These limitations have, for example, restricted comprehensive assessment of multiple dimensions of wetting and drying related to climate change and hampered understanding of why widespread changes in precipitation extremes are uncorrelated with changes in streamflow extremes. Here, we address this knowledge gap and underlying data gap by developing a new data synthesis product and using that product to detect trends in the frequencies and magnitudes of a comprehensive set of hydroclimatic and hydrologic extremes. CHOSEN (Comprehensive Hydrologic Observatory Sensor Network) is a database of streamflow, soil moisture, and other hydroclimatic and hydrologic variables from 30 study areas across the United States. An accompanying data pipeline provides a reproducible, semi-automated approach for assimilating data from multiple sources, performing quality assurance and control, gap-filling and writing to a standard format. Based on the analysis of extreme events in the CHOSEN dataset, we detected hotspots, characterized by unusually large proportions of monitored variables exhibiting trends, in the Pacific Northwest, New England, Florida and Alaska. Extreme streamflow wetting and drying trends exhibited regional coherence. Drying trends in the Pacific Northwest and Southeast were often associated with trends in soil moisture and precipitation (Pacific Northwest) and evapotranspiration-related variables (Southeast). In contrast, wetting trends in the upper Midwest and the Rocky Mountains showed few univariate associations with other hydroclimatic extremes, but their latitudes and elevations suggested the importance of changing snowmelt characteristics. On the whole, observed trends are incompatible with a 'drying-in-dry, wetting-in-wet' paradigm for climate-induced hydrologic changes over land. Our analysis underscores the need for more extensive, longer-term observational data for soil moisture, snow and evapotranspiration.

[1] Zhang, L., Moges, E., Kirchner, J. W., Coda, E., Liu, T., Wymore, A. S., et al. (2021). Chosen: A synthesis of hydrometeorological data from intensively monitored catchments and comparative analysis of hydrologic extremes. *Hydrol. Process.* 35 (11). doi:10.1002/hyp.14429

2.2 Introduction

Climatic and hydrologic extremes pose severe risks to human society and infrastructures and trigger irreversible transitions in ecosystems (AghaKouchak et al., 2020; Ainsworth et al., 2020; Hughes et al., 2019; McClymont et al., 2020). The magnitude and frequency of these extremes are increasing as a result of climate change (e.g., Ahn & Palmer, 2016; Pagán et al., 2016; Swain et al., 2018; Wentz et al., 2007), which results from basic physical principles. In accordance with Clausius-Clapeyron scaling, warmer air holds more moisture, which is associated with projected increases in rainfall intensity (Sillmann et al., 2013), the intensity and frequency of tropical cyclones (Marsooli et al., 2019), and the amount of water conveyed in atmospheric rivers (Gao et al., 2015; Payne et al., 2020). Warmer temperatures also increase potential evapotranspiration and are linked to increasing drought severity (Cook et al., 2015; Diffenbaugh et al., 2015). The balance between processes that promote catchment drying (e.g., enhanced evapotranspiration) and those that promote wetting (e.g., increased precipitation extremes) varies among catchments. Therefore, it can be difficult to generalize outcomes of increasing precipitation and temperature extremes for hydrological processes.

The difficulty in predicting how increased climatic extremes will impact hydrologic extremes is particularly apparent in the discrepancy between the projected and observed association between precipitation and discharge extremes. While climate models predict a strong correlation between extreme precipitation and extreme flood magnitude (e.g., Pall et al., 2011), observations show low correlation spatially and temporally (e.g., Archfield et al., 2016; Berghuijs et al., 2016; Blöschl et al., 2017; Do et al., 2020), except for rare floods with recurrence intervals longer than 10 years (Wasko & Nathan, 2019). Specifically, flood trends are not changing in accordance with climate model predictions (Sharma et al., 2018). The need to understand the link between changing precipitation and changing flooding has been argued to be one of the grand challenges in hydrology (Sharma et al., 2018).

Measurements of soil moisture and other variables indicative of water balance stores and fluxes may provide clues critical to reconciling Sharma et al. 's (2018) grand challenge, and, more broadly, understanding how shifting climate translates into a range of hydrological outcomes. Results of modeling and observational studies that derive (Berghuijs et al., 2016; Byun et al., 2019; Heidari et al., 2020; Ivancic & Shaw, 2015) or account for measured soil moisture (Wasko & Nathan, 2019) or changes in subsurface storage (Slater & Villarini, 2016) suggest that changes in hydrologic extremes are attributable to simultaneous shifts in several hydrologic variables, with soil moisture or subsurface storage of critical importance. One gap in these analyses is that, with the exception of Wasko and Nathan's (2019) study of Australian catchments, they rely on simple models or proxies for soil moisture rather than actual measurements. Meanwhile, the role of soil moisture, snow storage, and actual evapotranspiration in governing low-flow extremes remains underexplored. Exploration of causes of hydrologic extremes requires hydrologic

databases that synthesize variables beyond the precipitation, temperature, and streamflow measurements that are more typically available.

Long-term observational records play an important role in understanding and projecting the impact of climate change on hydrological systems. They provide important ground truth for hydroclimatic models, highlighting uncertainties in their representation of certain processes (e.g., rainfall-runoff processes). Trends detected in the observational record are also commonly reliable indicators of future hydroclimatic change (Batibeniz et al., 2020). Despite their potential importance, long-term and spatially extensive databases that contain a range of hydrologic variables relevant to water-balance partitioning (e.g., soil moisture, snow data, vapor pressure deficit) are virtually nonexistent. One reason for limited spatial coverage is that extensive measurements of soil moisture and snow-water content are impractical to measure with gauging stations and uncertain when inferred from current remote sensing techniques, with estimates characterized by limited volumetric representativeness and high uncertainty (Ford & Quiring, 2019). Further, hydrologically comprehensive datasets are available at only a limited, albeit growing, number of catchments, often referred to as hydrologic observatories. Synthesis across these observatories has been hindered by a lack of standardization in variable naming conventions, file formats, time steps, metadata, and data processing procedures, which in turn has slowed the development of the subfield of comparative hydrology (Gupta et al., 2014).

Here we respond to the dearth of long-term, regionally extensive, hydrologically comprehensive databases by presenting CHOSEN (DOI: 10.5281/zenodo.4060384), the Comprehensive Hydrologic Observatory SEnsor Network database, a compilation of publicly available hydrometeorological and hydrological measurements from 30 LTER (Long-Term Ecological Research observatories; Servilla & Brunt, 2011), CZO (Critical Zone Observatories; Zaslavsky et al., 2011), and university field stations in the United States (Kakalia et al., 2021; McNamara, 2017; R. S. Petersky & Harpold, 2018). We developed CHOSEN using a novel operational pipeline that overcomes the challenges associated with a lack of standardization across observatories. The data pipeline ensures accessibility and reproducibility of the data cleaning procedures including quality control, gap-filling, and file formatting, thereby facilitating the expansion of CHOSEN to additional times and catchments. An open-source Jupyter Notebook tutorial with a user interface facilitates the modification of this pipeline to suit the needs of other investigators. Reproducible data analysis pipelines such as this one are an essential part of a modern practice of environmental science that requires rapid data assimilation capabilities to enable rapid response (Fer et al., 2021).

Although CHOSEN was developed to facilitate a range of comparative hydrology studies, we demonstrate another application here in evaluating associations between observed trends in streamflow extremes (both wet and dry) and a wide range of climatic and other hydrologic extremes from a water-balance perspective. Given the limited number of hydrologic

observatories and the well-known difficulty in performing attribution analysis on trends in the observational record (Sillmann et al., 2013), this phenomenological analysis represents early progress toward resolving the challenge of understanding the relationship between hydrological and climatic extremes. The primary contributions of this work are to establish a baseline trend assessment for extreme values (high and low, for both magnitude and frequency of the extreme events) and to provide ground-truthing for extreme event detection and attribution analyses that rely on modeled/derived water-balance quantities.

We use CHOSEN to ground-truth four main predictions. First, both low extremes and high extremes in discharge and associated hydroclimatic variables are increasing in magnitude and frequency over a broad spectrum of study areas, with significant trends in frequency more common than trends in magnitude, as has been observed in streamflow records (e.g., Archfield et al., 2016; Hirsch & Archfield, 2015; Mallakpour & Villarini, 2015).

Second, with respect to "hotspots" of hydrologic and hydroclimatic extremes, we expect that northern latitudes and high-elevation study areas will exhibit the largest proportion of monitored variables with trends in magnitude, given the expectation that climatic forcing at these locations will exceed the envelope of historical variability earlier (Batibeniz et al., 2020). Extreme event frequency trends will reflect climate model projections and previously reported hydrologic observations, with many significant trends concentrated within the eastern, southern, and upper-Midwest portions of the US (Archfield et al., 2016; Batibeniz et al., 2020; Mallakpour & Villarini, 2015). Because climate change forcing may alter water-balance partitioning in competing directions (e.g., enhancing rainfall while also enhancing evapotranspiration), regional hotspots for trends in discharge extremes will not necessarily coincide with regional hotspots for trends in other hydroclimatic extremes.

Third, trends toward wetter conditions will predominantly occur in humid locations, whereas trends toward drier conditions will predominantly occur in more arid locations. This prediction originates from the "drier-in-dry, wetter-in-wet" (DIDWIW) hypothesis from climate models (Feng & Zhang, 2015), which replaces the wet-gets-wetter, dry-gets-drier paradigm (Held & Soden, 2006; Knutson & Manabe, 1995; Wentz et al., 2007) commonly applied to oceans but now thought inapplicable to the terrestrial setting (Byrne & O'Gorman, 2015; Hu et al., 2018).

Fourth, based on findings that discharge extremes result from interactive processes (Byun et al., 2019), changes in the magnitude and frequency of discharge extremes will be associated with changes in the magnitude and frequency of extremes in other hydroclimatic variables in a regionally coherent manner that reflects their contribution to water-balance processes (Table 1). Given that climate-induced changes in water balance stores and fluxes may have opposing effects, associations among trends that accord with the signs in Table 1 will be indicative of dominant water-balance processes triggering changes in discharge extremes. We expect that

extremes in antecedent moisture, as represented through soil moisture or snow variables, will exhibit associations with both high and low discharge extremes at many study areas.

Table 1. Hypothesized sign* of correlation between trends in extreme discharge frequency and magnitude and trends in extremes of associated hydroclimatic variables, based on analysis of seasonal anomalies.

Correlated extreme	Sign of correlation, frequency comparison	Sign of correlation, magnitude comparison	Associated hydrological process
Expected correlates to low-flow extremes			
Low precipitation (unseasonably dry)	+	+	Precipitation
Low solar radiation (unseasonably cloudy)	-	-	Evapotranspiration
Low relative humidity (unseasonably dry air)	+	+	Evapotranspiration
Low SWE (low snow water content)	+	+	Snow storage
Low snow depth (low snowpack)	+	+	Snow storage
Low soil moisture (unseasonably dry soils)	+	+	Soil storage
High air temperature (unseasonably hot)	+	-	Evapotranspiration
High solar radiation (unseasonably sunny)	+	-	Evapotranspiration
High relative humidity (unseasonably humid)	-	+	Evapotranspiration
High SWE (high snow water content)	-	+	Snow storage
High snow depth (high snowpack)	-	+	Snow storage
High soil temperature (unseasonably hot soils)	+	-	Evapotranspiration
High soil moisture (unseasonably wet soils)	-	+	Soil storage
Expected correlates to high-flow extremes			
Low precipitation (unseasonably dry)	-	+	Precipitation
Low SWE (low snow water content)	-	+	Snow storage
Low snow depth (low snowpack)	-	+	Snow storage
Low soil moisture (unseasonably dry soils)	-	+	Soil storage
High precipitation (unseasonably wet)	+	+	Precipitation
High SWE (high snow water content)	+	+	Snow storage
High snow depth (high snowpack)	+	+	Snow storage
High soil moisture (unseasonably wet soils)	+	+	Soil storage

* The "+" sign of correlation for frequency comparison represents the same direction (both positive or negative) of significant trends (p-value≤0.05) in frequencies of two extremes. The "+" sign of correlation for magnitude comparison represents the positive Pearson correlation coefficient (>0.7) with significance (p-values≤0.05) of trends in magnitudes of two extremes.

2.3 Data pipeline

The data synthesis followed the workflow (Figure 1) of data cleaning (downloading, quality control, data aggregation, naming standardization), gap-filling (section 2.2), and compilation (section 2.3). We implemented this workflow by using a set of Jupyter Notebooks as a pipeline on data from each study area (e.g., Harris et al., 2020). To make the pipeline reproducible, we provided an interactive Jupyter Notebook as a tutorial for data gap-filling which allows users to

interactively tune parameters in the gap-filling functions and graphically view the result. The data products and Jupyter Notebooks are available on the Zenodo (DOI:10.5281/zenodo.4060384) and GitLab platforms.

2.3.1 Data cleaning

First, any available time series of streamflow, precipitation, air temperature, solar radiation, relative humidity, wind direction, wind speed, SWE, snow depth, vapor pressure, soil moisture, soil temperature, and water isotopes were downloaded for each study area. Subsequent quality control consisted of exclusion of erroneous values (i.e., unrealistic values such as negative precipitation or relative humidity greater than 100%, obvious typos or errors due to equipment malfunction), and cross-checking with pre-flagged entries in the downloaded product. Next, we aggregated time series data to daily time steps if the original time series were on a sub-daily scale: cumulative variables were summed for the day, and rate variables were averaged for the day. Finally, variable names were standardized using the format suggested by (Addor et al., 2020) for large sample hydrology datasets.

2.3.2 Gap-filling methods

Gaps in the cleaned and aggregated daily data (excluding isotope data) were filled using one of three techniques, depending on the length of the gap and availability of complementary data. We applied the three techniques sequentially, meaning gaps not filled by the first technique would undergo the second method, etc. (Figure 2). First, for gaps of less than seven days, linear interpolation was applied. Though using linear interpolation may be improper for variables like precipitation, we made this operational decision for reasons of internal consistency, noting that our data processing pipeline gives researchers the necessary information to implement alternative processing conventions.

To fill gaps longer than seven days, we applied spatial regression for study areas that have multiple adjacent stations, and then applied temporal regressions for study areas that have long records (Pappas et al., 2014). To implement spatial regression, we first evaluated the correlation coefficients between the station with missing values and all the other stations in the same study area. We then used the data from the station with the highest correlation coefficient to estimate the linear regression parameters. If the highest correlation coefficient was less than 0.7, or if no data were available from other stations contemporarily, the missing values were reconstructed by the climate catalog technique. In the climate catalog (i.e., temporal regression) method, we filled gaps using data from the most highly correlated year at the same site, selected from among years with at least 9 months of data and a correlation coefficient greater than 0.7 to the missing-data year. Gaussian random noise was added to the resulting regression-based estimate, scaled by the standard deviation of the record of each date in the gap across all years, in order to maintain the variation statistics of the original time series. However, this technique may not be useful for

reconstructing non-random variations in time series that are large-scale (i.e. wet and dry years) or small-scale (i.e. before and after a storm).

To assure the quality of the gap-filled data, we deleted any values that exceeded the maximum or fell below the minimum of the original time series. Finally, flags were generated to differentiate between raw, missing, and filled data, indicating the technique used to create each reconstructed data point.

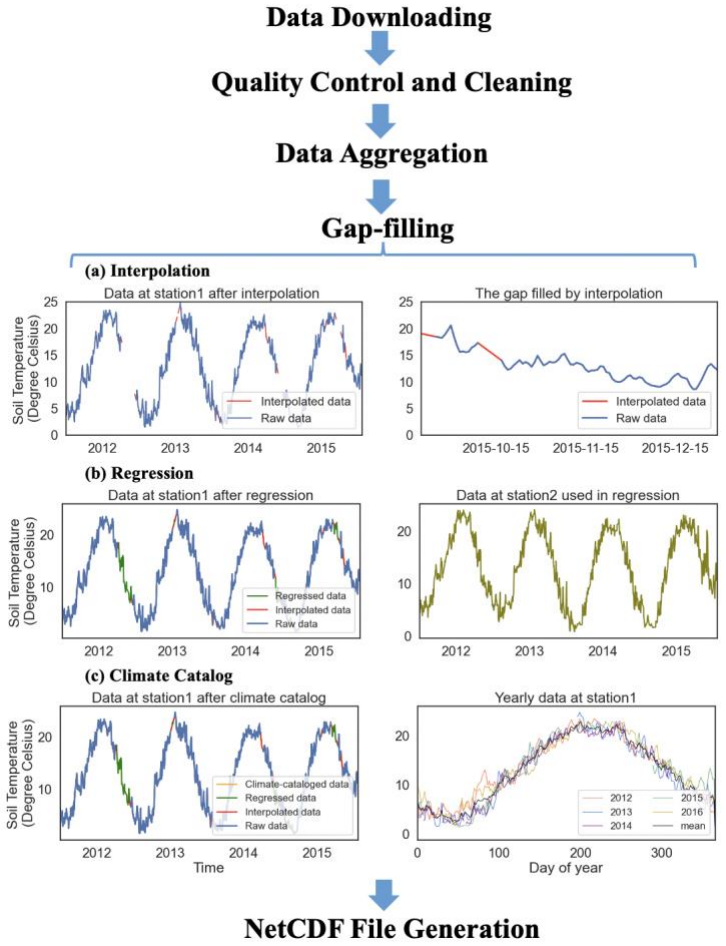


Figure 2. Data pipeline and visualizations of cleaning methods: a) interpolation, b) spatial regression and c) climate catalog (i.e., temporal regression).

2.3.3 NetCDF data product

We stored and published the processed data in NetCDF (Network Common Data Form) format. NetCDF data have hierarchical structures and are self-explanatory, which means the descriptions of the attributes of the data tables are accessible from the file by different programming interfaces, for example, C++, Java, Python, and MATLAB. NetCDF is emerging as the data standard for large-sample hydrology, as well as for other large-sample products across the geosciences, particularly climate science and remote sensing (Liu et al., 2016; Signell et al.,

2008). The NetCDF library is designed to read and write multi-dimensional scientific data in a well-structured manner. The library enables writing data in several coordinate dimensions, accommodating multiple measurement stations.

We stored the data and metadata from each study area in one NetCDF file. In the NetCDF files, the hydrometeorological variable data and associated data flags are two-dimensional arrays (i.e., by time and location). There is a timestamp variable for conveniently checking the first starting date and last ending date for data in this study area. The grid variable contains information about monitoring stations, providing the names, latitudes, and longitudes and elevations if available.

2.4 Dataset description

We synthesized data from 30 intensively monitored study areas across the United States (Figure 3). Sixteen of the 30 study areas are from the LTER network (Servilla & Brunt, 2011), 11 from the CZO network (Zaslavsky et al., 2011), and the remaining three are East River, Dry Creek, and Sagehen Creek (Kakalia et al., 2021; McNamara, 2017; Petersky & Harpold, 2018). Table S1 includes additional information about the study areas in the CHOSEN dataset such as data source links, geographical information, and climate conditions.

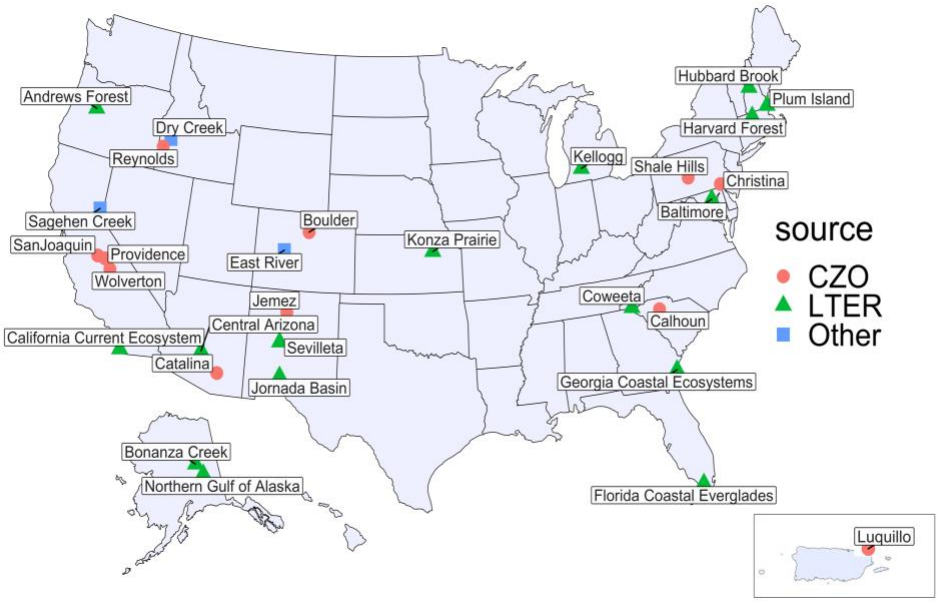


Figure 3. Geographical distribution of the study areas. “CZO” represents Critical Zone Observatories; “LTER” represents Long-term Ecological Research Stations; “Other” represents observatories managed by other entities.

The availability of different variables in CHOSEN varies by site. The H.J. Andrews and Bonanza LTER datasets contain all 13 variables, with most other datasets having around 10 variables. Discharge record lengths range from three years at Calhoun to 78 years at the San Diego River (California Current Ecosystem LTER), with a median of 19 years (Figure 4).

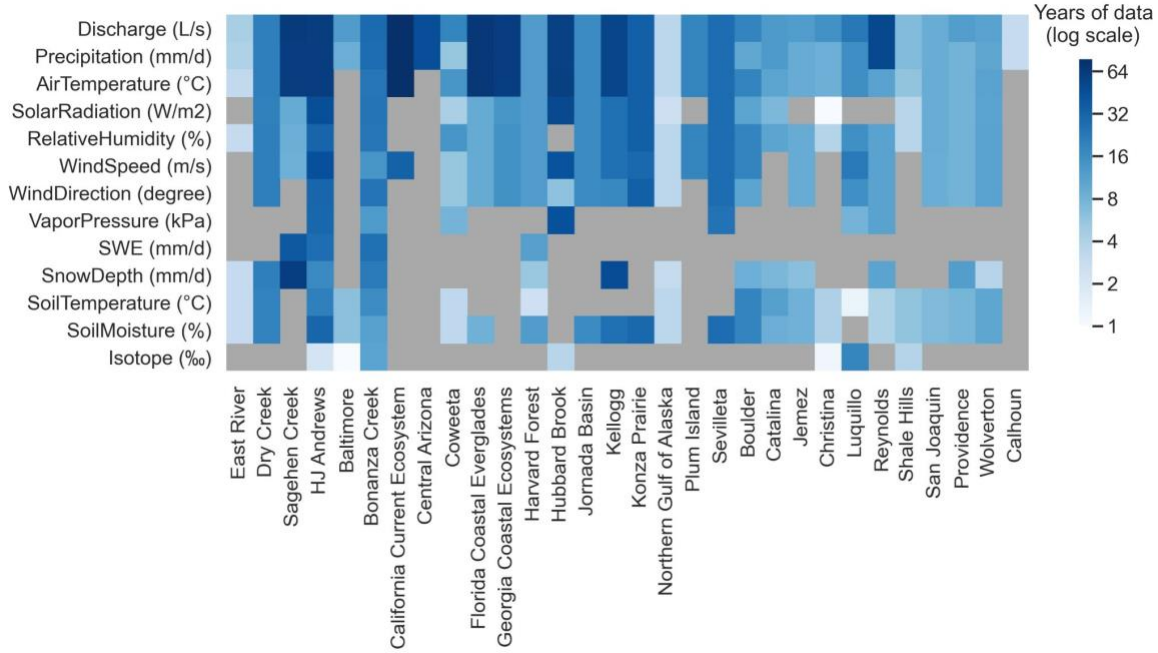


Figure 4. The span of time series availability and duration across study areas.

Discharge and precipitation time series are available in all CHOSEN study areas, and seven catchments have soil moisture and snow measurements with records exceeding five years. Although publicly available water isotope data are limited, we identified six study areas with water isotope time series longer than one year (Figure 5). The measured isotopes include ^{18}O and deuterium in streamflow, precipitation, and snowpack. Note that, unlike other variables, the resolution of isotope data is sparse, usually weekly or biweekly.

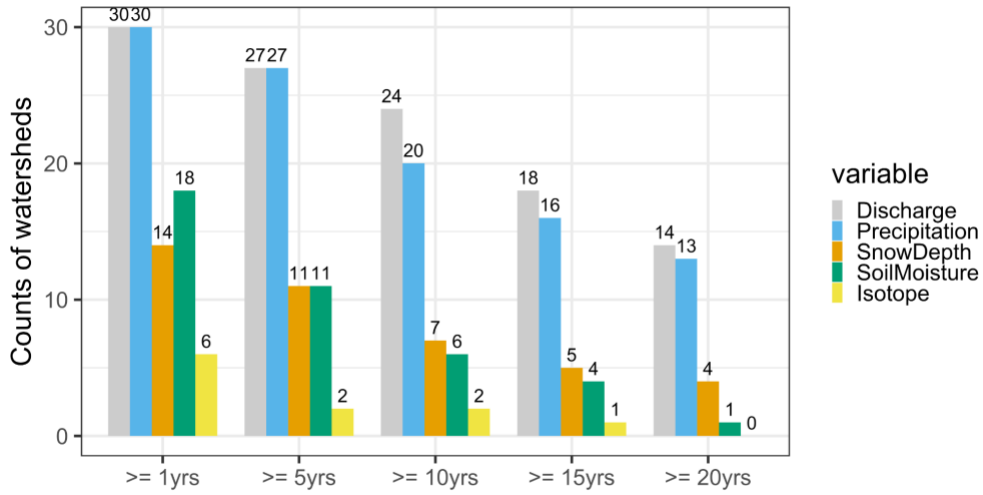


Figure 5. Distributions of record spans for a selection of variables in CHOSEN.

2.5 Extreme events analysis with CHOSEN data

2.5.1 Methods

Extreme events are occurrences above or below certain thresholds of exceedance over a period of time. In this paper, we evaluated extreme events based on seasonal anomalies, in which we first removed the seasonality calculated by a moving average of 30 days (Appendix A). Then we picked out local minima/maxima in the time series as independent events and identified the high or low extremes as the independent events above the value at the percentile ranking of 95% or below the value at the percentile ranking of 5%. We studied the extreme events of each hydro-meteorological variable with a record longer than 10 years for each study area in CHOSEN, with the exception of water isotopes. If the study area had multiple measurement records for a single variable, we chose the longest.

We used the Mann-Kendall trend test (M-K test) to identify the significance (with a p-value less than 0.05) and sign (increasing or decreasing) of monotonic trends in extreme event magnitudes and frequencies over time (Kendall, 1975; Mann, 1945). For convenience, we refer to significant test results as trends, though we recognize that the M-K test is specific only to the monotonicity, rather than to the magnitude of the trend. The M-K tests were performed on two kinds of statistics: annual counts and the annual median of the extreme event magnitudes, to detect trends in frequency and magnitude, respectively. The analyses were conducted for both high and low extremes and implemented using the python package scikit-learn (Pedregosa et al., 2021). It is worth noting that autocorrelated time series remain a challenge in the M-K test. Autocorrelated

time series may artificially inflate test statistics, resulting in false positives in the trend detection (Storch & Navarra, 1999; Yue et al., 2002). Our usage of the annual interval for the statistics decreases the likelihood of within-water-year autocorrelation that would arise from using shorter intervals.

Following the M-K trend analyses, the percentage of extreme-value time series available at each study area (for all variables, excluding isotopes, with record length longer than 10 years) with significant trends was computed as a first step in identifying locations that are "hotspots" for change across multiple hydrologic and hydroclimatic variables. For example, this value would be 25% for a study area with sufficient precipitation and discharge record lengths that exhibited a trend only in high-flow extremes (because one of the four possible extremes -- high flow, low flow, high precipitation, and low precipitation -- exhibited a trend). Hotspots were operationally defined as study areas that exhibited trends in over two-thirds (66.67%) of the available extreme-value time series.

Next, based on significant trends in the magnitude and frequency of extreme discharge, study areas were classified as "wetting" or "drying" with respect to discharge. Specifically, increases in the magnitude of extreme high or low-discharge, decreases in the frequency of extreme low-discharge, and increases in the frequency of extreme high-discharge were all classified as "wetting" trends, and vice-versa. We caution readers that these labels are not intended to apply to total water availability within the study area and that they are not necessarily representative of water availability outside of extreme flow events.

Last, we evaluated whether wetting or drying trends with respect to discharge were associated with trends indicative of wetting or drying in other water-balance stores and fluxes in a manner consistent with a simple water-balance explanation (i.e., Table 1). Namely, we evaluated correlations between significant trends in extreme discharge and significant trends in other monitored hydroclimatic variables. A positive correlation means that both variables trended in the same direction; a negative correlation means they trended in opposite directions. We compared these correlations with our predictions in Table 1 and counted how many correlations matched the predictions. Meanwhile, we identified counterfactuals to the predictions. Here, a counterfactual is an observed trend in the extremes of an associated hydroclimatic variable that has a sign opposite that predicted in Table 1 and, for sites where a trend in high or low-discharge extremes was also detected, is likewise inconsistent with the high or low-flow predictions. This complex definition accounts for the fact that trends in extremes contain no information about within-year timing, and that high-discharge and low-discharge extremes may be sensitive to different hydroclimatic extremes that occur at different times of the year. For example, increasing frequency of low-SWE events associated with an increasing frequency of high-discharge events is a counterfactual if there is no significant trend in low-discharge. However, it is not a counterfactual if that same catchment also shows an increasing frequency of low-

discharge events; indeed, low flows may be most sensitive to wintertime delivery of snow, whereas high-flow events may be most sensitive to warm-season rainfall.

2.5.2 Results

Among 26 study areas with records longer than 10 years, trends in the magnitude and frequency of extreme hydro-climatological and hydrological events were common. All variables in CHOSEN exhibited significant trends in magnitude and in frequency for at least one study area. These trends were distributed among 23 unique sites, with 22 sites exhibiting trends in frequency and 22 sites exhibiting trends in magnitude (Figure 6). On the whole, 81 trends in frequency and 101 trends in magnitude were observed.

Observed trends were indicative of changes to the full suite of water-balance stores and fluxes considered (evapotranspiration, snow storage, soil moisture storage, precipitation, discharge), with "hotspots" of change (defined here as areas with significant trends in over two-thirds of the observed variables) in the southeast (Florida Coastal Everglades), northeast (Hubbard Brook), Pacific Northwest (H.J. Andrews), and Alaska (Bonanza; Figure 6). These hotspots were geographically consistent across magnitude and frequency trends, except for Bonanza, which fell just short of the hotspot threshold for magnitude.

Within sites, trends in frequency and magnitude of extremes generally provided similar information about changes in water balance processes (i.e., Figure 6A compared to Figure 6B). Across sites, trends indicating changes in evapotranspiration were most common (19 study areas), followed by changes in precipitation (15 study areas), discharge (11 study areas), snow storage (two study areas), and soil moisture storage (two study areas). Most trends commonly associated with controls on evapotranspiration suggested increases, though at many sites, increasing high-relative-humidity events that accompanied increasing high-temperature events (Appendix B1) exerted competing influences. Trends indicative of changes in extreme runoff, snow storage, and soil moisture storage showed more geographic and temporal heterogeneity (e.g., increases in "high" extremes coupled with decreases in "low" extremes) in the direction of the change compared with trends related to the evapotranspiration. We have repeated this experiment excluding the climate-catalog data and found consistent results (Appendix B2).

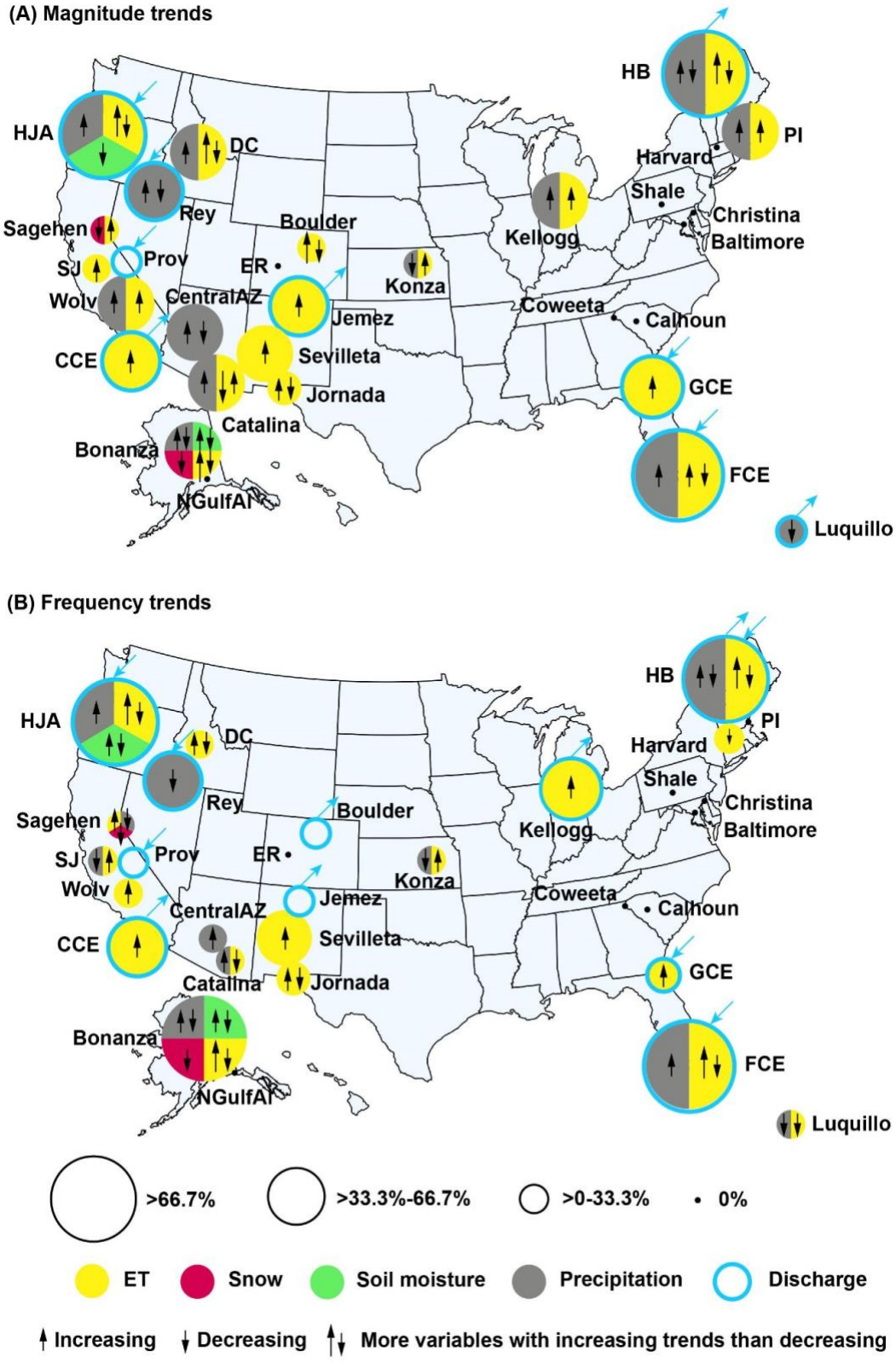

(A) Magnitude trends
HB
PI
HJA
DC
Harvard
Shale
Rey
Christina
Baltimore
Sagehen
Boulder
Kellogg
Prov
SJ
ER
Konza
Wolv
CentralAZ
Coweeta
Jemez
Calhoun
CCE
Sevilleta
GCE
Jornada
Catalina
Bonanza
FCE
NGulfAI
Luquillo
(B) Frequency trends
HB
PI
HJA
DC
Harvard
Shale
Rey
Christina
Boulder
Baltimore
Sagehen
Kellogg
Prov
SJ
ER
Konza
Wolv
Coweeta
Jemez
Calhoun
CentralAZ
CCE
Sevilleta
GCE
Catalina
Jornada
Bonanza
FCE
NGulfAI
Luquillo
>66.7%
>33.3%-66.7%
>0-33.3%
0%
ET
Snow
Soil moisture
Precipitation
Discharge
Increasing
Decreasing
More variables with increasing trends than decreasing

Figure 6. Distribution of significant magnitude (A) and frequency (B) trends among study sites and related hydrological stores and fluxes. The size of the bubbles indicates the percentage of the extreme-value time series available at a study area (for variables, excluding isotopes, with record lengths longer than 10 years) that exhibit significant trends in frequency or magnitude. Variables were classified as relevant to evapotranspiration (soil and air temperature, relative humidity, solar radiation), snow (SWE, snow depth), soil moisture, precipitation, or discharge, and those stores and fluxes exhibiting significant trends are indicated by color; note that size of the colored regions within each bubble has no relation to the proportion of trends relevant to that process. Arrows denote whether the trends suggest an increase or decrease in the time-averaged behavior of the associated store or flux (e.g., more frequent low-precipitation extremes would suggest a decrease in precipitation and be denoted with a down arrow in B). When multiple variables associated with the store or flux (e.g., temperature and relative humidity for evapotranspiration) or trends in high and low extremes for the same variable suggest opposing behavior (e.g., increasing low-precipitation and high-precipitation extremes), both arrows are depicted, though a larger arrow indicates the direction implied by a majority of trends.

Similar to the whole suite of variables (Figure 6), extreme discharge exhibited more trends in magnitude (17) than frequency (14), though more study areas (10) exhibited trends in frequency than in magnitude (nine; Appendix B1). Study areas often exhibited significant trends in both low and high discharge events that indicated either consistent wetting or drying (e.g., increasing magnitude of both low- and high-discharge events, or increasing frequency of high-discharge events coupled with decreasing frequency of low-discharge events), with the exception of Hubbard Brook, which exhibited increasing frequency of both high- and low-discharge extremes. Similarly, for study areas showing trends in both frequency and magnitude, the trends pointed consistently toward wetting or drying (Figure 7). Namely, study areas that had trends toward drier conditions with respect to discharge were clustered in the Southeast (Georgia and Florida) and Northwest (Oregon, Idaho, and central California). Meanwhile, study areas that exhibited trends toward wetter or predominantly wetter conditions in terms of discharge were located in the high-elevation West (Colorado, New Mexico), coastal Southwest (southern California), upper Midwest (Michigan), and Northeast (New Hampshire). With respect to the drying observed in the southeast and wetting observed in the montane west, this geographic pattern diverged from the DIDWIW prediction.

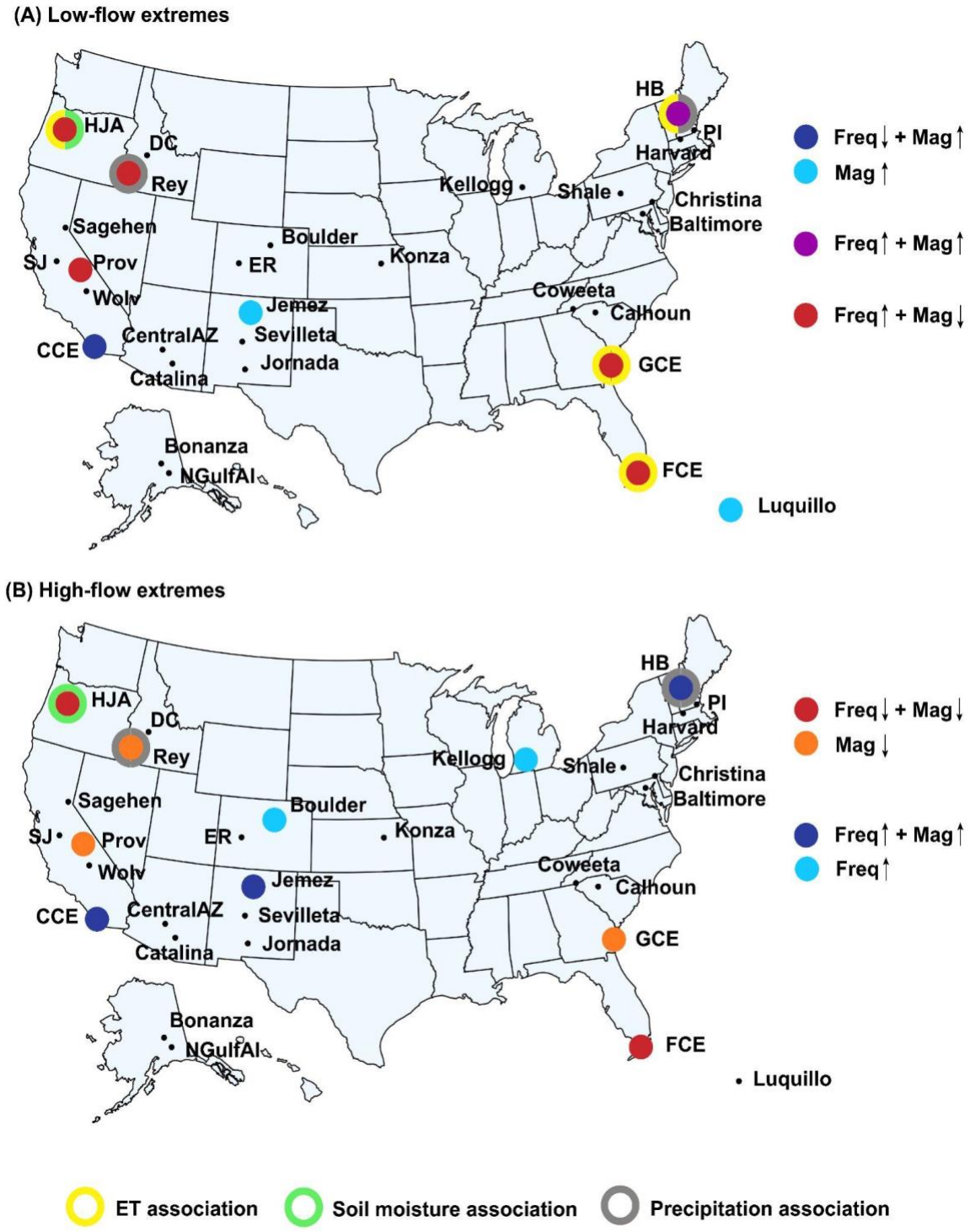


Figure 7. Trends in low-flow (A) and high-flow (B) extremes and associated trends in hydroclimatic stores and fluxes consistent with water-balance explanations of how those stores and fluxes impact streamflow (Table 1). Shades of blue suggest wetting trends (with respect to the particular discharge extreme plotted, based on trends in frequency and/or magnitude), while shades of red suggest drying trends. Purple represents a combination of a wetting trend (i.e.,

increased low-flow magnitude) and drying trend (i.e., increased low-flow frequency). Colored outlines show one or more associated significant trends in other hydroclimatic extremes that are consistent with a simple water-balance explanation (i.e., Table 1). Circles without outlines imply that the study area has no univariate associations between extreme discharge and other hydroclimatic extremes consistent with a simple water-balance explanation.

Five out of nine study areas exhibiting significant trends in low-flow events and three of the 10 study areas exhibiting trends in high-flow events showed associated trends in other hydroclimatic variables consistent with the predictions in Table 1. In the Southeast, drying trends in discharge extremes were associated with trends indicative of increased evapotranspiration, while the drying trends in discharge extremes observed in the northwest had more diverse associations: decreased precipitation, decreased soil moisture, and increased evapotranspiration (Figure 7; Tables 2 and 3). In the northeast, more frequent low-discharge extremes were associated with trends indicative of increased evapotranspiration and more frequent low-precipitation extremes. Meanwhile, wetter discharge extremes had almost no associations with trends in hydroclimatic variables, with the exception of Hubbard Brook, where wetter high-flow extremes were associated with more frequent and higher-magnitude precipitation extremes.

Overall, observed associations between discharge extremes and extremes in other hydroclimatic variables partially upheld our fourth prediction. Specifically, in many locations, trending extremes in discharge could be associated with trending extremes in one or more water balance processes. As expected, interactions among these processes were complex and often confounding; study areas with associations consistent with changing hydroclimatic inputs also commonly exhibited counterfactuals (Tables 2 and 3). Just two study areas exhibited associations that were only counterfactual to the water-balance expectations; both California Current Ecosystem and Jemez exhibited wetter low- and high-flow extremes based on trends in discharge frequency and magnitude (Figure 7), despite trends indicative of higher evapotranspiration. The remainder of the study areas with significant trends in discharge extremes exhibited no other trends in hydroclimatic variables. In contrast to our third prediction, widespread associations between variables indicative of antecedent moisture (i.e., soil moisture, snow depth, SWE) and discharge extremes were not observed. Only at H.J. Andrews was an association between soil moisture and discharge extremes observed.

Table 2. Hypothesis testing of correlation between trends in extreme discharge frequency and trends in the frequency of extremes of associated hydroclimatic variables. Only study areas with significant trends in discharge are included. Counterfactuals are compiled across the low-flow and high-flow analyses and represent correlations between significant trends in discharge and significant trends in other monitored variables that have signs opposite those depicted in Table 1.

Study Areas	Discharge record (yrs)	# Total trends	Low-flow extremes frequency	# Consistent with low-	Low-flow related processes	High-flow extremes frequency	# Consistent with high-	High-flow related processes	# Counterfactuals	Counterfactuals

				flow hypotheses			flow hypotheses			related processes
H.J. Andrews	62	13	increasing	4	ET, Soil Storage	decreasing	1	Soil Storage	4	P, ET, Soil Storage
California Current Ecosystem	78	4	decreasing	0		increasing			1	ET
Florida Coastal Everglades	68	7	increasing	1	ET	decreasing	0		3	P, ET
Georgia Coastal Ecosystems	61	2	increasing	1	ET					
Hubbard Brook	59	7	increasing	2	P, ET	increasing	1	P	1	ET
Kellogg	55	5				increasing	0		0	
Reynolds Creek	52	3	increasing	1	P				0	
Boulder Creek	19	1				increasing	0		0	
Jemez	12	1				increasing	0		0	
Providence	12	1	increasing	0					0	

Table 3. Hypothesis testing of correlation between trends in extreme discharge magnitude and trends in the magnitude of extremes of associated hydroclimatic variables. Only study areas with significant trends in discharge are included. Counterfactuals are compiled across the low-flow and high-flow analyses and represent correlations between significant trends in discharge and significant trends in other monitored variables that have signs opposite those depicted in Table 1.

Study Areas	Discharge record (yrs)	# Total trends	Low-flow extremes magnitude	# Consistent with low-flow hypotheses	Low-flow related processes	High-flow extremes magnitude	# Consistent with high-flow hypotheses	High-flow related processes	# Counterfactuals	Counterfactuals related processes
H.J. Andrews	62	13	decreasing	5	ET, Soil Storage	decreasing	2	Soil Storage	3	P, ET
California Current Ecosystem	78	4	increasing	0		increasing	0		1	ET
Florida Coastal Everglades	68	7	decreasing	1	ET	decreasing	0		3	P, ET

Georgia Coastal Ecosystems	61	4	decreasing	1	ET	decreasing	0		0	
Hubbard Brook	59	8	increasing	2	ET	increasing	1	P	2	P, ET
Jemez	12	3	increasing	0		increasing	0		1	ET
Luquillo	23	2	increasing	0					0	
Reynolds	52	4	decreasing	1	P	decreasing	1	P	1	P
Providence	12	2	decreasing	0		decreasing	0		0	

2.5.3 Discussion

CHOSEN contains an uncommon breadth of variables that allows for analysis of trends in multiple extremes, which is typically beyond the scope of observational extreme events studies. One advantage of analyzing multiple types of extremes simultaneously is the potential to evaluate multiple types of wetting or drying processes that affect different hydrological stores and fluxes. Such an analysis addresses the critique that the pronouncement of "wetting" or "drying" based on trends in a single variable (e.g., discharge, soil moisture, evapotranspiration flux) may be misleading (Roth et al., 2021). Indeed, our overall portrait of trends in hydrologic and hydroclimatic extremes (Figure 6) confirms that processes typically assigned the label "drying" or "wetting" may coexist within single locations (e.g., co-occurrences of "up" arrows for precipitation and "down" arrows for discharge or soil moisture). Further, with respect to single variables within single locations, trends in extremes often indicated both "wetting" and "drying" by exhibiting an increase in the magnitude of high extremes coupled to a decrease in the magnitude of low extremes. With respect to discharge, however, trends in low and high extremes tended to point toward consistent wetting or drying within individual study areas (i.e., Figure 7A compared to 7B), evidencing a shift in the whole distribution of streamflow, as has also been overwhelmingly observed at the global scale (Gudmundsson et al., 2019).

Though most observational studies have been limited to one type of extreme, climate modelers have used a multivariate Climate Extremes Index (Gleason et al., 2008) to identify likely "hotspots" of combined wet, dry, hot, and cold extremes from downscaled global climate models (Batibeniz et al., 2020), which our observations largely corroborate. Consistent with our finding of multivariate extreme "hotspots" in south Florida, Oregon, and New Hampshire, study indicated that by 2050, Florida, New England, and the Pacific Northwest are likely to develop the most extreme conditions across a suite of variables (Batibeniz et al., 2020; Appendix C). Note that they did not consider Alaska, our fourth hotspot, but they did find that the extreme conditions would extend into the Rocky Mountain west, where our observations indicated a less comprehensive set of trends to date. Furthermore, the study found that these patterns were primarily driven by warming and drying conditions, as the majority of areas did not exceed the

historical envelope of variability for intense precipitation events until 2050. Namely, the Florida hotspot primarily arose from extreme warm conditions, consistent with the decreased discharge/increased evapotranspiration associations that we observed. Meanwhile, the Pacific Northwest and New England hotspots predominantly arose from extremely dry conditions, consistent with our observed decreased soil moisture and increased evapotranspiration trends at H.J. Andrews and increased evapotranspiration trend at Hubbard Brook, together with an increased frequency of low-discharge extremes.

Although our observations generally upheld climate model-based projections of extreme event hotspots, they deviated from projections and previous observations in a few ways. First, our analysis resolved no trends in extremes for any of the five sites in the Mid-Atlantic region (Figure 6), in contrast to projected drying trends in streamflow extremes (Naz et al., 2016), observed wetting trends in high-streamflow extremes (Archfield et al., 2016), projected increases in hurricane-related flood hazards (Marsooli et al., 2019), and observed increasing trends in the climate extremes index for the 1981-2005 period, encompassing both drought and intense wet events (Batibeniz et al., 2020). Our lack of trends in the Mid-Atlantic region was likely strongly driven by the limited data record length (among the shortest of all sites for variables other than discharge) for most of the Mid-Atlantic observatories (Figure 4). To test whether short record length had impeded our ability to detect trends, we carried out two-sample t-tests. Results showed that the time series with identified trends for both frequency and magnitude of extreme events were significantly longer ($p<0.01$) than those with no trends. For most of the study areas, the record lengths for discharge, precipitation, and air temperature were sufficient, whereas, for other hydroclimatic variables, the scarcity of long records substantially restricted the trend analysis.

In addition to insufficient record lengths for some variables and study areas, geographic undersampling may also explain discrepancies between our findings and the literature. In the Mid-Atlantic region, both high-flow (Archfield et al., 2016) and low-flow (Kam & Sheffield, 2016) trends exhibit strong variability in sign and significance, making it likely that observations from just a few sites would not be representative of the regional mean. Undersampling of the Midwest in CHOSEN might also explain why we observed just one study area with a significant change in the frequency of high flows in this region (i.e., Kellogg), despite the prevalence of increased flood frequency observed for the region in other observational studies (Ahn & Palmer, 2016; Hirsch & Archfield, 2015; Mallakpour & Villarini, 2015).

The geographic undersampling inherent in CHOSEN may additionally provide an explanation for why our second prediction--that we would observe more trends in extreme event frequency than magnitude, as observed in geographically extensive discharge records (Hirsch & Archfield, 2015)--was not upheld. In contrast to this prediction, we observed a comparable number in trends in magnitude as in frequency (Figure 6). Small-sample bias may have been exacerbated in

CHOSEN by the preferential siting of many of the observatories in areas where rapid climate-driven change is expected. Furthermore, given that observed trends in extreme discharge are highly variable in sign and significance throughout the US (Ahn & Palmer, 2016; Archfield et al., 2016), it is not unexpected that the slight dominance of magnitude trends among our subset of sites would emerge from chance. A second potential explanation for the surprisingly large number of trends in magnitude is that many of these trends involved temperature or variables thought to be directly driven by temperature (Figure 6), and recent climate models (Batibeniz et al., 2020) suggest near-term (median: by 2025) emergence from the envelope of historic variability for temperature for most of the US.

Though undersampling provides a partial explanation for why aspects of our first and second predictions were not upheld, discrepancies from the DIDWIW prediction are likely not attributable to random sampling artifacts. Consistently across sites and variables, study areas in the arid Southwest showed trends toward wetter extremes, reflected in precipitation and discharge magnitude and frequency trends, while those in the humid Southeast showed trends toward drier extremes, reflected in discharge and evapotranspiration-related trends (Figure 6). This discrepancy underscores the importance of considering multiple variables in assessing wetting and drying trends (sensu Roth et al., 2021); the DIDWIW hypothesis was developed based on analysis of long-term, remotely sensed soil moisture changes between 1979 and 2013 (Feng & Zhang, 2015), whereas the increase in intense precipitation events forecasted for the Southwest (Batibeniz et al., 2020) may trigger high-flow extremes through Hortonian overland flow without a long-term increase in soil moisture, which would be consistent with our limited observations. Meanwhile, in humid environments like the Southeast, evapotranspiration may impact peak flow volumes while soils remain moist. Further, the soil moisture observations from 1979 to 2013 in Feng and Zhang (2015) may not have captured more recent changes in the Southwest present in CHOSEN. In fact, it is likely that the trends detected in this analysis are recent, as a 1981-2005 observational study of historical trends in intense precipitation events also shows no significant trends for the region (Batibeniz et al., 2020). Our results, taken together with model projections (e.g., Batibeniz et al., 2020), suggest that the DIDWIW paradigm will become less applicable as climate change advances.

Our ability to attribute observed trends in discharge to changes in dominant water balance processes was limited by the logical incongruity of correlative associations and causality and by a lack of long-term records of soil moisture and/or snow storage in most study areas. Nonetheless, the associations depicted in Figure 7 are generally consistent with previous studies that attribute changes in extreme discharge to underlying hydrological processes. In a statistical study based on precipitation and temperature measurements and modeled soil moisture and snowmelt, Berghuijs et al. (2016) found that increasing soil moisture storage is a strong predictor of extreme high-discharge throughout the Pacific Northwest, consistent with the soil moisture/high-flow association we found at H.J. Andrews (Figure 7B). Further, the association

between precipitation extremes and high-flow extremes that we found at Hubbard Brook and Reynolds Creek (where snow data records were too short for trend analysis) may be indicative of the importance of extreme precipitation for the rain-on-snow events found to be the dominant factor explaining trends in high flow for these regions (Berghuijs et al., 2016). Meanwhile, climate-model based attribution of decreasing magnitude of low flows in the Southeast to warmer temperatures (Hayhoe et al., 2007) is consistent with our observations (Figure 7A), as is a statistically based attribution of decreased low flows in Idaho to decreased precipitation inputs (Kormos et al., 2016). However, in contrast to the Kormos et al. study, we found no association between precipitation and low-flow extremes in Oregon (H.J. Andrews). Instead, we found associations to soil moisture and evapotranspiration extremes, the former of which was not considered in their study.

Attributional studies in the literature suggest mechanisms that may explain observed trends in discharge extremes that were not associated with other trends in our study (Figure 7). Increasing frequency and/or magnitude of high-flow extremes observed at the Kellogg (Michigan), Boulder Creek (Colorado), and Jemez (high-elevation New Mexico) observatories may be attributable to increasingly rapid snowmelt events triggered by warmer temperatures or rain on snow (Mallakpour & Villarini, 2015). These mechanisms would not be captured by our data, which lacked long-term snow records for these sites, or our analysis, which did not consider multivariate interactions between temperature or precipitation and snow storage. Meanwhile, less snow storage over time as a result of precipitation falling increasingly as rain instead of snow may explain drying trends in both high- and low-flow extremes at the Providence observatory (McCabe & Wolock, 2009; Miller et al., 2003). Lastly, climate models suggest that the wetting trends projected for the Southwest (e.g., California Current Ecosystem) are attributable to increased total precipitation delivery (Heidari et al., 2020), which might not be reflected in precipitation extremes.

Attributional studies typically assume that evapotranspiration plays no role in high-discharge extremes (e.g., Berghuijs et al., 2016 and Table 1 of this study). However, this assumption may not be valid for coastal and low-gradient parts of the Southeast, where watershed areas are large, flows are slow-moving, and the highest flows occur during the warmest part of the year and are not associated with snowmelt or frontal systems. At both the Georgia Coastal Ecosystem and Florida Coastal Everglades observatories, decreasing trends in the magnitude and/or frequency of high flow extremes are observed despite increasing (Florida) or no significant (Georgia) trends in high-precipitation extremes (Figure 6). Both of these areas, however, have exhibited increasing temperature trends (Appendix B1) that are among the strongest in the US (Batibeniz et al., 2020).

In summary, though our study was not attributional, it supports other attributional studies in suggesting that drying shifts in extreme streamflow in the Pacific Northwest and Southeast are likely linked to decreased precipitation inputs, decreased soil moisture, and increased

evapotranspiration due primarily to warming. Wetting shifts in streamflow extremes are more challenging to explain via simple statistical analyses, as evidenced by a prevailing lack of associations to other hydroclimatic variables (Figure 7). Though our findings fall short of reconciling Sharma et al.'s grand challenge (2018) to attribute changing streamflow extremes to changes in hydroclimatic forcing, they suggest three hypotheses that are potentially addressable through more sophisticated statistical analyses or longer periods of record as CHOSEN continues to grow. First, the preferential location of wetting high-flow extremes in regions with snowpack suggests that these trends may be linked to increasingly rapid snowmelt, due to interactions between temperature or precipitation and snow storage. Second, higher rates of evapotranspiration may decrease high-flow extremes in locations without a snowmelt peak or dominantly frontal mechanisms of precipitation delivery. And finally, given the modeling results of Berghuijs et al. (2016) and the observed association at H.J. Andrews, changes in soil storage (Dymond et al., 2014) likely also drive changes in streamflow extremes in many regions.

2.6 Conclusion

To the best of our knowledge, the CHOSEN database is the largest open-source collection of comprehensive data from hydrological observatories, containing variables important to understanding water-balance partitioning that are not typically present in existing large-sample databases. It thus fulfills critical data needs for comparative hydrology. In particular, it lays a foundation for studies that establish hydrologic baselines, synthesize information on multiple aspects of "wetting" and "drying," ground-truth model projections of highly uncertain, derived hydrological quantities, and attempt to attribute observed changes to underlying hydrological processes.

Our simple synthesis of trends in hydroclimatic extremes generated generally consistent results with model projections and statistical studies that use derived quantities for soil moisture, instilling confidence in model projections. Consistency was strong in the identification of geographic hotspots for multivariate change in extremes and in the hydrologic stores and fluxes dominantly associated with those extremes. However, observations were less consistent with projections of discharge trends (Naz et al., 2016). Namely, many areas where we resolved drying trends in high-flow extremes (i.e., red points in Figure 7B) were projected to exhibit wetting trends by 2050, with the exception of south Florida, where both model projections and observed trends indicated drying. We propose that this inconsistency may reflect late emergence (i.e., around or after 2050) from the historic envelope of variability for wet extremes in most regions of the US (Batibeniz et al., 2020) rather than fundamental flaws of the model.

Impending emergence from the envelope of historical variability for both wet and dry extremes underscores the need for synthesis products from hydrologic observatories that can document baselines for wetting or drying across different components of the water balance. Our analysis,

for example, suggests that in the Southwest, which is projected to show wetter extremes by 2050 (Batibeniz et al., 2020; Naz et al., 2016), a signature of wetting extremes in both precipitation and streamflow (Figure 6) is emerging. It further suggests that this emergence is recent, as these trends were absent in 1981-2005 observations (Batibeniz et al., 2020). The emergence of this wetting trend, together with drying in the Southeast with respect to discharge and evapotranspiration extremes, suggest that the WIWDID paradigm may be inadequate to describe ongoing climate-induced hydrological change across a suite of variables.

Lastly, though simple associations between hydroclimatic and hydrologic extremes were often consistent with a water-balance framework (Table 1) and prior attributional studies (Section 4.3), they were not sufficient to attribute most wetting trends in streamflow extremes to underlying mechanisms. These shortcomings underscore the need for analyses based on longer-term (i.e., >10 years), comprehensive, and openly available records of soil moisture and snow variables. The data record lengths in CHOSEN will continue to grow, and calls for more soil moisture data nationally are increasingly being heard (e.g., Sungmin & Orth, 2021; Petersky & Harpold, 2018; Wasko & Nathan, 2019). We echo that call and build upon it, highlighting that comprehensive observations related to changes in evapotranspiration (e.g., relative humidity, solar radiation, soil and air temperature, wind speed, and/or direct moisture flux data) may be relevant to explaining a wider range of hydrologic extremes than previously thought.

Chapter 3

A Physics-informed Machine Learning Model for Streamflow Prediction

3.1 Abstract

In various contexts, deep learning models have demonstrated superior performance over physically-derived process-based models in predicting streamflow. Despite their efficacy, these models often face criticism for their lack of interpretability and insight into the underlying physical processes governing streamflow response. To address this challenge, we introduce a novel approach employing a long short-term memory (LSTM) model that integrates water balance constraints for streamflow prediction. Our proposed physics-informed LSTM (PILSTM) combines a discharge-storage model with the LSTM architecture. We apply this hybrid model to eight intensively-monitored watersheds in the United States. Additionally, we conduct a comprehensive comparison of the LSTM, physical, and PILSTM models, evaluating their performance under scenarios simulating climate change and data-scarce conditions and with and without pretraining on large datasets. We find that for most watersheds, greater performance gains arise from pretraining LSTM-based models on large datasets than from integrating physical constraints, with reduced sensitivity to variability in input climate data. However, for watersheds with characteristics poorly represented in the pretraining database, pretraining can deteriorate performance relative to site-specific models. For those watersheds, integrating physical information can serve as a safeguard against poor performance from pretraining, particularly when data length for the target watershed is limited. Further, they can produce seasonal patterns of sensitivity to precipitation and evapotranspiration, better aligning with physical understanding. Although LSTM-based models trained on wet conditions generally perform well for dry conditions, for dry watersheds with long storage time and limited streamflow variability, physical models produce the best performance.

3.2 Introduction

Streamflow prediction is crucial for managing water resources, including water supply, reservoir operations, irrigation, energy production, flood and drought mitigation, and ecosystem management. The accuracy of these predictions is becoming increasingly important due to uncertainties associated with climate change, particularly in areas that rely on snowpack, experience flash flooding, or suffer from ecosystem degradation. The hydrologic science community has also invested in development of strategies to apply learning from well-instrumented watersheds to those with comparative data scarcity (Hrachowitz et al., 2013; Kratzert, Klotz, Shalev, et al., 2019).

There are two primary categories of traditional streamflow prediction models: process-based and data-driven. Process-based models can be further classified based on their degree of spatial distribution, which includes simple spatially lumped conceptual models, semi-distributed models, and fully distributed models, with the latter having the highest degree of spatial distribution (Fleming & Gupta, 2020). These models are typically interpretable because they incorporate physically meaningful components or simulate the physical processes within watersheds. However, process-based model calibration can be computationally costly and frequently requires prior knowledge of catchments and large amounts of data (Duan et al., 1992). When implementing these models in areas influenced by preferential flow paths, inter-catchment groundwater exchanges, fragmented hydrological connectivity, or regions with uncertain water balance estimates, the complexity could increase substantially due to the violation of physical assumptions under such conditions (Beven, 1986; Liu et al., 2020; Weiler & McDonnell, 2007).

Data-driven models used for streamflow prediction encompass a wide spectrum of complexities, often categorized as statistical models and machine learning (ML) models (Fleming & Gupta, 2020). Among statistical models are parsimonious approaches like simple and multiple linear regression (Garen, 1992; Loague & Freeze, 1985), as well as time-series methods like ARMA (AutoRegressive Moving Average) and ARIMA (Integrated Autoregressive Moving Average) (Delleur & Kavvas, 1978; Spolia & Chander, 1974). ML models, widely adopted for streamflow forecasting, include the support vector machine (SVM) (Asefa et al., 2006; Liong & Sivapragasam, 2002), decision trees (Schoppa et al., 2020; Wang et al., 2015) and artificial neural networks (ANNs). ML models have become increasingly favored for streamflow prediction research due to their robust predictive capabilities, relative simplicity of the parameter calibration process and an absence of biases which often affect physical models (Mosavi et al., 2018). Nevertheless, despite the superior accuracy achievable with ML models, their lack of physical components makes it challenging to enhance process understanding of watersheds through them. Hypothesis testing using ML models is still in its early stages (Nauta et al., 2019), whereas physical models have a well-established history of being employed for hypothesis testing and understanding the role of individual hydrological components in the rainfall/runoff relationship and streamflow prediction.

ML models, especially those employing deep learning techniques, are gaining traction due to the increasing availability of large hydrological datasets for streamflow and streamflow forecasting. Pioneering researchers of applying ANNs for streamflow prediction have demonstrated its enhanced predictive capabilities over simpler machine learning and widely-used conceptual physical models (Dawson & Wilby, 1998; Hsu et al., 1995; Minns & Hall, 1996; Shamseldin, 1997; Tokar & Johnson, 1999). ANNs come primarily in two forms: feed-forward neural networks (FNNs) and recurrent neural networks (RNNs). RNNs, by design, can retain historical input information through their sequential input processing, making them better suited for

sequential data and time-series analysis tasks (Rumelhart et al., 1988). For hydrological prediction, Nagesh Kumar et al. (2004) shows that RNNs outperform FNNs in forecasting river flow. Within the RNNs category, Long Short-Term Memory (LSTM) networks introduced by Hochreiter & Schmidhuber (1997) stand out for their ability to grasp long-term dependencies and resistance to vanishing gradients, positioning them at the forefront for rainfall-streamflow modeling (Kratzert et al., 2018). While few studies have compared the performance of LSTMs with other ML methods for streamflow prediction (e.g. Rahimzad et al., 2021), the LSTM models are widely regarded as being the state-of-the-art for these types of applications compared to other ML models, which also suggests traditional physically-based models are not fully exploiting the information available in the data (Nearing et al., 2021).

Nowadays, only a few ML-based hydrologic prediction approaches are used by government agencies in an operational context for hydrologic forecasting, resource management, or decision support. An often-cited barrier is the perceived lack of interpretability of ML models in terms of physical understanding (Fleming, Watson, et al., 2021). To address the complexity of interpreting neural networks, various methods have been developed, including integrated gradients (Sundararajan et al., 2017), contextual decomposition (CD) (Murdoch et al., 2018), agglomerative contextual decomposition (ACD) (Singh et al., 2019), interpreting transformations (TRIM) (Singh et al., 2021), and penalizing explanations (CDEP) (Rieger et al., 2020). Despite efforts to explain neural networks in hydrology, interpretability remains an ongoing concern in the field. Given this challenge, many agencies choose to apply ML models for testing and evaluation purposes in parallel with established operational forecasting protocols, or to emulate computationally expensive physical models. Examples include the application of an ANN ensemble to Englishman River, a flood-prone stream on Vancouver Island, BC, which was operationally successfully tested during the 2013-2014 storm season (Fleming et al., 2015); the development of a diverse ML-based prototype ensemble for water resources forecast which was used for live operational testing by the Natural Resources Conservation Service (NRCS) at a number of sites in the western US (Fleming et al., 2023; Fleming, Garen, et al., 2021); the use of an ANN to replace the salinity component of the California Department of Water Resources (DWR) Delta Simulation Model II (DSM2) model (CADWR, 2013), allowing its integration into CalSim3, a complex model used for water resources operations in California (Jayasundara et al., 2020); and a deep learning (DL) flood forecasting system developed to provide accurate real-time flood warnings to agencies and the public which became operational in India and Bangladesh during the 2021 monsoon season (Nevo et al., 2022).

To address the challenge of limited interpretability and unknown conformity to physical processes while leveraging the predictive power of ML models, hybrid approaches to forecasting have emerged, known as physics-informed or theory-guided ML models. These physics-aware ML models combine the strengths of machine learning to extract patterns from observational data with the domain knowledge and physical constraints enforcement, and, for hydrology and earth science, are regarded as a potential solution for bringing ML into operational use (Fleming,

Watson, et al., 2021; Slater et al., 2022). The motivation for developing physics-guided machine learning models arises from the recognition that purely data-driven models may inadvertently capture spurious relationships from training data, resulting in physically inconsistent predictions and poor generalization performance (Karpatne et al., 2017). In the context of hydrological prediction, the generalization ability of models pertains to their capability to accurately forecast water-related variables for conditions and geographic areas not encountered during the training phase, including data-scarce watersheds. A hydrological model with strong generalization ability can provide reliable predictions across different watersheds and under varying environmental conditions. The limited generalization capability of ML models is due to extrapolation and observational biases, particularly when data are noisy and insufficient. Adding previous physical knowledge and strong theoretical restrictions is one method to reduce inductive biases in ML models and address this problem (Karniadakis et al., 2021). In general, integrating ML models with physical models holds promise for enhancing physical interpretability and incorporating ML models into operational use (Karniadakis et al., 2021; Karpatne et al., 2017).

The integration of physics with ML has been actively pursued in many fields, including materials science, quantum chemistry, biomedical science, turbulence modeling and earth science (Karpatne et al., 2017; Willard et al., 2020). One integration strategy is to apply ML to a physical model for parameter calibration or error correction. In differential programming (Baydin et al., 2018), an example of the former, a differentiable hydrologic model (i.e., one in which the derivatives of the output with respect to the inputs can be computed analytically) is implemented within a DL platform, in which the DL model learns from data the parameters of the hydrologic model. Demonstration of this approach on the Hydrologiska Byråns Vattenbalansavdelning (HBV) hydrologic model showed drastic improvement compared to the standalone HBV model and suggested that the approach can reduce reliance on large training datasets (Tsai et al., 2020). An extension of the approach to enable DL to replace components of the HBV model (Feng et al., 2023) further improved model performance for the watersheds in the CAMELS database (A. J. Newman et al., 2015). In both of these examples, a standalone LSTM model produced better performance than the hybrid model, but unlike the standalone LSTM, the hybrid models conserved mass and provided information on internal fluxes and stores such as soil moisture and surface runoff. Alternatively, DL can be leveraged to learn and correct for the error in physical model outputs, an approach often called post-processing. With post-processing, outputs of the physical model are provided to the DL model as inputs, together with accessory time-series such as hydrometeorological predictors and hydrologic stores and fluxes internal to the physical model. The resulting hybrid models often demonstrate enhanced performance compared to standalone process-based and LSTM models (Konapala et al., 2020; Adera et al, in review), including for ungauged basins (Lu et al., 2021). However, there are exceptions. While post-processing the output of the National Water Model resulted in performance gains, the hybrid model performed nearly equivalently to a standalone LSTM model and worse than the standalone LSTM model for ungauged basins (J. M. Frame et al., 2021).

The other strategy for integrating physics with ML is to incorporate physics into the training or architecture (Willard et al., 2020) of ML models. For example, physical models can be used to generate synthetic data for training ML models. In hydrology, this approach has been used to increase the representativeness of training data for extreme storm events, with demonstrated improvement in performance (Xie et al., 2021). Physical constraints can also be introduced in the training of ML models, through modification of the loss function in a way that penalizes divergence from physical laws. For example, in a hybrid model of lake water temperatures, a loss function that prioritized energy balance and monotonicity of density in adjacent depth layers produced improvement over both a standalone LSTM model and state-of-the-art physical model (Jia et al., 2019). Lastly, more advanced techniques that involve incorporating physics directly into the architecture of DL models include modifying model cells, conceptualized as water stores within a catchment, to conserve mass (Hoedt et al., 2021). While application of this technique to the CAMELS dataset produced higher performance than other mass-conserving hydrological models, the mass-conserving LSTM exhibited a slightly lower Nash-Sutcliffe Efficiency (NSE) than a standalone LSTM on the CAMELS dataset. Further, the mass-conserving LSTM performed worse than the standalone LSTM for extreme events but better than two physical models (J. M. Frame et al., 2022).

Building upon previous studies that have integrated physical models with ML models, we present a novel physics-informed LSTM model (PILSTM) designed for predicting streamflow. The PILSTM model dynamically combines a reduced-complexity, process-based hydrological model (Kirchner, 2009) with the LSTM model. We incorporate an additional term into the loss function to penalize deviations from the model predictions relative to the physical model outputs. Furthermore, we employ a term to adjust the weight of this loss term, preventing the hybrid model from overly adopting biases from the physical model outputs. By integrating the physical term into the loss function, we constrain the model to make predictions aligning with the fundamental assumptions of the physical model, which are the principles of mass balance and the dynamic correlation between discharge and storage.

Our analysis compares the performance of the PILSTM model with that of a pure physical model and a stand-alone LSTM model under data-scarce and non-stationary scenarios for eight intensively monitored watersheds distributed across the United States. Additionally, we employ integrated gradients methods (Sundararajan et al., 2017) to interpret model predictions and evaluate the influence of each input feature. We predict that the PILSTM model will produce more generalizable and physically-consistent predictions of daily discharge in a wide range of catchments compared to a standalone LSTM. Furthermore, we anticipate that the PILSTM model will outperform both the stand-alone LSTM and the physical model when data are limited and when the training data have different patterns with the testing data.

3.3 Study areas and data

We evaluate our model on eight watersheds sourced from the CHOSEN dataset (Zhang et al., 2021). The CHOSEN dataset is a compilation of observational data from intensively monitored sites within the Long-Term Ecological Research Network (LTER) and the Critical Zone Observatory (CZO) program. The dataset encompasses a wide array of observations, including discharge, precipitation, air temperature, solar radiation, wind speed, relative humidity, vapor pressure, and other hydroclimatic and hydrologic variables. The selected eight catchments from the CHOSEN dataset provide at least 12 years of daily data and exhibit a diverse range of geologic and climatic conditions (Figure 8 and Table 4). Unlike large-sample datasets such as CAMELS, the CHOSEN dataset contains hydrometeorological variables that are measured rather than modeled or interpolated from observations and variables that can be used directly in equations predictive of evapotranspiration (ET) fluxes–both assets that we valued in our model evaluation process. Importantly, the CHOSEN watersheds also provide test cases for the application of LSTM-based models outside the dominant dataset (CAMELS) for which LSTM models have already been extensively tested tuned (e.g., Frame et al., 2021; Kratzert et al., 2018, 2019) and may better represent "new" watersheds under consideration for forecast development.

Given the limited record length of the CHOSEN dataset compared to other large-sample hydrologic datasets, we applied transfer learning to improve model performance. Transfer learning, which involves pre-training a model on a large dataset and then fine-tuning it on a more specific dataset, can significantly boost a model's generalization capabilities (Tan et al., 2018). For model pretraining, we use data from 531 CAMELS-US catchments (Addor et al., 2017; A. J. Newman et al., 2015). These catchments align with those utilized in previous studies by Kratzert et al. (2018) and Newman et al. (2017).

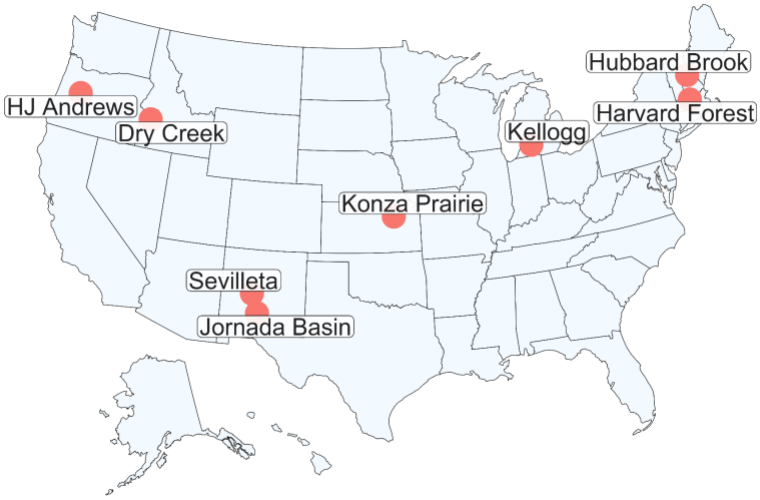


Figure 8. Geographical distribution of the eight study areas.

ET data are required in many water-balance models and are also one of the inputs in the reduced-complexity rainfall-runoff model that we use in this study. We calculate potential evapotranspiration (PET) using the FAO-56 Penman-Monteith method (Allen, 2005), incorporating time series data of precipitation, air temperature, relative humidity, solar radiation, and wind speed available from the CHOSEN dataset. The PyEto python package is employed for PET computation based on these provided variables (Zotarelli et al., 2010). For pretraining with the CAMELS dataset, where data for relative humidity, solar radiation, and wind speed are not available, PET (Potential Evapotranspiration) data is estimated using the Priestley-Taylor equation instead.

Table 4. Basic physiographic information for the eight study areas. The statistics for precipitation, streamflow, and potential evapotranspiration are the annual water fluxes averaging across 12 years. The climate is based on the Koppen climate classification scheme.

Study areas	Dry Creek	H.J. Andrews	Harvard Forest	Hubbard Brook	Jornada Basin	Kellogg	Konza Prairie	Sevilleta
Drainage area (km2)	27	62	0.65	0.77	1976.2	101.8	222.7	13745.1
Elevation (m)	1036	422	330	590	1315	247	382	1478
Precipitation (mm/y)	680	2255	1277	1540	255	1010	835	186
Streamflow (mm/y)	154	1754	633	1167	0.56	336	125	1.35
Reference evapo-transpiration (mm/y)	977	663	1118	917	1722	1221	784	1215
Climate	Cold Semi-arid Climate (BSk)	Warm-summer Mediterranean Climate (Csb)	Humid Continental Mild Summer, Wet All Year (Dfb)	Humid Continental Mild Summer, Wet All Year (Dfb)	Cold Semi-arid Climate (BSk)	Humid Continental Mild Summer, Wet All Year (Dfb)	Humid Continental Hot Summers With Year Around Precipitation (Dfa)	Cold Semi-arid Climate (BSk)
Data time range	2005-10-01, 2017-09-30	2005-10-01, 2017-09-30	2008-01-01, 2019-12-31	1998-01-01, 2009-12-31	2000-01-01, 2011-12-31	2007-01-01, 2018-12-31	2008-10-01, 2020-09-30	2008-01-01, 2019-12-31

3.4 Methods

3.4.1 Introduction of a reduced-complexity rainfall-runoff model

The physical model we use in the PILSTM is a single-equation rainfall-runoff model based on water balance:

$$\frac{dS}{dt} = P - E - Q. \tag{1}$$

P: Averaged precipitation in the catchment (Length/Time)
E: Averaged evapotranspiration in the catchment (Length/Time)
Q: Watershed area-normalized discharge at the outlet of the catchment (Length/Time)
S: Water stored in the catchment per unit area (Length)

The main assumption of this model is the monotonic relationship between discharge (Q) and total water storage (S) in the catchment, represented by a sensitivity function (g(Q)), in which Q is an increasing single-valued function of S (dQ/dS > 0 for all Q and S):

$$\frac{dQ}{dS} = g(Q). \tag{2}$$

We can substitute the g(Q) function into the storage term in eqn.1 to obtain a single-equation rainfall-runoff model (Kirchner, 2009):

$$\frac{dQ}{g(Q)dt} = P - E - Q. \tag{3}$$

This model can be integrated numerically to estimate the value of discharge at the current time step:

$$Q^t - Q^{t-1} = g(Q^{t-1})(P^{t-1} - E^{t-1} - Q^{t-1}); \tag{4}$$

$$Q^t - Q^{t-\Delta t} = g(Q^{t-\Delta t})(P^{t-\Delta t} - E^{t-\Delta t} - Q^{t-\Delta t})\Delta t.$$

We employ the methodology outlined by Kirchner (2009) to estimate the g(Q) function. Specifically, we select for data points where precipitation (P) and evapotranspiration (E) are considerably smaller than Q. This allows us to omit the terms P and E (eqn. 3) for these data points, treating them as approximate recession data (eqn. 5). As a result, we estimate g(Q) solely based on these recession data, assuming a quadratic functional form in the empirical relationship between discharge and the flow recession rate (-dQ/dt) (eqn. 6). This approach circumvents issues related to uncertainty and scaling of the precipitation and evapotranspiration data, which will be revisited later.

$$g(Q) = \frac{dQ}{dS} \approx \frac{-dQ/dt}{Q}\Big|_{P << Q, E << Q} ; \quad (5)$$

$$ln(g(Q)) = c1 + c2 ln(Q) + c3\, ln(Q)^2 . \quad (6)$$

Since the input data series represent single-station PET and precipitation, we follow Kirchner (2009) in using two additional parameters (kE and kP, respectively) to scale the values to approximate actual evapotranspiration and catchment-averaged precipitation:

$$dlnQ = g(Q)\left(\frac{kP \cdot P - kE \cdot E}{Q} - 1\right) . \quad (7)$$

We jointly calibrate these two parameters with the three parameters of (6). The resulting rainfall-runoff conceptual model becomes:

$$lnQ^{t+1} = g(Q^t)\left(\frac{kP \cdot P^t - kE \cdot E^t}{Q^t} - 1\right) + lnQ^t . \quad (8)$$

For each site, the five parameters of the physical model are estimated using a Bayesian optimization algorithm (Pelikan et al., 1999) based on the training dataset.

3.4.2 Basics of the LSTM model

We build the LSTM model with the PyTorch python package (Paszke et al., 2019). During the training process, the coefficients in the model are updated through backpropagation to minimize the loss, using a stochastic gradient-based optimization algorithm (Kingma & Ba, 2017). The LSTM structure enabling feedback from former timesteps is particularly advantageous for emulating streamflow generation processes. We implemented the LSTM model using the code from the NeuralHydrology package, considered as the state-of-the-art and benchmarking code for using LSTM in predicting discharge (Kratzert et al., 2022). More details of the LSTM model are illustrated in Appendix D.

3.4.3 The PILSTM model

We introduce a Physics-Informed LSTM (PILSTM) model (Figure 9) that integrates the outputs from a physical model into the LSTM model. The discrepancy between the PILSTM model prediction and the output of the physical model, which embodies a mass-conserving solution, is incorporated into the loss function as a term with a tunable weight (eqn. 9). This loss function is then employed to train the parameters of the PILSTM model. Through the introduction of this additional loss term, we constrain the model's predictions to better adhere to the conservation-of-mass principle and the dynamic relationship between discharge and storage outlined by the

physical model. The incorporation of this additional loss term aims to ensure that the model generates more physically consistent results, thereby improving the physical interpretability of its predictions. Furthermore, the constraints derived from mass balance have the potential to assist the model in making more reliable predictions under climate variability and data scarcity, thereby enhancing the model's generalization ability.

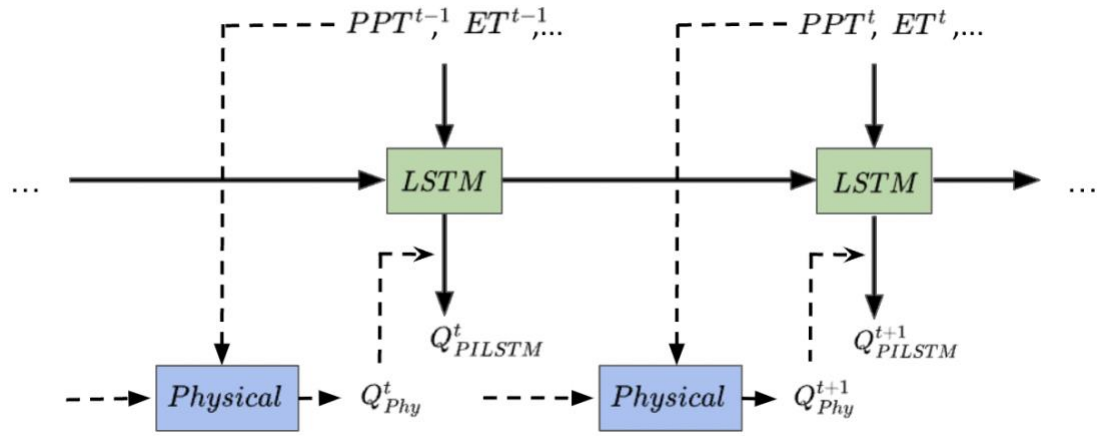


Figure 9. An overview of the PILSTM model. The blue box represents a reduced-complexity rainfall-runoff model. The green box represents an LSTM neural network. PPT and ET are precipitation and evapotranspiration, respectively. Q^t_{obs} and Q^t_{phy} represent observed discharge and the output from the physical model at the t timestamp respectively. $Q^t_{\square}$ is the output of the PILSTM model.

However, the physical model may exhibit bias and provide poor predictive power if its assumptions are not suitable for specific catchments. Thus, we introduce an extra hyperparameter λ, allowing us to optimize for the weighting of the physical term in the loss function (eqn.9). The value of the hyperparameter is fine-tuned using the validation data during training, ranging from 0 to 1 with an increment of 0.1.

$$Loss = (1-\lambda)\, MSE(Q^t_{PILSTM}, Q^t_{Obs}) + \lambda\, MSE(Q^t_{PILSTM}, Q^t_{Phy}) \quad . \tag{9}$$

λ : hyperparameter in the loss function
Q^t_{PILSTM} : prediction from the PILSTM model at timestamp t
Q^t_{Obs} : observation of discharge at timestamp t
Q^t_{Phy} : integrated discharge from the physical model at timestamp t
MSE : mean squared error

In the workflow for the PILSTM model with pretraining, we initially train the LSTM model on CAMELS sites, utilizing data from October 1, 1999, to September 30, 2008. We employ an early-stopping method to determine the optimal number of training epochs, indicated by the highest median NSE across sites, using validation data from October 1, 1980, to September 30,

1989. For other hyperparameters, we adopt values following the pretraining example in the NeuralHydrology codebase (Kratzert et al., 2022) (Table 5). Subsequently, we fine-tune the model for each CHOSEN site individually, employing early stopping to identify the optimal training epochs with validation data. Regarding hyperparameters in the fine-tuning phase, we maintain the same set as in the pretraining phase, except for a reduced piecewise learning rate. For the PILSTM model without pretraining, we skip the pretraining step and directly train the model for each individual watershed with the hyperparameters in the fine-tuning phase.

Table 5. Hyperparameters in the LSTM model

Training phase	Hidden size	Input sequence	Learning rate	Dropout	LSTM layers #	Maximum training epochs #	Batch size
Pretraining	128	365	0-10: 1e-3 11-20: 5e-4 21-30: 1e-4	0.4	1	30	256
Fine-tuning	128	365	0-10: 5e-4 11-20: 1e-4 21-30: 5e-5	0.4	1	30	256

3.4.4 Integrated gradients

We employ the integrated gradients method (Sundararajan et al., 2017) to quantify the importance of each input feature to predictions in order to study how meteorological forcings affect discharge in the models. The method calculates the gradient of the prediction with respect to the model input integrated from a chosen baseline:

$$IntegratedGrads_i(x) = \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha. \quad (10)$$

In comparison to previous deep neural network interpretation methodologies, this metric emphasizes feature importance sensitivity, which is the influence computed along the pathway from the baseline. Following Kratzert, et al. (2019), we use zero values as the baselines (i.e., 0 mm of precipitation and evapotranspiration) and an integration step of 1000. We perform the integrated gradients analysis on the testing dataset for a single run of the base experiment for each watershed. feature importance was computed for each data point and then averaged across the day of the year. In order to accentuate the effect of physical information in the comparison

between the PILSTM and the LSTM, we set λ to one for all of the PILSTM models used in this experiment.

3.4.5 Experiments

In this study, we evaluate the performance of five different models:

(1) a stand-alone LSTM model with precipitation (PPT) and air temperature (AT) as inputs
(2) a stand-alone LSTM model with PPT and potential evapotranspiration (PET) as inputs
(3) a PILSTM model with PPT and AT as inputs
(4) a PILSTM model with PPT and PET as inputs
(5) a stand-alone physical model with PPT and PET as inputs

In particular, we compare the performance of the PILSTM model to that of the LSTM model with the identical set of inputs. Although the physical model uses PET as a necessary input, because PET is a derived time series, models 1 and 3 are used as additional benchmarks that make no a priori assumptions about the partitioning of PPT.

We evaluate model performance in three types of experiments (Table 6). The first assesses the relative performance of the five models and the optimal weighting of physical information in each watershed. The second evaluates model performance under alternate ways of sampling climate variability. Given projections indicating more frequent and intense heat and precipitation extremes with vanishing cold extremes under ongoing global warming (Li et al., 2021), we anticipate corresponding shifts in discharge levels toward drier or wetter scenarios. As a result, we partition the data to simulate scenarios ranging from dry training years and wet testing years to the opposite scenario of wet training years and dry testing years. In the third experiment, we evaluate model performance under situations with limited training data. All experiments were performed under both site-specific and pretraining scenarios, utilizing an ensemble approach initialized with distinct random seeds. The reported results are the averages obtained from 10 ensemble runs on the testing dataset.

Table 6. Experimental Design and Corresponding Methodological Results.

Experiment	Methodological Details
Base Experiment	For each CHOSEN site, the model is trained on the initial six years of data, validated on the subsequent two years, and tested on the last three years.
Non-stationary Scenario	We implement a sliding window approach, moving the five years of validation and testing data across the entire data span to create seven different data splits. We calculate the average discharge levels for the

	training and testing data, employing a dry-wet-index to indicate the relative dryness ranking in the training data compared to the testing data. For instance, a dry-wet-index of 1 means that the data split has the relatively driest training years and wettest testing years, while a dry-wet-index of 7 indicates the opposite scenario with the wettest training years and driest testing years.
Data Scarcity Scenario	We progressively decrease the training data duration from six to one year while preserving the last five years of the data sequence for validation and testing. Within the five years of data, the initial two years are designated as the validation dataset, and the subsequent three years are utilized as the testing data.

3.4.6 Evaluation Metrics

To evaluate model performance, we use the Nash-Sutcliffe Efficiency (NSE), the Pearson correlation coefficient (r), and bias term in the Kling–Gupta efficiency (Gupta et al., 2009). We also calculate the bias in the flow duration curve (FDC) high-segment volume (FHV), bias in FDC mid-segment slope (FMS), and bias in FDC low-segment volume (FLV) metrics (Appendix E; Yilmaz et al., 2008). We additionally develop new indices that represent model overall accuracy, the stability of models and their physical consistency with the water balance.

The overall model accuracy is measured by the mean NSE across 10 ensemble runs. Model stability is quantified using the standard deviation of the NSE values in the 10 ensemble runs.

We assess the physical consistency of our predictions by examining the closure of the water balance within each water year cycle. The physical consistency score is defined based on the water storage deficit averaged across all water years (eqn.11), where a score of 1 indicates a perfect match between the output mass and the input mass. Although it is common to assume water balance closure on annual timescales, it is widely recognized that many watersheds–particularly small ones–violate this assumption (Safeeq et al., 2021) and hence this metric serves as an imperfect means to quantify physical consistency. Nevertheless, we include it as an expedient way to assess tradeoffs among multiple dimensions of model performance.

$$\text{Physical consistency score} = \frac{1}{n}\sum_{water\ year\ idx=1}^{n} 1 - \frac{|\sum(PPT-ET-Q)|}{\sum PPT} \quad . \qquad (11)$$

In the base experiment, we employ additional metrics to assess the model's overall performance tradeoffs concerning prediction accuracy, stability, and its adherence to the water balance on the testing data on the pretrained model. We visualize these metrics across a spectrum of λ values and check model performance with changing λ values. To facilitate a comparison of metrics and

determine the optimal λ value considering all three metrics, we normalize the model accuracy, stability, and physical consistency scores to a range of zero to one by subtracting the minimum value and then dividing by the maximum (Figure 14). This normalization allows for a consistent scale, enabling a comprehensive analysis of the model's performance across different evaluation criteria.

3.5 Results

3.5.1 Model performance comparison for base experiment

Pretraining improved the accuracy of LSTM-based models, boosting median NSE values from well below to above 0.5 (Figure 10). The distribution of LSTM-based model performance with pretraining lay within the distribution of model performance for LSTM models with static inputs in the CAMELS database (Kratzert, Klotz, et al., 2019), but with a lower median value than the 0.73 from the CAMELS benchmark. Nonetheless, three out of eight watersheds met or exceeded the benchmark median (H.J. Andrews and Harvard Forest; Table 7B), suggesting that performance of the LSTM-based models with pretraining approached state-of-the-art. Pretraining mitigated underestimation of values of high- and medium flow (captured by the FHV and FHM statistics) in most watersheds but resulted in biases greater than one (meaning the mean simulated streamflow is higher than the mean observed streamflow), compared to less than one for site-specific models.

In both pretraining and site-specific scenarios, the PILSTM model consistently outperformed its LSTM counterpart, although the advantages of PILSTM over LSTM models were less pronounced with pretraining (Figure 10). With pretraining, the PILSTM generally outperformed the physical model, though without pretraining, the physical model exhibited better performance with respect to *r*, bias, and FHV and similar performance with respect to the NSE. Even compared to pretrained LSTM-based models, the physical model exhibited the best performance with respect to bias. The pretrained PILSTM model, when incorporating PPT and PET inputs, demonstrated superior performance compared to its counterpart using PPT and AT inputs. Conversely, the site-specific PILSTM model featuring PPT and AT inputs exhibited slightly better performance compared to the configuration with PPT and PET inputs.

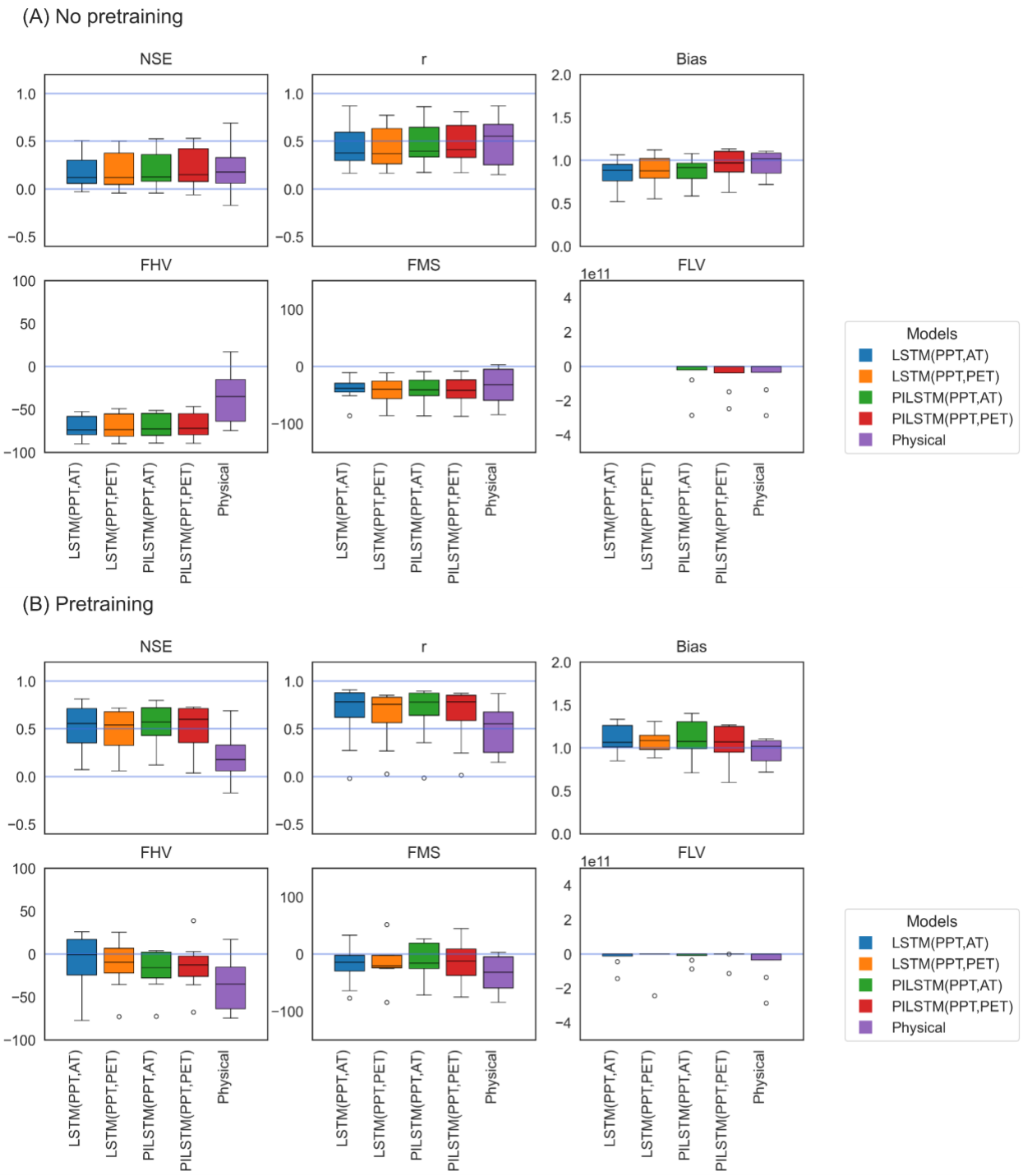


Figure 10. Model performance in the base experiment across eight watersheds (A) without pretraining (B) with pretraining. PPT stands for precipitation, while AT and PET refer to air temperature and potential evapotranspiration, respectively.

While site-specific models that leveraged physical information resulted in the best performance for nearly all watersheds, the value of physical information diminished substantially with pretraining, which alone resulted in large performance gains. Specifically, without pretraining,

PILSTM models showcase the optimal performance for three out of eight watersheds. The physical model exhibits the best performance for four watersheds (tied with the PILSTM model for one of those four), and the LSTM model excels in delivering the best performance for the Jornada Basin (Table 7A). With pretraining (Table 7B), the best performing model NSE values for each watershed increase substantially, with the exception of Jornada Basin and Sevilleta. The PILSTM exhibits the best performance for one watershed, while the LSTM model performs best in five watersheds. The physical model performs best in the remaining two watersheds, which happen to be the driest (Jornada Basin and Sevilleta). In contrast to the overall improvement in model accuracy for most watersheds, the NSE for the LSTM-based models for Jornada Basin decreases substantially with pretraining, while those for Sevilleta increases only marginally. Meanwhile, whether the use of PET or AT inputs results in better performance varies across watersheds.

Table 7A. Model NSE in the base experiment across eight watersheds without pretraining. Boldface indicates the best-performing model for each watershed.

Watershed	LSTM (PPT,AT)	LSTM (PPT,PET)	PILSTM (PPT,AT)	PILSTM (PPT,PET)	Physical
Dry Creek	0.51	0.5	0.53	0.53	**0.69**
H.J. Andrews	0.38	0.41	0.43	0.45	**0.67**
Harvard Forest	0.15	0.13	0.16	0.18	**0.22**
Hubbard Brook	0.09	0.11	0.1	**0.12**	0.03
Jornada Basin	**-0.03**	-0.04	-0.04	-0.06	-0.17
Kellogg	0.28	0.37	0.34	**0.41**	0.17
Konza Prairie	0.06	0.05	**0.07**	**0.07**	**0.07**
Sevilleta	0.05	0.03	0.08	0.08	**0.19**

Table 7B. Model NSE in the base experiment across eight watersheds with pretraining. Boldface indicates the best-performing model for each watershed. Where the best-performing PILSTM has an optimal λ value of zero, the LSTM is denoted as best performing, even when its NSE is a few percentage points lower than the PILSTM (as a result of stochasticity in training and optimization).

Watershed	LSTM (PPT,AT)	LSTM (PPT,PET)	PILSTM (PPT,AT)	PILSTM (PPT,PET)	Physical
Dry Creek	0.69	0.68	0.7	**0.73**	0.69
H.J. Andrews	**0.82**	0.61	0.8	0.72	0.67
Harvard Forest	**0.79**	0.72	**0.79**	0.71	0.22

Hubbard Brook	**0.54**	0.48	**0.55**	0.46	0.03
Jornada Basin	-1.51	-4.29	-1.36	-3.48	**-0.17**
Kellogg	0.45	**0.69**	0.53	0.65	0.17
Konza Prairie	**0.58**	0.42	**0.6**	0.55	0.07
Sevilleta	0.07	0.06	0.12	0.04	**0.19**

Optimal λ values change with pretraining, across watersheds, and with model inputs (Table 8). With pretraining, the PILSTM model relies less on physical information, as evidenced by a decrease in the optimal λ values for most cases, with the exception of Dry Creek and Kellogg. Across all watersheds, Dry Creek and H.J. Andrews consistently show higher λ values. Coincidentally, their physical model accuracies are also superior to those of other watersheds. In comparing models with PET vs. AT as inputs, there was no consistency in which set of inputs resulted in a higher optimal λ.

Table 8. Optimal λ values for PILSTM models. The values are determined based on the validation data, where the NSE achieves its highest point in the comparison between model outputs and observed discharge. Boldface corresponds to models that were the best-performing within each set of experiments (i.e., without pretraining, with pretraining), based on NSE (Tables 7A and 7B).

	Without pretraining		With pretraining	
Watershed	PILSTM (PPT,AT)	PILSTM (PPT,PET)	PILSTM (PPT,AT)	PILSTM (PPT,PET)
Dry Creek	0.33	0.55	0.09	**0.86**
H.J. Andrews	0.57	0.6	0.03	0.49
Harvard Forest	0.35	0.26	**0**	0
Hubbard Brook	0.18	**0.08**	**0**	0.02
Jornada Basin	0.28	0.26	0.36	0.03
Kellogg	0.16	**0.21**	0.14	0.25
Konza Prairie	**0.14**	**0.1**	**0**	0
Sevilleta	0.11	0.32	0.04	0.03

3.5.2 Experiments under non-stationary scenarios

Without pretraining (Figure 11A), all the PILSTM models slightly outperform their LSTM counterparts in the non-stationary scenarios. The median of the PILSTM models also generally exceeds that of the physical model, though the maximum value of NSE for the physical model is

generally highest. Although the performance of the physical models generally declines with wetter training years/drier testing years, no trend is apparent in the LSTM-based models. Likewise, no trend is apparent in the optimal λ values across the wet-dry indices. These values, consistently near 0.21, indicate models that are primarily data-driven but with some influence of the water-balance constraint.

Compared to the site-specific model results, pretraining increases the median NSE of the LSTM-based models but decreases the minimum and lower quartile (Figure 11B). The lower NSE values in the pretraining experiments mainly stem from the Jornada Basin watershed, which also demonstrates poor model accuracy in the base experiment. As with the site-specific results, there is no trend in optimal λ values across the wet-dry indices for the PILSTM models that use AT inputs. For those that use PET, a slight increase in optimal λ values as the training years become wetter indicates slightly greater leveraging of physical information under these conditions.

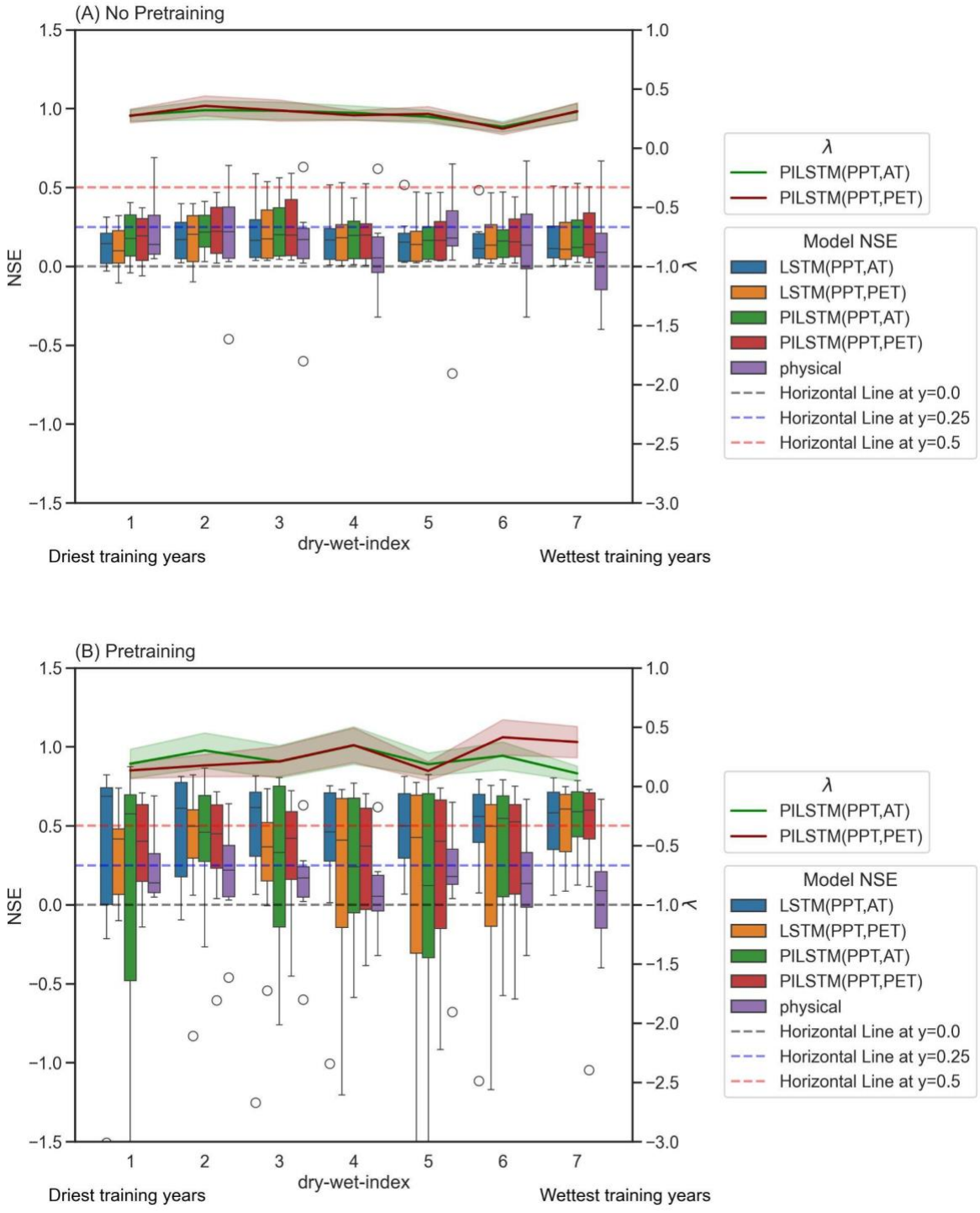


Figure 11. NSE values across eight watersheds for models tested on different splits of training and testing years. A dry-wet-index of 1 means the training years are the driest and the testing years are the wettest based on the annual discharge volume, and a dry-wet-index of 7 means the training years are the wettest and the testing years are the driest. The lines represent the means and standard errors of the optimized values of λ in the PILSTM models across eight watersheds.

Figure A shows results of models without pretraining. Figure B shows results of fine-tuning models with pretraining.

3.5.3 Experiments under data-scarce scenarios

In evaluating the effectiveness of site-specific LSTM-based models and a physical model in scenarios with limited data availability, we observed an improvement in prediction accuracy with an increase in the amount of training data (Figure 12A). This enhancement is evident in the rising median NSE values across diverse watersheds. The performance advantage that the PILSTM models exhibit over the LSTM models diminishes as the record length decreases and is negligible for four training years or less. Compared to the LSTM-based models, the physical model exhibits more stability of the median and upper-quartile NSE values over the range of data lengths, but minimum and lower-quartile NSE values drop substantially for two training years or less. Meanwhile, optimal λ values for the PILSTM remain consistently near 0.26 over the range of record lengths.

In scenarios where pretrained models undergo fine-tuning for each watershed, we observe a significant enhancement in median and upper-quartile model accuracy compared to site-specific models (Figure 12B). Similar to site-specific models, LSTM-based models benefit from an increasing number of training years. However, relative to the site-specific models, the minimum and lower-quartile NSE values decrease substantially with pretraining under data-scarce conditions. This decrease in performance is less pronounced for the PILSTM models (particularly with AT inputs) than for the LSTM models when more than one year of training data is available and is less pronounced for the LSTM using PET instead of AT inputs. Meanwhile, with regard to the median and upper-quartile NSE values, the PILSTM has an advantage over the LSTM model for more than five training years. As with the base case, optimal λ values are more variable for the pretrained models compared to the site-specific models. PILSTM models using AT inputs leverage more physical information as the data length decreases, while those using PET inputs exhibit no trend in optimal λ values but leverage more physical information than their AT-input counterparts for longer record lengths.

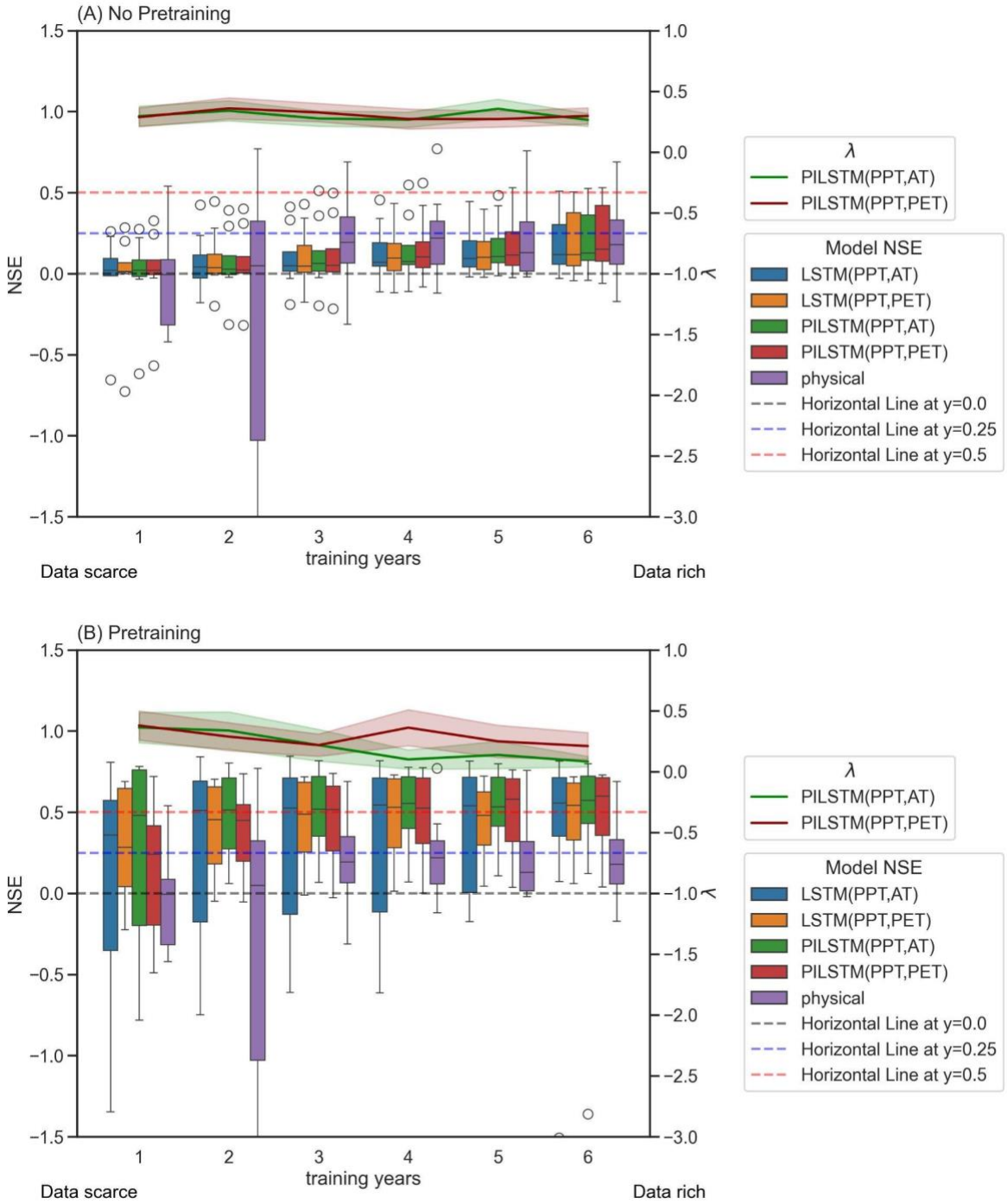


Figure 12. NSE values across eight watersheds for models trained on varying record lengths in (A) Site-specific and (B) Pretraining cases. The lines depict the means and standard errors of the optimized λ values in the PILSTM models. Both NSE values and λ values are averaged across 10 ensemble runs.

3.5.4 feature importance

The feature importance of PPT and PET on discharge varies across watersheds and over time (Figure 13). In the site-specific models (Appendix F), feature importance patterns are noisy, with PPT typically having a positive influence on discharge and PET typically having a negative influence. Generally, feature importance patterns for the site-specific models do not exhibit pronounced seasonal variations. However, in the pretrained models, many watersheds exhibit a similar annual pattern in the feature importance of PPT, in which PPT has a low but positive influence on discharge in autumn and winter and an increasing influence through the snowmelt season, peaking in late spring or early summer. For most watersheds (Dry Creek, Harvard Forest, Hubbard Brook, Kellogg, Konza Prairie), the feature importance of PPT is lower and less seasonal (i.e., flatter) for the LSTM model than for the PILSTM model. Exceptions include Jornada Basin, for which the feature importance of PPT is similar between the LSTM and PILSTM, Sevilleta, for which the LSTM has a stronger and more seasonal feature importance, and HJ Andrews, for which the LSTM generally has a stronger but highly noisy feature importance. Another difference between the LSTM and PILSTM models is that for some watersheds (Hubbard Brook, Jornada Basin, Sevilleta), the feature importance of PPT is negative during the spring for the PILSTM model but not the LSTM model.

The feature importance of PET also exhibits a characteristic temporal pattern that is generally consistent across watersheds, with the most negative values (i.e., the most negative influence on discharge) in spring through summer and near-zero values through the winter. For several of the watersheds, PET exhibits a positive feature importance during the snowmelt season, when it potentially serves as a proxy for AT, which was not included separately as a model input. Patterns in the feature importance of PET are generally similar for the LSTM and PILSTM models, with a few exceptions. In Hubbard Brook and Harvard Forest, the PILSTM model has a more negative feature importance during the summer, whereas, in contrast, feature importance for the LSTM model remains positive in Hubbard Brook. Meanwhile, in HJ Andrews, feature importance is substantially more variable (i.e., higher magnitude positive and negative spikes) for the LSTM model than for the PILSTM model.

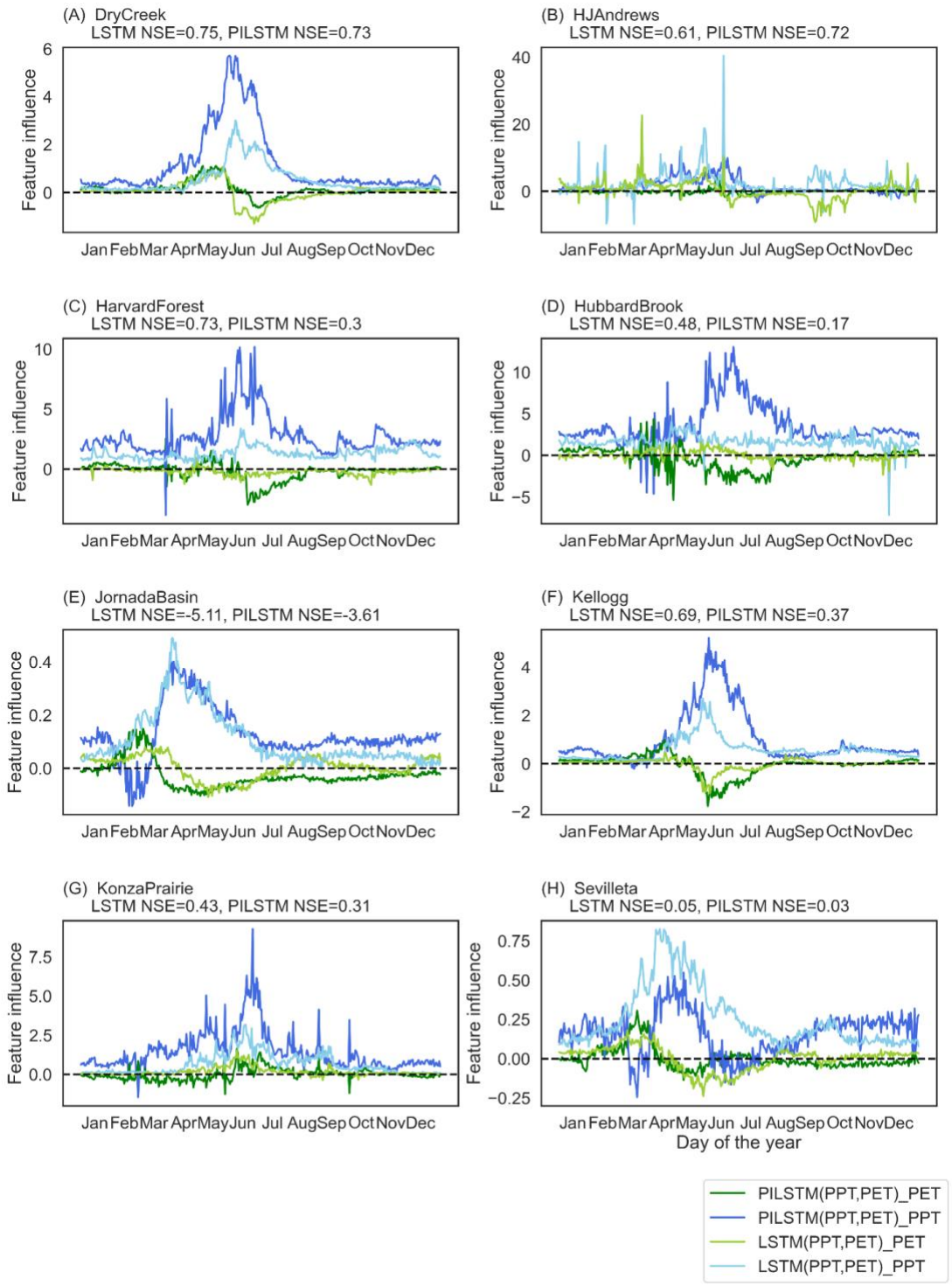


Figure 13. Impact of input features using the integrated gradients method for pretrained PILSTM models. All PILSTM models employ a λ value of one.

3.5.5 Tradeoffs among model accuracy, stability, and physical consistency across a gradient of data-driven to physical models

As the weight on the physical constraint in the loss function (λ) varies from zero to one, the model accuracy, stability, and physical consistency scores exhibit unique variations and tradeoffs for each watershed in the base case with pretraining (Figure 14). Except for the Jornada Basin watershed, the optimal λ value at which the combined value of the three scores is the highest (i.e., the "optimal") is greater than zero, suggesting that a PILSTM typically performs best in a multi-objective tradeoff. Regarding individual scores, trends with λ vary widely. In two out of the eight watersheds, accuracy increases as λ increases. Three watersheds show highest stability at the purely data-driven end of the spectrum ($\lambda = 0$), another three at the physical end of the spectrum, and the remaining two at intermediate values. Meanwhile, physical consistency achieved its maximum value at the physical end of the spectrum for just three watersheds and was maximal at the data-driven end of the spectrum for the two driest watersheds (Jornada Basin and Sevilleta).

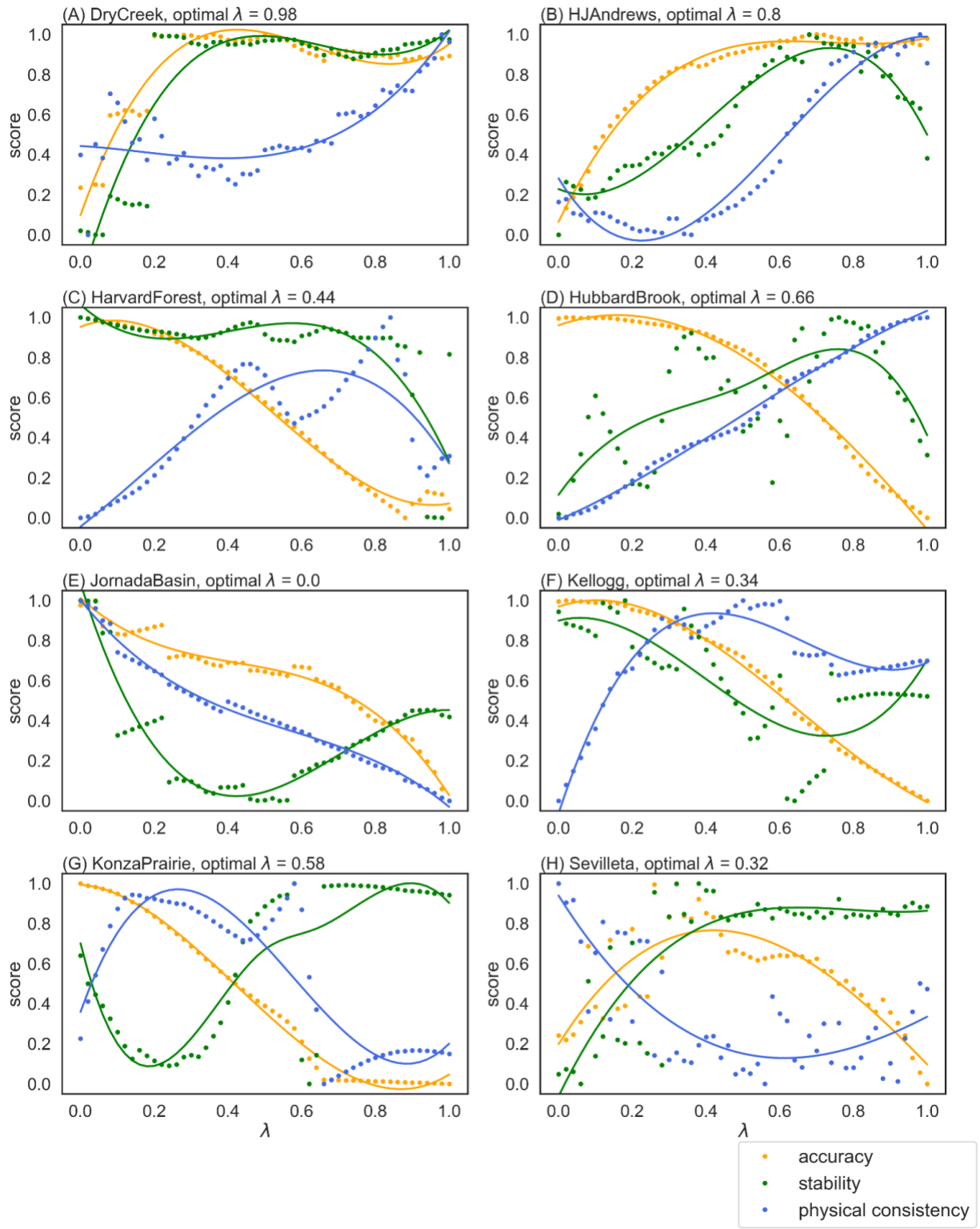


Figure 14. Normalized scores depicting model accuracy, stability, and physical interpretability across varying λ values in the base case with pretraining and PET inputs. A λ value of 0

indicates a purely data-driven model, while a λ value of 1 represents an approximation of a physical model.

3.6 Discussion

Whether and how physical knowledge should be used to assist and improve data-driven models has been an ongoing discussion (Beven, 2020; Nearing et al., 2021). It is acknowledged that in hydrology, the applicability of a one-size-fits-all physical model is limited, and certain physical assumptions may hold true for specific watersheds rather than universally (Hrachowitz et al., 2013). Consequently, the quest for a universal hydrological theory that seamlessly supports data-driven models poses a significant challenge. While acknowledging the imperfections inherent in individual models, we propose that by judiciously combining them, their collective power could surpass the performance of each model in isolation. In this study, we demonstrate that integrating a physical model with the LSTM model for discharge prediction often yields improved performance compared to both the stand-alone LSTM model and the physical model. It is noteworthy, however, that the efficacy of this hybrid model is more apparent in certain scenarios and less pronounced in others.

Diagnosing and predicting differences in model performance across experiments and watersheds

Throughout the sets of experiments, the watersheds in the CHOSEN dataset exhibited great variability in multi-metric model performance, sensitivity to undersampling of climate variability and data scarcity, and in the extent to which integration of physical information improved model performance. We found that, for places with frequent storms, where recession parameters are uncertain, and/or snow (which the physical model did not represent) falls and melts throughout the winter, the pure data-driven model did best. In dry places that are baseflow dominated, where storage processes longer than the timescales over which the LSTM is sensitive to were dominant, the purely physical model did best. Meanwhile, the PILSTM performed best in semiarid to wet environments with distinctive seasonality in PPT and/or streamflow (Table 7B).

Feature importance plots provide a means of "looking under the hood" of LSTM-based models to diagnose and understand these discrepancies and differences in model performance. Our examination of the feature importance plots for the LSTM and PILSTM model with λ set to unity allowed us to visualize how models at opposite ends of the data-driven to physical spectrum represented seasonal forcing of stream discharge by PPT and AT or PET. These visualizations allowed us to see, for example, that seasonal forcings are much smoother when models leverage pretraining data from 531 CAMELS-US catchments than when they do not (Appendix F compared to Figure 13), suggesting that even with the six or more years of data associated with the CHOSEN watersheds, LSTM-based models that do not involve pretraining face data limitations that can potentially lead to overfitting, a widely recognized problem with

neural networks (Srivastava et al., 2014). These limitations, however, can be largely overcome with pretraining.

Among those three pretrained models that benefited from leveraging substantial amounts of physical information (i.e., Dry Creek, HJ Andrews, and Kellogg watersheds, which had moderate to high optimal values of λ), the feature importance plots provide information about how physical information might have produced improvements. The seasonal pattern in PPT having maximal influence on discharge toward the end of the snowmelt season, when watersheds are the most connected, is strongest for the PILSTM in both Dry Creek and Kellogg and potentially more reflective of reality. Although the physical model does not explicitly represent snow dynamics, its representation of a single dominant storage reservoir that releases water slowly is likely adequate for these snowmelt-dominated watersheds and provided additional information about long-term storage and storage release (for example, the requirement to route PPT to storage when storage is depleted) that the pure LSTM was unable to learn. For Dry Creek, the timing of when the influence of PET became negative (reflecting when it switched from a proxy for air temperature and snowmelt to when it dominantly represented actual evapotranspiration and water depletion) also differed between the PILSTM and LSTM. The later shift in the PILSTM better coincides with the end of the snowmelt peak in discharge and may be more physically representative. Meanwhile, the feature importance plots for HJ Andrews were not as seasonally variable as in Dry Creek and Kellogg but still reflected more physically realistic conditions for the PILSTM model. Namely, the PILSTM model was much smoother than its LSTM counterpart, suggesting that at HJ Andrews, the incorporation of physical information may have suppressed the overfitting apparent in the LSTM.

Feature importance plots also provide insight into how physically informed models may be inadequate for those watersheds with optimal λ values of zero or near-zero. In Harvard Forest, Hubbard Brook, and the Konza Prairie, the PILSTM exhibited a more seasonal influence on discharge than the LSTM. In these three watersheds that receive substantial convective as well as frontal PPT, that do not accumulate large snowpack, and that are humid, it is expected that hydrologic connectivity would remain high through the year and that the feature importance of PPT would thus be flatter. Greater difficulty in calibrating parameters for the physical model in these watersheds—including the challenge of shorter intervals between storms during which streamflow recession behavior can be observed and fewer data points that meet the assumptions of calibration ($P \ll Q$ and $ET \ll Q$) likely contribute to the uninformativeness of physical information in these catchments. With the limitations on the calibration data, it is likely that parameter calibration is overrepresenting the behavior of large storage reservoirs that release water slowly and underrepresenting smaller, faster draining storage reservoirs, resulting in a more pronounced seasonal pattern in the feature importance plot for the PILSTM with λ equal to one. If this analysis were, however, performed on subdaily data, we expect that the physical model would be easier to calibrate, given the availability of nighttime data for which ET can

effectively be assumed equal to zero, as in Kirchner's (2009) original analysis. We expect that optimal λ values for models developed with subdaily data would be substantially higher for watersheds like these three, and that the overall performance advantage of PILSTM models would be higher.

Diagnosing the poor performance of Sevilleta and Jornada Basins, both of which also have an optimal λ value of zero, requires insight beyond the feature importance plots. Despite their optimal λ value, the physical models for those watersheds paradoxically performed best, and none of the models had good performance (Table 7B). We can resolve the paradoxical optimal λ value of zero by concluding that the LSTM architecture did not emulate the physical model well for these sites, potentially because it was unable to capture the longer storage times driving baseflow in these landscapes. The feature importance plots support the conjecture of poor physical model emulation by showing a negative influence of PPT in the PILSTM for February-March, which is not physically realistic. This negative influence of PPT is paired with a positive influence of PET on discharge, which, in contrast to watersheds where PET serves as a proxy for AT and melt rate during this time of year, is unrealistic for these watersheds that receive little to no snow. In general, we expect that parameter optimization for LSTM-based models would be particularly poorly constrained in these watersheds because of limited variability in flow over the data record. The poor performance of all types of models for these watersheds may also reflect their large area and the inadequacy of single-station values of PPT and PET at that scale.

Extrapolating these findings to other watersheds, we predict that when a physical model is not representative and particularly when there is high uncertainty in parameterization, a purely data-driven approach would be preferable when training data are sufficient. On the other hand, when watersheds are arid and baseflow-driven, with limited streamflow variability, and when the timescale that the LSTM's memory captures is not commensurate with the timescale of dominant watershed processes, a purely physical model would be preferable. Otherwise, a PILSTM is probably best, and, as will be discussed in the subsequent sections, is likely most robust to shortcomings in the training data, including short record length and undersampling climate variability. We recommend more comprehensive studies that span many watersheds to develop guidance on how watershed characteristics relate to the weighting of physical vs. data-driven modeling elements (i.e., the value of λ).

Greater advantages for model pre-training than incorporation of physics, with caveats
Across the base, non-stationary, and data scarcity experiments, it was clear that pretraining LSTM-based models on a geographically broad dataset resulted in large multi-metric performance gains for most watersheds that were far greater than gains from integrating physics into LSTM-based models or changing the inputs to those models. Leveraging the spatiotemporally rich data in the CAMELS database resulted in less overfitting of LSTM models, as evidenced in much smoother feature importance plots (Figure 13 compared to Appendix F),

and likely made up for missing information in the training data from the target sites. Pretraining produced advantages in nearly all performance metrics and resulted in diminished sensitivity of median performance to record length of and climate variability in the fine-tuning data, indicating that the learning leveraged in pre-training resulted in fewer incidences of extrapolation. These findings reinforce recent recognition (Qualls, 2022) that pretraining may help overcome long-recognized problems in ML of predicting discharge during dry periods when models are trained in wet periods (e.g., Vaze et al., 2010) or other out-of sample conditions. In the site-specific (i.e., non-pretrained) LSTM-based models, a decline in performance was apparent when just a single training year was withheld from the fine-tuning data (Figure 11A). In contrast, in the pretrained LSTM-based models, the performance of the median and upper-quartile sites was remarkably stable with declining training years, until two or fewer years were available (Figure 11B). Even with just one or two years of fine-tuning data, though, median and upper-quartile performance was better than for the site-specific LSTM-based models, suggesting that most of the dynamics relevant to these watersheds were already present in the pretraining dataset. Meanwhile, the decline in performance below two years of fine-tuning data suggests that site-specific training information is still useful for improving performance.

However, for those watersheds at the lower end of the distribution of model performance, pretraining can worsen performance under conditions of undersampling climate variability or data scarcity. This worsened performance likely results from the model "learning" behaviors from the pretraining dataset that are not applicable to the target site when the target site is poorly represented in the pretraining data and fine-tuning being insufficient to correct for counterproductive transfer learning. In the CHOSEN dataset, it was the dry sites with large watershed areas (Jornada Basin and Sevilleta) that consistently delivered poor performance in pretraining for the non-stationary and data scarcity experiments.

Safeguarding against poor performance through site-specific data and integration of physics

Our results suggest that the poor performance of some pretrained models may be alleviated when the fine-tuning data record is sufficiently long (i.e., six or more years for all LSTM-based models to consistently see performance gains through pretraining) and well-distributed (summarized in the decision tree in Figure 15). Here, well-distributed means that the fine-tuning dataset ideally samples nearly the full range of climate variability; when climatically similar years were excluded from training in the non-stationary experiments, lower-quartile performance values were lower than those from site-specific models. The only exception was when the driest years were excluded from training, suggesting that learning specifically from hydrologically active periods in the fine-tuning dataset (particularly important for the arid sites from which poor performances originated in our dataset) can alleviate poor performance arising from pretraining. Additional or alternative types of information can correct for this counterproductive transfer learning. Our results suggest that transfer learning may be more effective when based on direct water balance quantities rather than derived quantities. For example, in contrast to the

counterproductive transfer learning seen in the LSTM with AT inputs for cases with fewer than six years of fine-tuning data, for the LSTM with PET inputs, the most poorly performing watersheds exhibited better performance in the pretrained than in the site-specific models with just three or more years of fine-tuning data. This finding suggests that the globally trained LSTM had difficulty in extracting information from AT dynamics learned from other sites that was relevant to the water balance at the target site. Second, physical information, introduced via the PILSTM, can also improve the performance of lower-quartile sites in the pretrained models, with even greater gains apparent than in the LSTM with PET inputs. Still, a minimum of two- to four years of training data is needed to improve the worst performances, given the challenges of calibrating physical models and achieving high performance under data-scarce scenarios. Interestingly, PILSTM models developed with AT inputs had highest minimum and lower-quartile performances, suggesting that the imposition of assumed evapotranspiration values in the water-balance model coupled with flexibility for the LSTM to translate AT into discharge best corrected for counterproductive transfer learning in the pre-training process.

The findings in this study are complementary to other studies that have demonstrated that physically constrained ML models outperform standard ML models when testing data are distributed differently from training data. For example, Lu et al., 2021 used an LSTM and a physics-informed hybrid LSTM to forecast discharge in a data-scarce rural area and found that the hybrid model performed better when the variability of the prediction and calibration periods was substantially different. Similarly, Wang et al., (2020) suggested that theory-guided neural networks were more effective than standard deep neural networks for subsurface flow prediction due to their ability to generalize to scenarios outside of the training dataset. Collectively, these works and our experiments suggest that a combination of ML and physical models may be highly effective in handling nonstationarity when predicting discharge and that physics-guided ML models can leverage the information from limited training data and improve predictive power compared to both process-based and data-driven models.

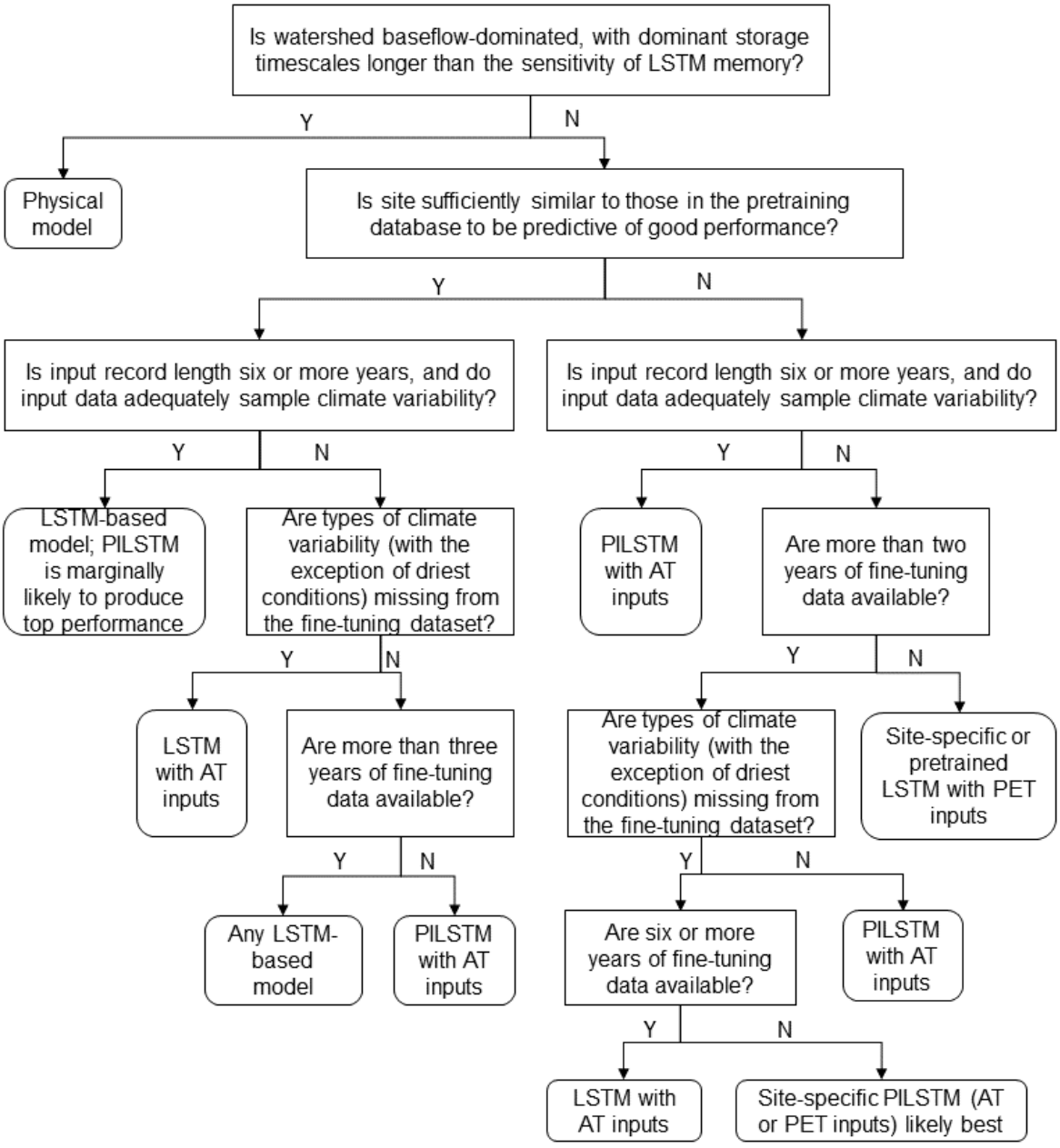


Figure 15. Phenomenological decision tree for developing the highest-performing predictive model of streamflow from precipitation (PPT) and either air temperature (AT) or potential evapotranspiration (PET) inputs, based on analysis of the CHOSEN watersheds. Recommended LSTM-based models are pre-trained unless otherwise indicated (i.e., referred to as site-specific models). Larger-sample studies would be needed to develop certainty around decision points involving specific record lengths, assessment of the degree of similarity between target watersheds and those in a pretraining database that is predictive of good (i.e., near-median and above) performance, and the adequacy of sampling climate variability in the test data.

Transfer learning with DL models has greatly accelerated the improvement of PUB (Kratzert, Klotz, et al., 2019), but remnant performance differentials between gauged and ungauged watersheds underscore the importance of site-specific data for fine tuning. Here we show that, for certain watersheds with characteristics poorly represented in the pretraining dataset, transfer learning can diminish performance (relative to site-specific DL models) as the amount of fine-tuning data decreases (to five years and below). However, incorporating water-balance constraints into the loss function and tuning the weight of those constraints can generally compensate for counterproductive transfer learning when two or more years of data are available (provided those years reasonably sample expected climatic variability, with the possible exception of the driest conditions), serving as a safeguard against poor performance (Figure 15). Similarly, training the model on input time-series directly relevant to the water balance (e.g., PET) can also compensate for poor performance (Figure 12B), though not to the same extent as adopting a PILSTM approach.

These experiments better define the state-of-the-art in predicting streamflow in data-scarce watersheds. First, transfer learning through pre-training a DL model on existing hydrologic databases before fine-tuning with available data from the target watershed is likely to result in substantial performance gains for most watersheds and boost generalization capabilities, as has been recognized previously (Tan et al., 2018). However, we show that for a subset of watersheds, leveraging this geographically broad information will diminish performance. More research is needed to identify characteristics of watersheds that fall into this subset and understand how they differ from the watersheds in the pretraining dataset. Nonetheless, when two or more years of data are available from the target watersheds, informing the DL model with physics and/or basing the model on water balance-relevant time series is likely to compensate for counterproductive transfer learning. Importantly, this finding pushes back against the idea that two or three years of data from a watershed is insufficient for streamflow forecast development and may greatly expand the number of watersheds for which robust, hybrid forecasts are available.

Many data-scarce watersheds for which discharge prediction would be desirable are located outside the geographic extent of the large-sample hydrology database CAMELS, though recent years have produced extensions to Great Britain (Coxon et al., 2020), Brazil (Chagas et al., 2020), Chile (Alvarez-Garreton et al., 2018), Switzerland (Höge et al., 2023) and Australia (Fowler et al., 2021). Without the immediate availability of a database for pretraining, our findings suggest that predictive models for these watersheds would benefit from a PILSTM approach (in terms of multidimensional predictive performance—when training data exceed five years—and stability in nonstationary conditions). Whether those watersheds would still benefit from transfer learning from DL models developed for other geographic areas is worthy of investigation and underscores the need for developing understanding of how geologic and

climatic similarity to catchments in large-sample hydrology databases is related to predictive performance. Because storm characteristics can play as large a role in runoff response behaviors as geologic characteristics (Moges et al., 2022), geographic proxies for climate (i.e., longitude, latitude) in the set of features considered in entity-aware DL may require substitution with a suite of more specific climate metrics that capture storm timing, frequency, intensity, and overall water delivery characteristics to effectively transfer learned behaviors outside the geography of the pretraining database. Our work suggests that, even if dissimilarity between the target watersheds and those in existing databases is predictive of poor performance in an LSTM, integrating physical constraints into the loss function in a PILSTM and/or using direct water balance quantities as input data when two or more years of fine-tuning data are available may yield models with suitable performance. However, these predictions need to be tested more broadly.

Although agencies have favored investment in physical models for operational continental-scale hydrological forecasting in the form of the National Water Model (NOAA 2016) and National Hydrologic Model (Regan et al., 2018), recent studies suggest that performance can be improved by leveraging hybrid (i.e., post-processing and/or mass-conserving) approaches or adopting an LSTM model pretrained on the CAMELS database (Frame et al., 2021; 2022;). Further, water management agencies such as the National Resource Conservation Service are beginning to adopt ensemble water supply forecasts for the western U.S. that leverage diverse types of machine learning and physics-aware artificial intelligence (Fleming et al., 2023). These studies hint that operational adoption of these approaches is within reach. Our findings point to a refinement of this future, in which watersheds are pre-classified according to whether they are likely to benefit from pre-training and whether fine-tuning with a PILSTM or LSTM may safeguard against possible performance deterioration arising from pretraining. Once broader studies have been done to refine our decision-tree approach (Figure 15), development of a modeling workflow to implement the decision tree in a geographically broad forecasting system will be straightforward.

3.7 Conclusion

In this study, we introduced a novel Physics-Informed Machine Learning (PILSTM) model, designed to combine the strength of ML models and the physical model to enhance accuracy under various scenarios. Additionally, we explored the implications of pretraining models on a large hydrological dataset, leveraging available data to unlock the full potential of the data-driven model.

The results indicate consistent superiority of the PILSTM model over LSTM-based models across various watersheds when employing the same model inputs for site-specific scenarios. In the absence of pretraining, the PILSTM model also demonstrates an advantage in data-scarce and

climate-change scenarios. Pretraining the model on a large dataset significantly enhances predictive accuracy, particularly for locations with limited data. In pretraining scenarios, the PILSTM models exhibit a marginal performance advantage compared to the LSTM model in most cases. Machine learning models pretrained on extensive datasets show a reduced need for additional inputs of physical information. However, caution is advised when applying pretrained models to watersheds with hydrological patterns not well represented in the pretraining dataset, as their predictive power may be compromised. Under certain circumstances, integrating physical constraints through a PILSTM may ameliorate counterproductive transfer learning to produce performance gains for those watersheds. The study emphasizes the dual benefits arising from leveraging extensive datasets and incorporating physical constraints in data-driven models. The integration of physically-based models with ML models emerges as a promising strategy, producing predictions that are not only accurate but also interpretable and better aligned with our understanding of water-balance processes.

Chapter 4

Utility of clustering for predictions in ungauged basins in the age of machine learning

4.1 Abstract

This study investigates the potential benefits of employing watershed pre-classification in conjunction with deep neural networks, specifically Long Short-Term Memory (LSTM) models, for streamflow estimation in ungauged basins. Utilizing data from the CAMELS dataset, we compare the performance of a global entity-aware LSTM model trained on meteorological time series data with that of local models trained on pre-clustered watersheds. Our approach involves experimenting with different features for watershed clustering, including site attributes and hydrological signatures. We also evaluate the use of meteorology-discharge transfer entropy statistics for clustering, leveraging directional information flows from precipitation to discharge. Our findings show that pre-clustering with these three types of features does not significantly enhance model performance. Instead, employing these features directly as inputs into LSTM models yields better results than using them to pre-cluster watersheds. In addition, the entity-aware LSTM model utilizing hydrological signatures demonstrates the highest predictive power, suggesting their effectiveness in distinguishing watershed characteristics and identifying hydrological processes. While our study acknowledges its limitations, it underscores the potential for future research to explore the integration of hydrological signatures as static inputs in LSTM models for streamflow prediction in ungauged basins.

4.2 Introduction

A significant portion of the world's rivers, stream reaches, and tributaries lack adequate gauging records, meaning there are insufficient hydrological observations both in terms of quantity and quality. This makes it challenging to do water supply or flood forecasting in many regions or relate streamflow to other environmental parameters of interest. The urgency to reduce uncertainty in predicting flow in ungauged basins is particularly critical for developing countries, where unreliable predictions significantly hamper efficient water resource management and the ability to mitigate the impacts of floods and droughts (Hrachowitz et al., 2013). Prediction in ungauged basins (PUB) is one of the longest-standing grand challenges in hydrology (Sivapalan et al., 2003). To catalyze research progress on PUB, the hydrologic science community named the decade 2003-2013 as the “Decade of PUB,” (Hrachowitz et al., 2013) with funding and research incentives intended to galvanize major progress on how to transfer learning from individual, well-studied watersheds to other watersheds. While important advances were made (Hrachowitz et al., 2013), the Decade of PUB predated the data and machine learning revolution,

which is now rapidly changing the nature of learning and prediction across nearly all disciplines. Here we investigate several strategies for knowledge transfer to improve PUB, comparing and blending long-standing methods in hydrology with discipline-agnostic approaches from data science.

In ungauged basins, the absence of observations makes it impractical to select or calibrate hydrological models using in situ data, as typically done in well-gauged basins. Consequently, watershed regionalization and classification methods have emerged as a solution to facilitate predictions in ungauged basins. Regionalization involves the extrapolation of hydrological information from gauged (observed) basins to ungauged basins within a region (G. Blöschl & Sivapalan, 1995). For example, some studies directly regress model parameters to physiographic attributes to subset variables for regionalization (Fouad et al., 2018; Ye et al., 2014). Meanwhile, classification is primarily employed to group similar watersheds based on their inherent characteristics or attributes and then develop individual hydrological models for each group. Regionalization and classification methods are often used in conjunction to improve hydrological predictions in various contexts. Both methods enable predictions in ungauged basins by leveraging calibrated hydrological models from well-observed watersheds with similarities in hydrological behaviors (Razavi & Coulibaly, 2013).

Utilizing an appropriate watershed classification methodology can enhance predictions in ungauged basins when combined with physical and process-based hydrological models. For instance, Kanishka & Eldho (2020) employed Isomap and principal component analysis (PCA) techniques to classify 30 watersheds in the Godavari river basin, India, identifying hydrologically similar watersheds. Applying these classification results to the Soil and Water Assessment Tool (SWAT), they observed an improved accuracy in streamflow estimation for ungauged basins. Additionally, Razavi & Coulibaly (2016) showed that specific combinations of watershed classification techniques, regionalization methods, and hydrologic models, such as MAC-HBV (McMaster University-Hydrologiska Byråns Vattenbalansavdelning) and SAC-SMA (Sacramento Soil Moisture Accounting Model), significantly enhanced the accuracy of estimated daily mean, low, and peak flows in ungauged basins.

A well-designed classification scheme not only facilitates model prediction but also transfers knowledge about hydrologic processes from gauged to ungauged basins and can enhance understanding of hydrological processes. For example, Wu et al., (2021) investigated regional patterns of dominant streamflow generation mechanisms, utilizing six hydrological signatures to classify 432 catchments into eight classes. Their findings provide detailed insights into the climatic and physiographic controls on regional streamflow patterns, contributing to a transferable understanding of these mechanisms. Jehn et al., (2020) analyzed 643 catchments from the CAMELS dataset (Addor et al., 2017), identifying hydrological clusters based on hydrological behaviors and revealing connections with catchment attributes. They discovered

that, when considering the complete dataset, climate stands out as the primary factor influencing hydrological behavior, and the manifestation of climatic forcing is more pronounced in the behavior of specific catchments than in others.

Meanwhile, the rise of artificial intelligence has propelled machine learning models to the forefront of streamflow prediction, owing to their exceptional predictive capabilities. Instead of pre-classifying watersheds based on site attributes, deep neural networks (DNNs) allow direct input of these attributes, enabling the model to autonomously discern differences between sites. In a study by Kratzert, Klotz, Herrnegger, et al., (2019), long short-term memory (LSTM) models forced with nearly ubiquitously available time series (i.e., temperature, precipitation) were trained with static catchment attributes on CAMELS sites using k-fold validation, treating the left-out fold as ungauged basins. The results demonstrated that the LSTM outperformed both the calibrated SAC-SMA (one of the most widely used physical models of streamflow) and the National Water Model in terms of median Nash-Sutcliffe Efficiencies when tested on 531 basins treated as ungauged. This research highlights the potential of LSTM models in PUB and suggests that commonly available catchment characteristics offer sufficient information for machine learning algorithms to distinguish between catchment-specific rainfall-runoff behaviors.

It remains uncertain whether pre-classifying watersheds can enhance DNN performance by leveraging their capacity to discern similarities and differences based on input time series and attributes. Watershed classification offers one pathway for incorporating expert knowledge about hydrological behaviors into prediction models. Here we hypothesize that certain classification methods may offer supplementary insights into watershed heterogeneity and similarity, beyond what the prediction model discerns, and can therefore enhance model accuracy. This study uses the state-of-the-art DNN models in in terms of streamflow prediction to address whether classification remains useful for PUB. It also examines which classification method can offer more insights into the hydrological behaviors of watersheds, despite the absence of comparative analyses on different watershed classification schemes and their impact on predictive models.

In this study, we experiment with a machine learning model (LSTM) and see whether classification can bring benefits to improve model prediction. We conduct our experiment using data from the CAMELS dataset (Addor et al., 2017; Newman et al., 2015). For PUB with LSTM, we train a global model with 5 meteorological time series data as inputs and compare it with local models trained using clustering results.For watershed clustering, we experimented with different features for clustering to see which one provides more useful information to model prediction. We use the site attributes of watersheds available from the CAMELS dataset and hydrological signatures provided in Wu et al., (2021). We assessed the use of meteorology-discharge transfer entropy statistics for clustering, considering the capacity of directional information flows from precipitation to discharge to offer insight into dominant mechanisms of catchment runoff generation (Moges et al., 2022).

4.3 Data

The data utilized in this study are sourced from the publicly available Catchment Attributes and Meteorology for Large-Sample Studies (CAMELS) dataset (Addor et al., 2017; Newman et al., 2015). Among all the CAMELS sites, we preselected 432 watersheds, based on three criteria established by Wu et al. in 2021. Firs, the catchment area for each selected site falls within the range of larger than 20 km² and smaller than 10,000 km². Second, no more than 30% of the total precipitation in the catchments occurs in the form of snow. Lastly, the Nash-Sutcliffe efficiency of streamflow predictions in the coupled Snow-17 model and the SAC-SMA model is constrained to be no less than 0.5 (Wu et al., 2021).

In this study, we employ data for training that spans January 1, 1997 to December 31, 2014, which aligns with the data range utilized for analysis in Wu et al., 2021. For testing, we utilize data from January 1, 1985, to December 31, 1996 at a daily time increment. The meteorological forcing variables in our analysis encompass precipitation, maximum and minimum air temperature, vapor pressure, and solar radiation. The forcing data, covering the period from January 1, 1985, to December 31, 2014, exhibit no missing values as they are continuous gridded weather data products developed by NASA's Earth Science Division.

After scrutinizing the discharge data, we detected missing values within the designated date ranges. To ensure data quality, we established a criterion stipulating that the number of missing values must not exceed 20% for both the testing and training periods. In accordance with this pre-screening standard, we proceed to retain data from 415 sites for subsequent analysis. Within this dataset, the average proportion of missing values is 1.3% for the training data and 0.3% for the testing data across these 415 sites.

4.4 Methods

4.4.1 Features for clustering

4.4.1.1 Site attributes

In this study, we experimented with three sets of features for watershed clustering. The first approach leverages 27 watershed attributes available from the CAMELS dataset. These 27 attributes encompass a range of details pertaining to topography, climate, vegetation, soil composition, and geological features (further elaborated in Appendix G1). These 27 attributes are used as static inputs in the LSTM model by Kratzert et al., 2019. In our study, we employ these identical 27 attributes for both watershed clustering purposes and for establishing static inputs, facilitating a comparative analysis with the methodology in Kratzert's research (Kratzert et al., 2019).

4.4.1.2 Hydrological signatures

The secondary clustering method employed in this study relies on the utilization of six hydrological signatures as outlined in Wu et al., 2021 (Appendix G2). These hydrological signatures are examined using CAMELS data spanning the period from 1997 to 2014. In addition to the daily precipitation time series, the analysis of hydrological signatures, as detailed in Wu et al., 2021, incorporates an hourly precipitation dataset sourced from the NCEP Stage II and Stage IV multisensor gridded data. It is noteworthy that this clustering method can be interpreted as the one most informed by expertise on physical hydrological processes.

4.4.1.3 Transfer entropy statistics

Our third clustering feature uses an information-theoretic method called transfer entropy (TE). This method quantifies the information transferred (i.e., reduction of uncertainty) from the past values of X to Y, given the historical values of Y (eqn.12; eqn.13).

$$TE(X \to Y)^{k,\tau,h} = \sum p\left(x_{t-\tau}^{k}, y_t, y_{t-h}\right) \log\left[\frac{p(y_t|y_{t-h}, x_{t-\tau}^{k})}{p(y_t|y_{t-h})}\right]; \quad (12)$$

$$x_{t-\tau}^{k} = \frac{1}{k}\sum_{p=0}^{k-1} x_{t-\tau-p} \ . \quad (13)$$

$TE(X \to Y)^{k,\tau,h}$ represents the transfer entropy from variable X to variable Y conditioned on Y with a history length of h. The variable X has a time lag of τ and an aggregation length of k.

We applied this method to examine the information flow from five meteorological time series (precipitation, maximum and minimum air temperatures, solar radiation and vapor pressure) to streamflow time series across varying time spans and delay days. Here we calculate the TE conditioned on the streamflow value of the previous day ($h = 1$). Note that we only use the data in the training period to compute the transfer entropy statistics.

We computed transfer entropy (TE) attributes using different aggregation intervals and lag days, specifically 1, 3, 7, 30, 60, 90, and 180 days. As a result, we generate a total of 245 TE attributes (i.e., combinations of seven aggregation intervals, seven lags and five meteorological time series) for each specific site. To ensure the significance of these TE attributes, we shuffle the meteorological time series 1000 times and calculate random TE statistics. This procedure yields a critical TE threshold at the 95th percentile, and we only carry forward the TE attributes exceeding these thresholds for the clustering analysis. To avoid collinearity, we delete TE features with correlations greater than 0.90. After preprocessing the TE features with above steps, we end up with 39 significant TE features for clustering.

4.4.2 Watershed clustering and local model weights

To mitigate collinearity among the three sets of features outlined earlier prior to watershed clustering, we performed a Principal Component Analysis (PCA). Subsequently, we retained the

PCA components that account for 95% of the variance within the data. The clustering procedure was then executed using these selected PCA components.

In the process of determining the optimal number of clusters, we used the Silhouette Score (Rousseeuw, 1987). When clustering with site attributes, the Silhouette Score indicated that the optimal number of clusters is four. In the case of hydrological signatures, the Silhouette Score exhibited a local maximum at four clusters. For the TE method, we observed a decreasing Silhouette Score as the number of clusters increased from two, meaning the optimal number of clusters is two for the TE method. To fairly compare across three clustering methods, we decided to use four clusters for grouping watersheds.

We applied Gaussian mixture models (GMMs) for clustering, which provide us with the probability of each site belonging to a cluster. Subsequently, we utilized these clustering outcomes as weights for training local LSTM models. By 'local LSTM model,' we refer to the training of a model for each watershed cluster. We experimented with two approaches to map the GMMs outcomes to local model weights. The first approach involves directly using the probability as the model weights. The second approach (results provided in Appendix I) entails using the exponential of the probability as the model weights. Utilizing the exponential of the probability offers the advantage of enabling the local model to utilize and learn from more data points, even when a data point has a probability of 0 to belong to a particular cluster. However, employing the exponential probability mapping method tends to align the behavior of the local model more closely with the global model trained on all the sites.

4.4.3 PUB with LSTM model

To simulate predictions in ungauged basins, we first randomly divided the sites into four folds. The model was trained using the sites in three folds, and subsequently, the model was tested on the fold left out. The sites in this left-out fold were treated as "ungauged basins" in each experiment. This process was repeated four times, enabling us to test the model on all sites, treating them as "ungauged".

To evaluate the effectiveness of clustering results, we trained two types of models. The first, known as the global model, was trained on all sites across the three folds with uniform weights. The second type, leveraging clustering results, is denoted as regional/local models. These models are individually trained for each cluster, incorporating designated weights for every training site in the three folds. The assignment of weights is determined by the clustering results as explained in 4.3.2. Subsequently, these regional models are utilized to make predictions on the testing sites if the testing sites are identified as belonging to the respective cluster (Table 9).

In the global benchmark model, we employ five meteorological time series as data inputs, covering precipitation, solar radiation, minimum and maximum air temperature, and vapor

pressure. To align with the methodology proposed by Kratzert (2019) and facilitate a comparative analysis, we use an additional set of 27 site attributes—matching the features used for clustering—as supplementary static inputs in the LSTM global model. The global model, utilizing static attributes as inputs, is also referred to as an entity-aware LSTM model. This designation stems from its ability to distinguish between different watersheds by leveraging these static inputs, thereby enhancing its capacity to gather additional information. We also try with inputting other clustering features as static inputs in the model besides the site attributes. For each global model with static attributes as inputs, we train a corresponding local model to see whether pre-clustering can help with increasing model performance. We utilized the Mann-Whitney U test (Wilcoxon Rank-Sum test) to assess whether there are significant differences between the medians of the model results.

Table 9. Model Set-ups

No.	Model name	Input features	Using clustering results?	Clustering features	With static inputs?
1	global_benchmark	5 Meteorological variables	No	/	No
2	global_attr	5 Meteorological variables, site attributes	No	/	Yes, use site attribute
3	global_hydro	5 Meteorological variables, hydrological signatures	No	/	Yes, use hydrological signatures
4	global_TE	5 Meteorological variables, TE attributes	No	/	Yes, use TE statistics
5	local_attr	5 Meteorological variables, site attributes	Yes	site attributes	Yes, use site attribute
6	local_hydro	5 Meteorological variables, hydrological signatures	Yes	hydrological signatures	Yes, use hydrological signatures
7	local_TE	5 Meteorological variables, TE attributes	Yes	TE statistics	Yes, use TE statistics

4.4.4 Hyperparameter selection

We fine-tuned hyperparameters in the global benchmark experiment and maintained their consistency across other experiments. Specifically, we tuned the model's hidden size and learning rate and set the values of other hyperparameters according to Kratzert, Klotz, Herrnegger, et al., 2019 (Table 10).

In the hyperparameter tuning process, we divided the sites into four folds. The model underwent training on three folds of data, with validation performed exclusively on the last fold. This process exclusively utilized data from the training period. The results reveal that an optimal configuration is attained with a hidden size of 256 and a piecewise learning rate (Table 10).

Table 10. Hyperparameters in the LSTM model

Hidden size	Learning rate (for epoch)	Input sequence	Dropout	LSTM layers #	training epochs	Batch size
256	0-10: 1e-3 11-20: 5e-4 21-30: 1e-4	270	0.4	1	30	256

4.5 Results

4.5.1 Watershed classification results

An examination of clustering results, comparing outcomes across three distinct clustering methods is performed: clustering based on static attributes, TE (Transfer Entropy) statistics, and hydrological signatures. An analysis of the distribution of sites among clusters revealed consistent patterns across all three methods and a nonuniform distribution of sites across clusters. Notably, the largest cluster encompasses nearly half of the sites, while the smallest cluster

comprises approximately ten sites (Figure 16).

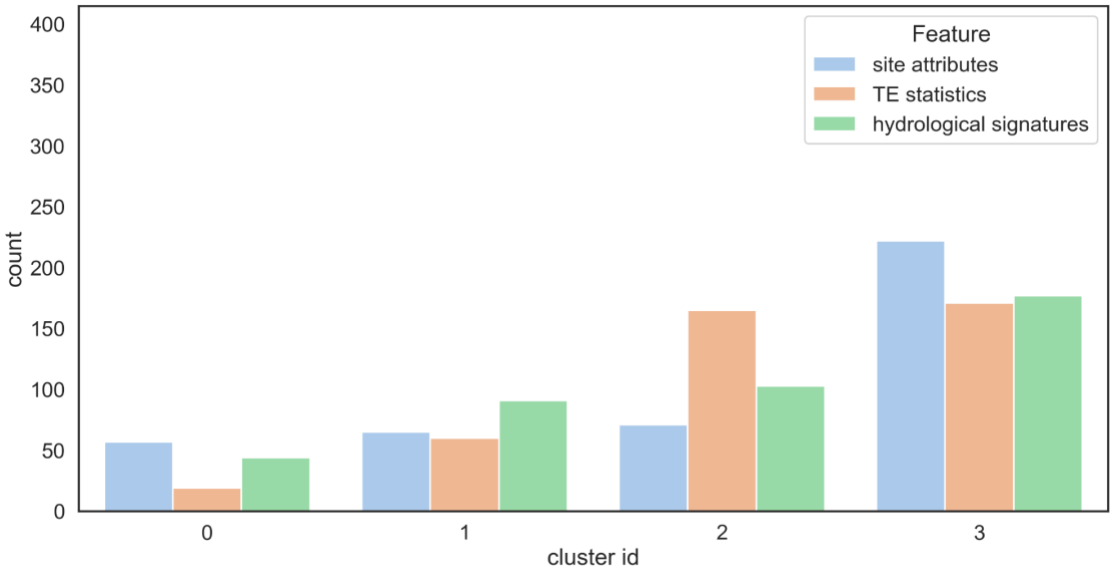
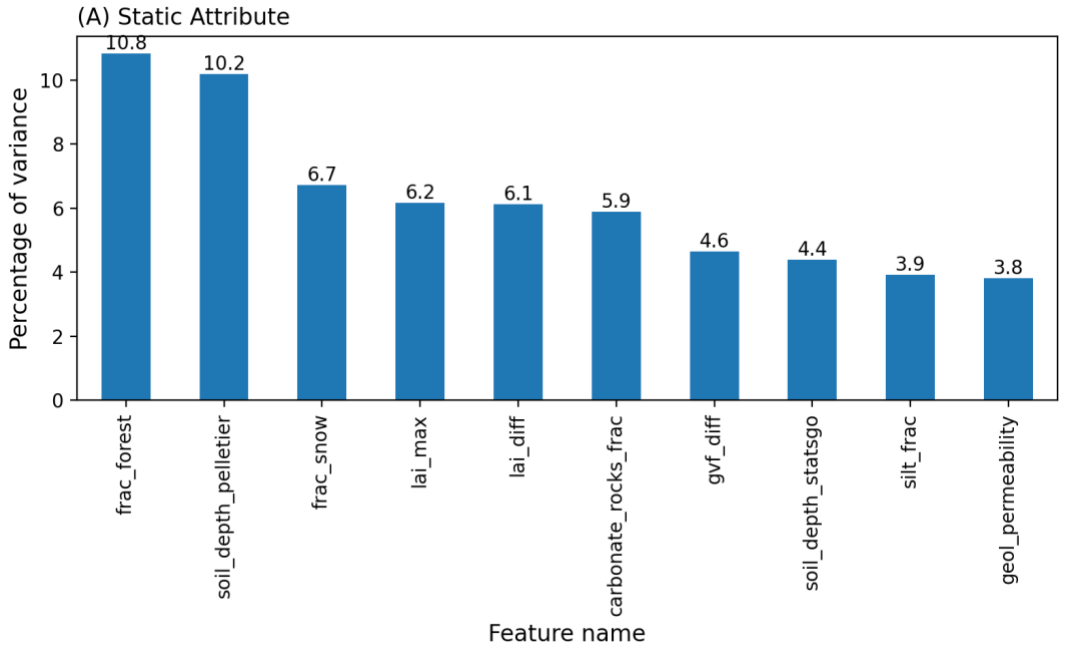
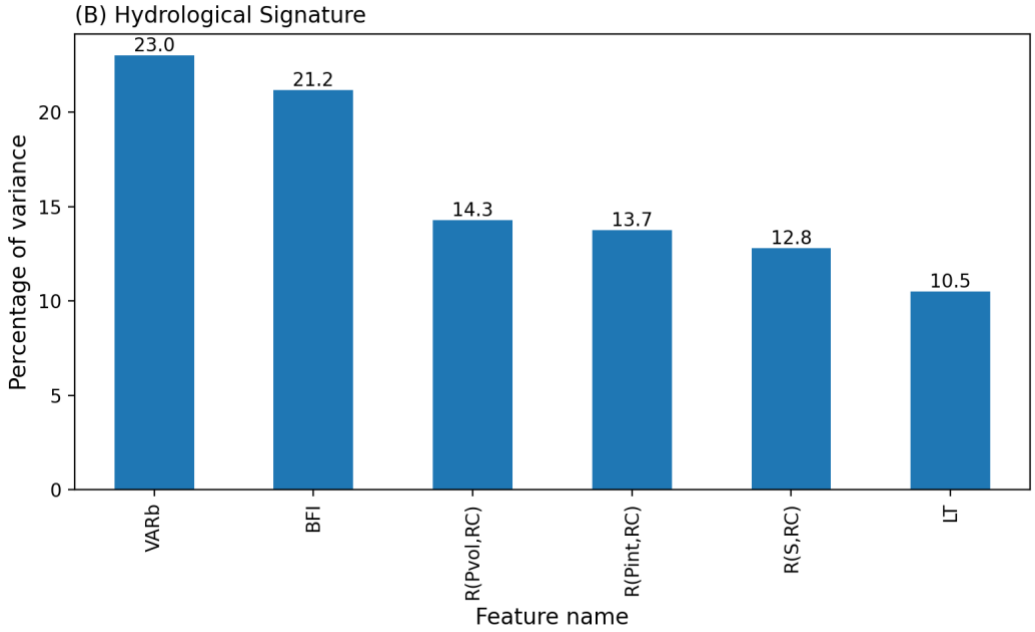


Figure 16. The number of sites in each cluster identified by different types of features.

We assess the variance of principal components by each feature, highlighting the most influential ones (Figure 17) which represent the primary contributors to site clustering. Among static attributes, the top feature shaping site clusters is the fraction of forest (frac_forest), related to land cover and vegetation. Following closely is soil depth to bedrock (soil_depth_pelietier), reflecting soil characteristics. The third significant feature is the fraction of snow, indicative of site climate and precipitation form. Additionally, leaf area index features, pertaining to land cover, are among the top contributors. In terms of hydrological signatures, the foremost features for clustering are the variability of base flow (VARb) and the base flow index (BFI), crucial for understanding streamflow partitioning and groundwater dynamics. In TE statistics, the top three features are all associated with precipitation, with a lag day of 1 and aggregation lengths of 1, 3, and 7 respectively. This indicates that, in comparison to interactions involving other meteorological variables and discharge, the relationships between precipitation and discharge are the most significant factors in distinguishing among watersheds during clustering.


(A) Static Attribute
Percentage of variance
Feature name
frac_forest
soil_depth_pelletier
frac_snow
lai_max
lai_diff
carbonate_rocks_frac
gvf_diff
soil_depth_statsgo
silt_frac
geol_permeability
10.8
10.2
6.7
6.2
6.1
5.9
4.6
4.4
3.9
3.8


(B) Hydrological Signature
Percentage of variance
Feature name
VARb
BFI
R(Pvol,RC)
R(Pint,RC)
R(S,RC)
LT
23.0
21.2
14.3
13.7
12.8
10.5

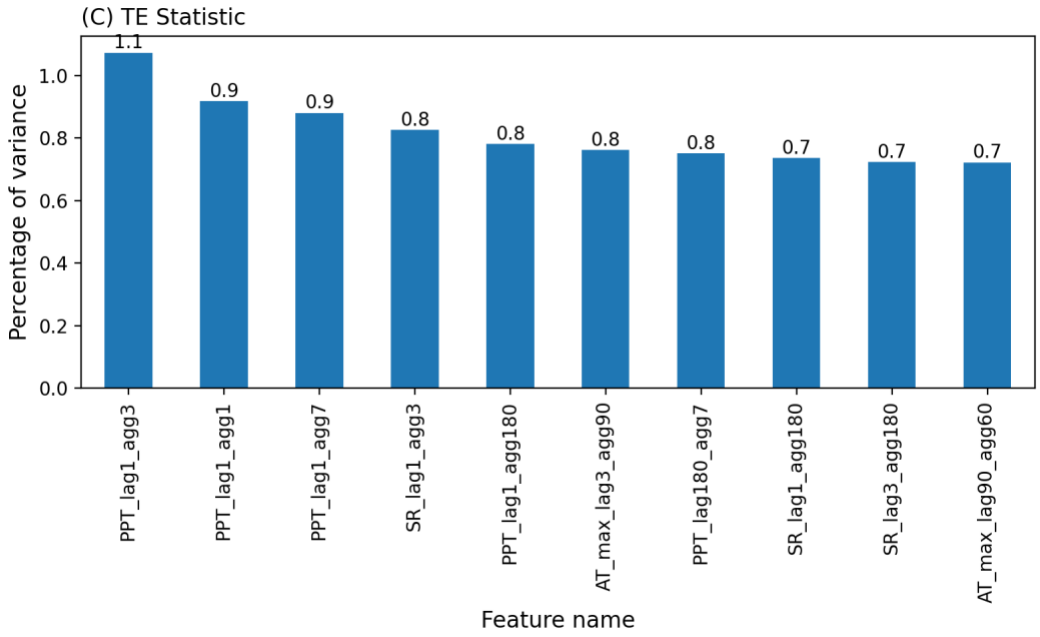
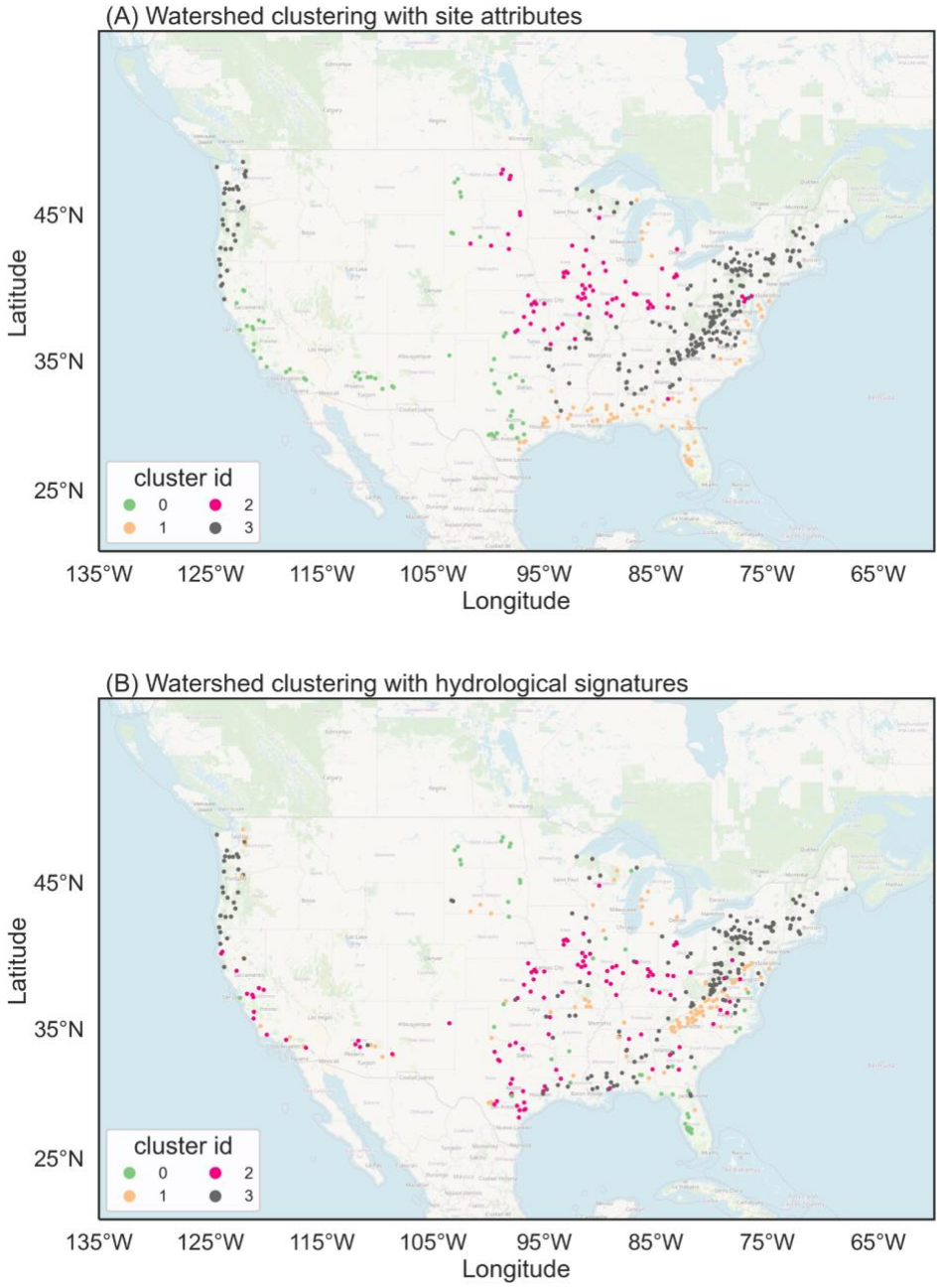


Figure 17. Breakdown of the variance by each feature. (A) Percentage of variance explained by each static attribute. (B) Percentage of variance explained by each hydrological signature. (C) Percentage of variance explained by each TE statistic.

When examining clusters identified based on watershed static attributes, a notable observation arises regarding their geographical distribution (Figure 18). Sites within the same cluster demonstrate a pronounced spatial coherence, with a tendency towards concentrated geographical proximity. The largest cluster encompasses most sites along the Appalachian Mountains and the northwest region, characterized by a predominant land cover of deciduous and evergreen forest. Meanwhile, the second largest cluster comprises sites in the central part of the US. Sites along the southern coastline form two distinct clusters, distinguished by their location on either the east or west coast, which suggests variations in land cover and soil characteristics between the two coastal regions. The clustering results based on hydrological signatures reveal less spatial cohesion, primarily influenced by proxies of precipitation partitioning. Clusters identified using

TE statistics exhibit an even more dispersed geographical arrangement.



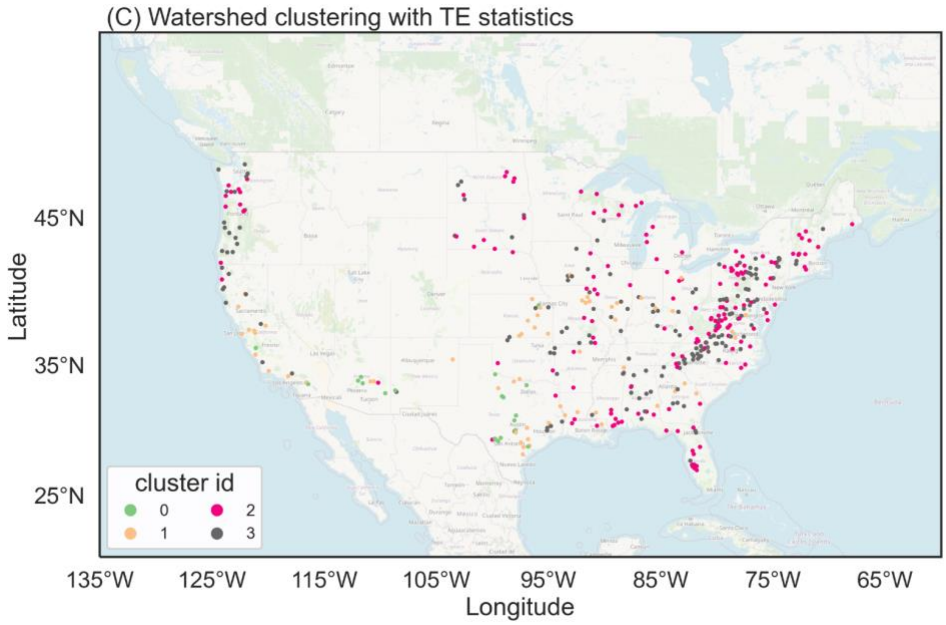


Figure 18. Geographical distribution of sites in different clusters identified by (A) site attributes (B) TE statistics (C) hydrological signatures.

4.5.2 LSTM model prediction results

Each of the global LSTM models outperforms or has the same performance as the corresponding set of local models using the same inputs (Table 11). Additionally, all the global models with different sets of static inputs outperform the global benchmark model. When utilizing clustering results, only the local model that incorporates hydrological signature information outperforms the benchmark model. That local model matches the performance of the global model that uses hydrological signatures as direct static model inputs.

Most of the sites achieved their best performance in models using hydrological signature information. Overall, the global model incorporating hydrological signatures as static inputs emerges as the most effective.

Table 11. Summary table of model performance. The model with boldface indicates the best performance overall. The Mann-Whitney U test (Wilcoxon Rank-Sum test) was utilized to assess significant differences in model performance. If the two models do not exhibit significantly different median values, the models will be denoted with the same superscript.

Model name	NSE median	NSE mean	Number of sites with NSE values<0	Number of sites where this model performs optimally
global_benchmark	0.58^{a}	0.48	25	19

global_attr	0.67^{b}	0.51	18	78
global_hydro	0.69^{c}	0.59	13	110
global_TE	0.64^{b}	0.49	15	66
local_attr	0.48	-24.33	123	35
local_hydro	0.68^{bc}	0.53	20	76
local_TE	0.57^{a}	-3.20	39	32

4.6 Discussion

This study aimed to evaluate whether pre-clustering watersheds can provide helpful information for LSTM model prediction in ungauged basins. We found that pre-clustering was generally not helpful; models utilizing clustering could achieve only the same or lower overall performance than counterpart global models that directly incorporate these clustering features as static inputs. However, among the different features used for clustering, hydrological signatures are the most effective in extracting information for use in the LSTM model.

Global Benchmark vs. Global Entity-Aware Model

Comparing the global benchmark with the other three global entity-aware models reveals that regardless of the features used as input for the LSTM model (static attributes, hydrological signatures, TE statistics), the global entity-aware models consistently outperform the benchmark (Table 11). This finding corroborates previous studies on entity-aware LSTM models, showcasing the advantage of integrating static attributes into the LSTM model (Kratzert et al., 2019a; Kratzert et al., 2019b). The results demonstrate that the entity-aware model can extract valuable information from these features and employ it effectively for prediction, indicating that all three types of features contain pertinent information about watersheds and their rainfall-runoff processes.

Global Entity-Aware Model vs. Local Entity-Aware Model

Despite our intention for the local model to cater to specific types of watersheds with similar hydrological behaviors, it does not exhibit comparable performance to the global entity-aware models (Table 11). Notably, both global and local entity-aware models utilize the same set of input features, differing only in the assignment of weights to watersheds. Analysis of weights in local models reveals that for most watersheds for which membership in the cluster of interest is not most probable, the weights are nearly or entirely zero, indicating that they do not contribute to learning. Consequently, each local model is trained on fewer watersheds overall compared to the global model. These results suggest that pre-clustering watersheds and restricting the model to digest information from specific watersheds are not as effective as training the model on a broader range of watersheds, allowing it to discern differences among them.

We find that the local model utilizing hydrological signatures for clustering performs nearly the same as the global model (Table 11). This finding underscores the effectiveness of hydrological signatures in capturing similarities in watershed behaviors, indicating that hydrological behaviors are valuable information for model prediction. Previous classification studies have noted that different sets of catchment attributes and climate can lead to very similar hydrological behaviors—a phenomenon known as equifinality in catchment response (Jehn et al., 2020). This study offers insight into why static attributes are less effective than hydrological signatures in watershed clustering, as clustering results based on static attributes may not necessarily reflect similarities in hydrological behaviors.

Although hydrological signatures demonstrate superior performance and provide the most useful additional information in the global entity-aware model, static attribute feature values are the most readily obtainable for ungauged basins. Static attributes encompass details about watershed topography, soil, climate, land cover, and geology but lack explicit information reflecting the interaction between climate and discharge time series. TE statistics reflect compound interactions between meteorological and discharge time series in the form of direct information flows, requiring prior observations of precipitation and discharge but computable with a few years of data. Hydrological signature classification provides information based on prior hydrological knowledge and physical rules, indicating the relationship between base flow and total discharge, stormflow behavior, and offering insights into the catchment's hydrograph. However, hydrological signatures necessitate extended observations and analyses to compile. To leverage hydrological signatures for PUB, forthcoming work should be focusing on estimating these metrics for unmonitored basins.

4.7 Conclusion

The purpose of this study was to determine whether using deep neural networks (LSTM) to estimate streamflow in ungauged basins can benefit from watershed pre-classification in any way. Three distinct feature types, each representing a different viewpoint on the watershed characteristics, have been tested for watershed clustering. Additionally, we employed Gaussian mixture models (GMMs) as our clustering technique. We preprocessed the 415 watersheds from the CAMELS dataset into five clusters in order to exploit the results of watershed pre-clustering. Next, we trained a local LSTM for each cluster. By contrast, we employed these attributes directly as static inputs to the LSTM model, bypassing the pre-clustering of the watersheds, and used the entity-aware model to estimate streamflow in ungauged basins.

The outcomes demonstrated that pre-clustering with these three attributes was ineffective in enhancing model performance. Furthermore, it turns out that using these watershed static data directly as inputs into LSTM models works better than pre-clustering. The hydrological signature-using entity-aware long short-term memory (LSTM) model has the highest prediction power of all of them, suggesting that hydrological signatures are the most effective way to distinguish among various watersheds and identify their hydrological processes.

It's important to note that the methods and features explored in this study are limited. Therefore, the results do not dismiss the possibility of pre-clustering watersheds to enhance prediction in PUB. Additionally, as the hydrological signatures are derived from observed time series, if we aim to utilize them as static inputs in the LSTM model for PUB, future work will involve estimating these features for ungauged basins.

Chapter 5

Conclusion

In conclusion, this dissertation contributes insights and methodologies to the field of hydrology through the exploration of large-scale hydrological datasets and the development of innovative modeling techniques.

The creation of the CHOSEN database provides researchers with access to comprehensive hydrological observatory data essential for doing comparative analysis. This database fulfills the critical data needs for comparative hydrology and also provides an example of cleaning and processing hydrological data. With the CHOSEN dataset lays the groundwork for studies aimed at establishing hydrologic baselines, analyzing information on wetting and drying trends, and attributing observed changes to underlying hydrological processes.

In parallel, the introduction of the Physics-Informed Machine Learning (PILSTM) model presents a novel approach to combining machine learning and physical models, resulting in enhanced accuracy under various scenarios. The study highlights the consistent superiority of the PILSTM model over traditional LSTM-based models, particularly in data-scarce and climate-change scenarios. Furthermore, the study emphasizes the importance of leveraging extensive datasets and incorporating physical constraints in data-driven models to produce accurate and interpretable predictions aligned with our understanding of water-balance processes.

Lastly, the investigation into the use of deep neural networks for streamflow estimation in ungauged basins reveals valuable insights into the effectiveness of watershed pre-classification. While pre-clustering with distinct watershed attributes was found ineffective in enhancing model performance, the study demonstrates that utilizing hydrological signatures directly as inputs yields the highest prediction power. This underscores the significance of hydrological signatures in distinguishing watershed characteristics and identifying hydrological processes.

In conclusion, the findings from these chapters underscore the importance of comprehensive datasets, innovative modeling techniques, and interdisciplinary approaches in advancing our understanding of hydrological processes and improving predictive capabilities. Moving forward, future research should focus on expanding datasets, refining models, and exploring novel methodologies to address the ongoing challenges in hydrology, thereby enhancing our ability to predict and manage water resources effectively.

Bibliography

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*(10), 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

Addor, N., Do, H. X., Alvarez-Garreton, C., Coxon, G., Fowler, K., & Mendoza, P. A. (2020). Large-sample hydrology: recent progress, guidelines for new datasets and grand challenges. *Hydrological Sciences Journal*, *65*(5), 712–725. https://doi.org/10.1080/02626667.2019.1683182

AghaKouchak, A., Chiang, F., Huning, L. S., Love, C. A., Mallakpour, I., Mazdiyasni, O., et al. (2020). Climate Extremes and Compound Hazards in a Warming World. *Annual Review of Earth and Planetary Sciences*, *48*(1), 519–548. https://doi.org/10.1146/annurev-earth-071719-055228

Ahn, K.-H., & Palmer, R. N. (2016). Trend and Variability in Observed Hydrological Extremes in the United States. *Journal of Hydrologic Engineering*, *21*(2), 04015061. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001286

Ainsworth, T. D., Hurd, C. L., Gates, R. D., & Boyd, P. W. (2020). How do we overcome abrupt degradation of marine ecosystems and meet the challenge of heat waves and climate extremes? *Global Change Biology*, *26*(2), 343–354. https://doi.org/10.1111/gcb.14901

Archfield, S. A., Hirsch, R. M., Viglione, A., & Blöschl, G. (2016). Fragmented patterns of flood change across the United States. *Geophysical Research Letters*, *43*(19), 10,232-10,239. https://doi.org/10.1002/2016GL070590

Ardabili, S., Mosavi, A., Dehghani, M., & Várkonyi-Kóczy, A. R. (2020). Deep Learning and Machine Learning in Hydrological Processes Climate Change and Earth Systems a Systematic Review. In A. R. Várkonyi-Kóczy (Ed.), *Engineering for Sustainable Future* (pp. 52–62). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-36841-8_5

Batibeniz, F., Ashfaq, M., Diffenbaugh, N. S., Key, K., Evans, K. J., Turuncoglu, U. U., & Önol, B. (2020). Doubling of U.S. Population Exposure to Climate Extremes by 2050. *Earth's Future*, *8*(4), e2019EF001421. https://doi.org/10.1029/2019EF001421

Berghuijs, W. R., Woods, R. A., Hutton, C. J., & Sivapalan, M. (2016). Dominant flood generating mechanisms across the United States. *Geophysical Research Letters*, *43*(9), 4382–4390. https://doi.org/10.1002/2016GL068070

Blöschl, G., & Sivapalan, M. (1995). Scale issues in hydrological modelling: A review. *Hydrological Processes*, *9*(3–4), 251–290. https://doi.org/10.1002/hyp.3360090305

Blöschl, Günter, Hall, J., Parajka, J., Perdigão, R. A. P., Merz, B., Arheimer, B., et al. (2017). Changing climate shifts timing of European floods. *Science*, *357*(6351), 588–590. https://doi.org/10.1126/science.aan2506

Byrne, M. P., & O'Gorman, P. A. (2015). The Response of Precipitation Minus Evapotranspiration to Climate Warming: Why the "Wet-Get-Wetter, Dry-Get-Drier" Scaling Does Not Hold over Land. *Journal of Climate*, *28*(20), 8078–8092. https://doi.org/10.1175/JCLI-D-15-0369.1

Byun, K., Chiu, C.-M., & Hamlet, A. F. (2019). Effects of 21st century climate change on seasonal flow regimes and hydrologic extremes over the Midwest and Great Lakes region of the

US. *Science of The Total Environment*, *650*, 1261–1277. https://doi.org/10.1016/j.scitotenv.2018.09.063
Cook, B. I., Ault, T. R., & Smerdon, J. E. (2015). Unprecedented 21st century drought risk in the American Southwest and Central Plains. *Science Advances*, *1*(1), e1400082. https://doi.org/10.1126/sciadv.1400082
Diffenbaugh, N. S., Swain, D. L., & Touma, D. (2015). Anthropogenic warming has increased drought risk in California. *Proceedings of the National Academy of Sciences*, *112*(13), 3931–3936. https://doi.org/10.1073/pnas.1422385112
Do, H. X., Mei, Y., & Gronewold, A. D. (2020). To What Extent Are Changes in Flood Magnitude Related to Changes in Precipitation Extremes? *Geophysical Research Letters*, *47*(18), e2020GL088684. https://doi.org/10.1029/2020GL088684
Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H. V., et al. (2006). Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. *Journal of Hydrology*, *320*(1), 3–17. https://doi.org/10.1016/j.jhydrol.2005.07.031
Feng, H., & Zhang, M. (2015). Global land moisture trends: drier in dry and wetter in wet over land. *Scientific Reports*, *5*(1), 18018. https://doi.org/10.1038/srep18018
Fer, I., Gardella, A. K., Shiklomanov, A. N., Campbell, E. E., Cowdery, E. M., Kauwe, M. G. D., et al. (2021). Beyond ecosystem modeling: A roadmap to community cyberinfrastructure for ecological data-model integration. *Global Change Biology*, *27*(1), 13–26. https://doi.org/10.1111/gcb.15409
Fleming, S. W., & Gupta, H. V. (2020). The physics of river prediction. *Physics Today*, *73*(7), 46–52. https://doi.org/10.1063/PT.3.4523
Fleming, S. W., Watson, J. R., Ellenson, A., Cannon, A. J., & Vesselinov, V. C. (2021). Machine learning in Earth and environmental science requires education and research policy reforms. *Nature Geoscience*, *14*(12), 878–880. https://doi.org/10.1038/s41561-021-00865-3
Ford, T. W., & Quiring, S. M. (2019). Comparison of Contemporary In Situ, Model, and Satellite Remote Sensing Soil Moisture With a Focus on Drought Monitoring. *Water Resources Research*, *55*(2), 1565–1582. https://doi.org/10.1029/2018WR024039
Fouad, G., Skupin, A., & Tague, C. L. (2018). Regional regression models of percentile flows for the contiguous United States: Expert versus data-driven independent variable selection. *Journal of Hydrology: Regional Studies*, *17*, 64–82. https://doi.org/10.1016/j.ejrh.2018.04.002
Frame, J. M., Kratzert, F., Raney II, A., Rahman, M., Salas, F. R., & Nearing, G. S. (2021). Post-Processing the National Water Model with Long Short-Term Memory Networks for Streamflow Predictions and Model Diagnostics. *JAWRA Journal of the American Water Resources Association*, *57*(6), 885–905. https://doi.org/10.1111/1752-1688.12964
Gao, Y., Lu, J., Leung, L. R., Yang, Q., Hagos, S., & Qian, Y. (2015). Dynamical and thermodynamical modulations on future changes of landfalling atmospheric rivers over western North America. *Geophysical Research Letters*, *42*(17), 7179–7186. https://doi.org/10.1002/2015GL065435

Gleason, K. L., Lawrimore, J. H., Levinson, D. H., Karl, T. R., & Karoly, D. J. (2008). A Revised U.S. Climate Extremes Index. *Journal of Climate*, *21*(10), 2124–2137. https://doi.org/10.1175/2007JCLI1883.1
Gudmundsson, L., Leonard, M., Do, H. X., Westra, S., & Seneviratne, S. I. (2019). Observed Trends in Global Indicators of Mean and Extreme Streamflow. *Geophysical Research Letters*, *46*(2), 756–766. https://doi.org/10.1029/2018GL079725
Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., & Andréassian, V. (2014). Large-sample hydrology: a need to balance depth with breadth. *Hydrology and Earth System Sciences*, *18*(2), 463–477. https://doi.org/10.5194/hess-18-463-2014
Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2
Hayhoe, K., Wake, C. P., Huntington, T. G., Luo, L., Schwartz, M. D., Sheffield, J., et al. (2007). Past and future changes in climate and hydrological indicators in the US Northeast. *Climate Dynamics*, *28*(4), 381–407. https://doi.org/10.1007/s00382-006-0187-8
Heidari, H., Arabi, M., Warziniack, T., & Kao, S.-C. (2020). Assessing Shifts in Regional Hydroclimatic Conditions of U.S. River Basins in Response to Climate Change over the 21st Century. *Earth's Future*, *8*(10), e2020EF001657. https://doi.org/10.1029/2020EF001657
Held, I. M., & Soden, B. J. (2006). Robust Responses of the Hydrological Cycle to Global Warming. *Journal of Climate*, *19*(21), 5686–5699. https://doi.org/10.1175/JCLI3990.1
Hirsch, R. M., & Archfield, S. A. (2015). Not higher but more often. *Nature Climate Change*, *5*(3), 198–199. https://doi.org/10.1038/nclimate2551
Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., et al. (2013). A decade of Predictions in Ungauged Basins (PUB)—a review. *Hydrological Sciences Journal*, *58*(6), 1198–1255. https://doi.org/10.1080/02626667.2013.803183
Hu, Z.-Y., Chen, X., Chen, D., Li, J., Wang, S., Zhou, Q., et al. (2018). "Dry gets drier, wet gets wetter": a case study over the arid regions of Central Asia. *International Journal of Climatology*, *39*. https://doi.org/10.1002/joc.5863
Hughes, T. P., Kerry, J. T., Connolly, S. R., Baird, A. H., Eakin, C. M., Heron, S. F., et al. (2019). Ecological memory modifies the cumulative impact of recurrent climate extremes. *Nature Climate Change*, *9*(1), 40–43. https://doi.org/10.1038/s41558-018-0351-2
Ivancic, T., & Shaw, S. (2015). Examining why trends in very heavy precipitation should not be mistaken for trends in very high river discharge. *Climatic Change*, *133*(4), 681–693.
Jehn, F. U., Bestian, K., Breuer, L., Kraft, P., & Houska, T. (2020). Using hydrological and climatic catchment clusters to explore drivers of catchment behavior. *Hydrology and Earth System Sciences*, *24*(3), 1081–1100. https://doi.org/10.5194/hess-24-1081-2020
Jia, X., Willard, J., Karpatne, A., Read, J., Zwart, J., Steinbach, M., & Kumar, V. (2019). Physics Guided RNNs for Modeling Dynamical Systems: A Case Study in Simulating Lake Temperature Profiles. *arXiv:1810.13075 [Physics]*. Retrieved from http://arxiv.org/abs/1810.13075

Kakalia, Z., Varadharajan, C., Alper, E., Brodie, E. L., Burrus, M., Carroll, R. W. H., et al. (2021). The Colorado East River Community Observatory Data Collection. *Hydrological Processes*, *35*(6), e14243. https://doi.org/10.1002/hyp.14243
Kam, J., & Sheffield, J. (2016). Changes in the low flow regime over the eastern United States (1962–2011): variability, trends, and attributions. *Climatic Change*, *135*(3), 639–653.
Kanishka, G., & Eldho, T. I. (2020). Streamflow estimation in ungauged basins using watershed classification and regionalization techniques. *Journal of Earth System Science*, *129*(1), 186. https://doi.org/10.1007/s12040-020-01451-8
Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., et al. (2017). Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10), 2318–2331. https://doi.org/10.1109/TKDE.2017.2720168
Kendall, M. G. (1975). *Rank correlation methods*. London: Griffin.
Knutson, T. R., & Manabe, S. (1995). Time-Mean Response over the Tropical Pacific to Increased C02 in a Coupled Ocean-Atmosphere Model. *Journal of Climate*, *8*(9), 2181–2199. https://doi.org/10.1175/1520-0442(1995)008<2181:TMROTT>2.0.CO;2
Konapala, G., Kao, S.-C., Painter, S. L., & Lu, D. (2020). Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US. *Environmental Research Letters*, *15*(10), 104022. https://doi.org/10.1088/1748-9326/aba927
Kormos, P. R., Luce, C. H., Wenger, S. J., & Berghuijs, W. R. (2016). Trends and sensitivities of low streamflow extremes to discharge timing and magnitude in Pacific Northwest mountain streams. *Water Resources Research*, *52*(7), 4990–5007. https://doi.org/10.1002/2015WR018125
Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. *Water Resources Research*, *55*(12), 11344–11354. https://doi.org/10.1029/2019WR026065
Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12), 5089–5110. https://doi.org/10.5194/hess-23-5089-2019
Lange, H., & Sippel, S. (2020). Machine Learning Applications in Hydrology. In D. F. Levia, D. E. Carlyle-Moses, S. Iida, B. Michalzik, K. Nanko, & A. Tischer (Eds.), *Forest-Water Interactions* (pp. 233–257). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-26086-6_10
Liu, H., van Oosterom, P., Hu, C., & Wang, W. (2016). Managing Large Multidimensional Array Hydrologic Datasets: A Case Study Comparing NetCDF and SciDB. *Procedia Engineering*, *154*, 207–214. https://doi.org/10.1016/j.proeng.2016.07.449
Lu, D., Konapala, G., Painter, S. L., Kao, S.-C., & Gangrade, S. (2021). Streamflow Simulation in Data-Scarce Basins Using Bayesian and Physics-Informed Machine Learning Models. *Journal of Hydrometeorology*, *22*(6), 1421–1438. https://doi.org/10.1175/JHM-D-20-0082.1

Mallakpour, I., & Villarini, G. (2015). The changing nature of flooding across the central United States. *Nature Climate Change*, *5*(3), 250–254. https://doi.org/10.1038/nclimate2516
Mann, H. B. (1945). Nonparametric Tests Against Trend. *Econometrica*, *13*(3), 245–259. https://doi.org/10.2307/1907187
Marsooli, R., Lin, N., Emanuel, K., & Feng, K. (2019). Climate change exacerbates hurricane flood hazards along US Atlantic and Gulf Coasts in spatially varying patterns. *Nature Communications*, *10*(1), 3785. https://doi.org/10.1038/s41467-019-11755-z
McCabe, G. J., & Wolock, D. M. (2009). Recent Declines in Western U.S. Snowpack in the Context of Twentieth-Century Climate Variability. *Earth Interactions*, *13*(12), 1–15. https://doi.org/10.1175/2009EI283.1
McClymont, K., Morrison, D., Beevers, L., & Carmen, E. (2020). Flood resilience: a systematic review. *Journal of Environmental Planning and Management*, *63*(7), 1151–1176. https://doi.org/10.1080/09640568.2019.1641474
McNamara, J. (2017). Long-Term, Continuous Stream Discharge Time Series from Measurement Sites in Dry Creek Experimental Watershed, Southwest Idaho. *Dry Creek Experimental Watershed Data*. https://doi.org/10.18122/B2VG6G
Miller, N., Bashford, K., & Strem, E. (2003). Potential Impact of Climate Change on California Hydrology. *JAWRA Journal of the American Water Resources Association*, *39*, 771–784. https://doi.org/10.1111/j.1752-1688.2003.tb04404.x
Moges, E., Ruddell, B. L., Zhang, L., Driscoll, J. M., & Larsen, L. G. (2022). Strength and Memory of Precipitation's Control Over Streamflow Across the Conterminous United States. *Water Resources Research*, *58*(3), e2021WR030186. https://doi.org/10.1029/2021WR030186
Murdoch, W. J., Liu, P. J., & Yu, B. (2018, April 27). Beyond Word Importance: Contextual Decomposition to Extract Interactions from LSTMs. arXiv. https://doi.org/10.48550/arXiv.1801.05453
Naz, B. S., Kao, S.-C., Ashfaq, M., Rastogi, D., Mei, R., & Bowling, L. C. (2016). Regional hydrologic response to climate change in the conterminous United States using high-resolution hydroclimate simulations. *Global and Planetary Change*, *143*, 100–117. https://doi.org/10.1016/j.gloplacha.2016.06.003
Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., et al. (2015). Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, *19*(1), 209–223. https://doi.org/10.5194/hess-19-209-2015
O., S., & Orth, R. (2021). Global soil moisture data derived through machine learning trained with in-situ measurements. *Scientific Data*, *8*(1), 170. https://doi.org/10.1038/s41597-021-00964-1
Pagán, B. R., Ashfaq, M., Rastogi, D., Kendall, D. R., Kao, S.-C., Naz, B. S., et al. (2016). Extreme hydrological changes in the southwestern US drive reductions in water supply to

Southern California by mid century. *Environmental Research Letters*, *11*(9), 094026. https://doi.org/10.1088/1748-9326/11/9/094026
Pall, P., Aina, T., Stone, D. A., Stott, P. A., Nozawa, T., Hilberts, A. G. J., et al. (2011). Anthropogenic greenhouse gas contribution to flood risk in England and Wales in autumn 2000. *Nature*, *470*(7334), 382–385. https://doi.org/10.1038/nature09762
Pappas, C., Papalexiou, S. M., & Koutsoyiannis, D. (2014). A quick gap filling of missing hydrometeorological data. *Journal of Geophysical Research: Atmospheres*, *119*(15), 9290–9300. https://doi.org/10.1002/2014JD021633
Payne, A. E., Demory, M.-E., Leung, L. R., Ramos, A. M., Shields, C. A., Rutz, J. J., et al. (2020). Responses and impacts of atmospheric rivers to climate change. *Nature Reviews Earth & Environment*, *1*(3), 143–157. https://doi.org/10.1038/s43017-020-0030-5
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (n.d.). Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*, 6.
Petersky, R., & Harpold, A. (2018). Now you see it, now you don't: a case study of ephemeral snowpacks and soil moisture response in the Great Basin, USA. *Hydrology and Earth System Sciences*, *22*(9), 4891–4906. https://doi.org/10.5194/hess-22-4891-2018
Petersky, R. S., & Harpold, A. A. (2018). A Long-Term Micrometeorological and Hydrological Dataset Across an Elevation Gradient in Sagehen Creek, Sierra Nevada, California [Data set]. University of Nevada Reno. https://doi.org/10.5281/zenodo.2590799
Razavi, T., & Coulibaly, P. (2013). Streamflow Prediction in Ungauged Basins: Review of Regionalization Methods. *Journal of Hydrologic Engineering*, *18*(8), 958–975. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000690
Razavi, T., & Coulibaly, P. (2016). Improving streamflow estimation in ungauged basins using a multi-modelling approach. *Hydrological Sciences Journal*, *61*(15), 2668–2679. https://doi.org/10.1080/02626667.2016.1154558
Rieger, L., Singh, C., Murdoch, W. J., & Yu, B. (2020, October 8). Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. arXiv. https://doi.org/10.48550/arXiv.1909.13584
Roth, N., Jaramillo, F., Wang-Erlandsson, L., Zamora, D., Palomino-Ángel, S., & Cousins, S. A. O. (2021). A call for consistency with the terms 'wetter' and 'drier' in climate change studies. *Environmental Evidence*, *10*(1), 8. https://doi.org/10.1186/s13750-021-00224-0
Servilla, M., & Brunt, J. (2011). The LTER Network Information System: Improving Data Quality and Synthesis through Community Collaboration, *2011*, IN51C-1598. Presented at the AGU Fall Meeting Abstracts.
Sharma, A., Wasko, C., & Lettenmaier, D. P. (2018). If Precipitation Extremes Are Increasing, Why Aren't Floods? *Water Resources Research*, *54*(11), 8545–8551. https://doi.org/10.1029/2018WR023749
Signell, R. P., Carniel, S., Chiggiato, J., Janekovic, I., Pullen, J., & Sherwood, C. R. (2008). Collaboration tools and techniques for large model datasets. *Journal of Marine Systems*, *69*(1), 154–161. https://doi.org/10.1016/j.jmarsys.2007.02.013

Sillmann, J., Kharin, V. V., Zwiers, F. W., Zhang, X., & Bronaugh, D. (2013). Climate extremes indices in the CMIP5 multimodel ensemble: Part 2. Future climate projections. *Journal of Geophysical Research: Atmospheres*, *118*(6), 2473–2493. https://doi.org/10.1002/jgrd.50188
Singh, C., Murdoch, W. J., & Yu, B. (2019, January 16). Hierarchical interpretations for neural network predictions. arXiv. https://doi.org/10.48550/arXiv.1806.05337
Singh, C., Ha, W., Lanusse, F., Boehm, V., Liu, J., & Yu, B. (2021, June 14). Transformation Importance with Applications to Cosmology. arXiv. https://doi.org/10.48550/arXiv.2003.01926
Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., et al. (2003). IAHS Decade on Predictions in Ungauged Basins (PUB), 2003-2012: Shaping an exciting future for the hydrological sciences. *Hydrological Sciences Journal*, *48*(6), 857–880. https://doi.org/10.1623/hysj.48.6.857.51421
Slater, L. J., & Villarini, G. (2016). Recent trends in U.S. flood risk. *Geophysical Research Letters*, *43*(24), 12,428-12,436. https://doi.org/10.1002/2016GL071199
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*(56), 1929–1958.
Storch, H. von, & Navarra, A. (Eds.). (1999). *Analysis of Climate Variability: Applications of Statistical Techniques Proceedings of an Autumn School Organized by the Commission of the European Community on Elba from October 30 to November 6, 1993* (2nd ed.). Berlin Heidelberg: Springer-Verlag. https://doi.org/10.1007/978-3-662-03744-7
Swain, D. L., Langenbrunner, B., Neelin, J. D., & Hall, A. (2018). Increasing precipitation volatility in twenty-first-century California. *Nature Climate Change*, *8*(5), 427–433. https://doi.org/10.1038/s41558-018-0140-y
Wasko, C., & Nathan, R. (2019). Influence of changes in rainfall and soil moisture on trends in flooding. *Journal of Hydrology*, *575*, 432–441. https://doi.org/10.1016/j.jhydrol.2019.05.054
Wentz, F. J., Ricciardulli, L., Hilburn, K., & Mears, C. (2007). How Much More Rain Will Global Warming Bring? *Science*, *317*(5835), 233–235. https://doi.org/10.1126/science.1140746
Werkhoven, K. van, Wagener, T., Reed, P., & Tang, Y. (2008). Characterization of watershed model behavior across a hydroclimatic gradient. *Water Resources Research*, *44*(1). https://doi.org/10.1029/2007WR006271
Willard, J., Jia, X., Xu, S., Steinbach, M., & Kumar, V. (2020). Integrating Physics-Based Modeling with Machine Learning: A Survey. *arXiv:2003.04919 [Physics, Stat]*. Retrieved from http://arxiv.org/abs/2003.04919
Wu, S., Zhao, J., Wang, H., & Sivapalan, M. (2021). Regional Patterns and Physical Controls of Streamflow Generation Across the Conterminous United States. *Water Resources Research*, *57*(6), e2020WR028086. https://doi.org/10.1029/2020WR028086
Xie, K., Liu, P., Zhang, J., Han, D., Wang, G., & Shen, C. (2021). Physics-guided deep learning for rainfall-runoff modeling by considering extreme events and monotonic relationships. *Journal of Hydrology*, *603*, 127043. https://doi.org/10.1016/j.jhydrol.2021.127043

Ye, S., Li, H.-Y., Huang, M., Ali, M., Leng, G., Leung, L. R., et al. (2014). Regionalization of subsurface stormflow parameters of hydrologic models: Derivation from regional analysis of streamflow recession curves. *Journal of Hydrology*, *519*, 670–682. https://doi.org/10.1016/j.jhydrol.2014.07.017

Yue, S., Pilon, P., & Cavadias, G. (2002). Power of the Mann–Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series. *Journal of Hydrology*, *259*(1), 254–271. https://doi.org/10.1016/S0022-1694(01)00594-7

Zaslavsky, I., Whitenack, T., Williams, M., Tarboton, D., Schreuders, K., & Aufdenkampe, A. (n.d.-a). The Initial Design of Data Sharing Infrastructure for the Critical Zone Observatory, 6.

Zaslavsky, I., Whitenack, T., Williams, M., Tarboton, D., Schreuders, K., & Aufdenkampe, A. (n.d.-b). The Initial Design of Data Sharing Infrastructure for the Critical Zone Observatory, 6.

Appendices

A. Extreme events detection based on seasonal anomalies

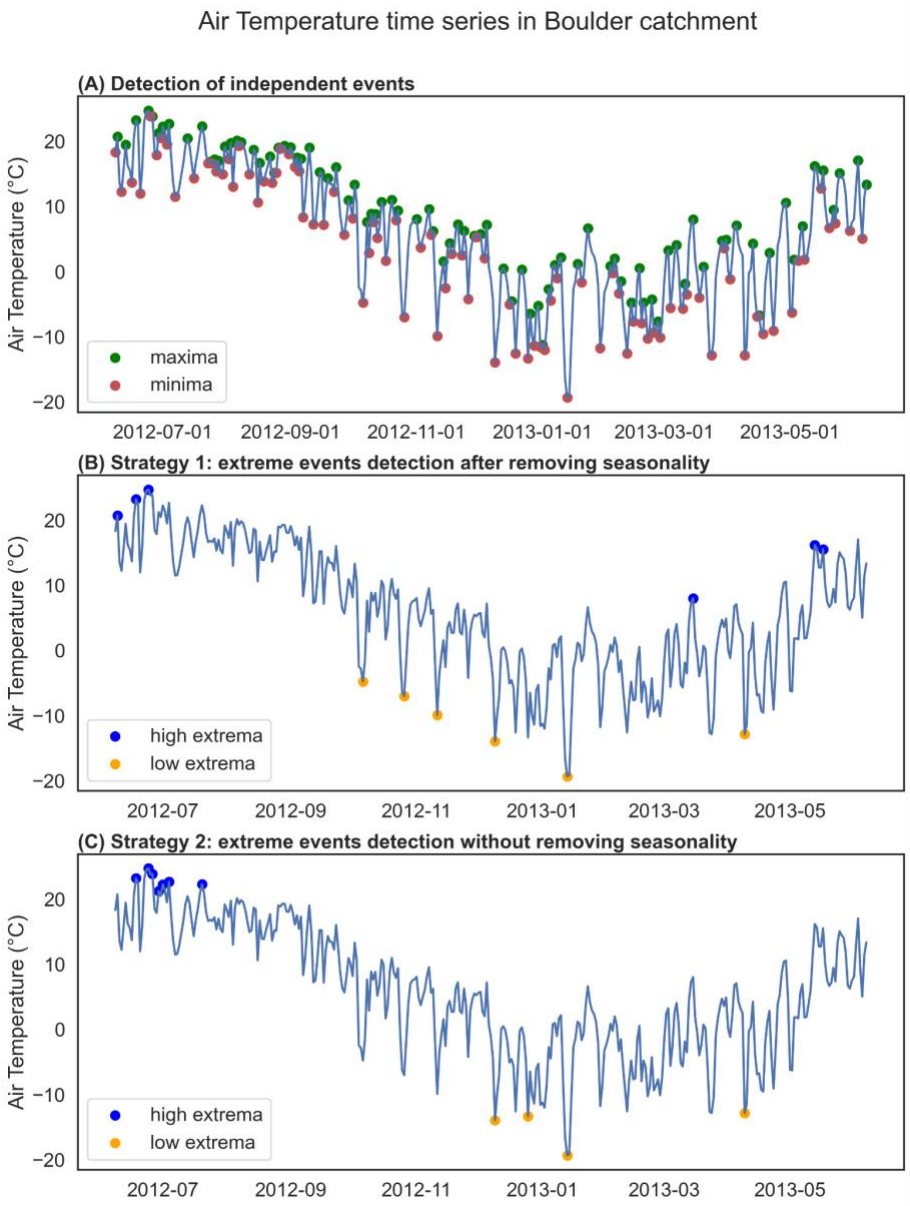


Figure A. (A) The detection of independent events. (B) and (C) show two strategies to detect extreme events. We used the first strategy to generate the results in Chapter 2. The data are air temperature time series from the Boulder study area.

B. Significant trends of frequency and magnitude of the extreme events

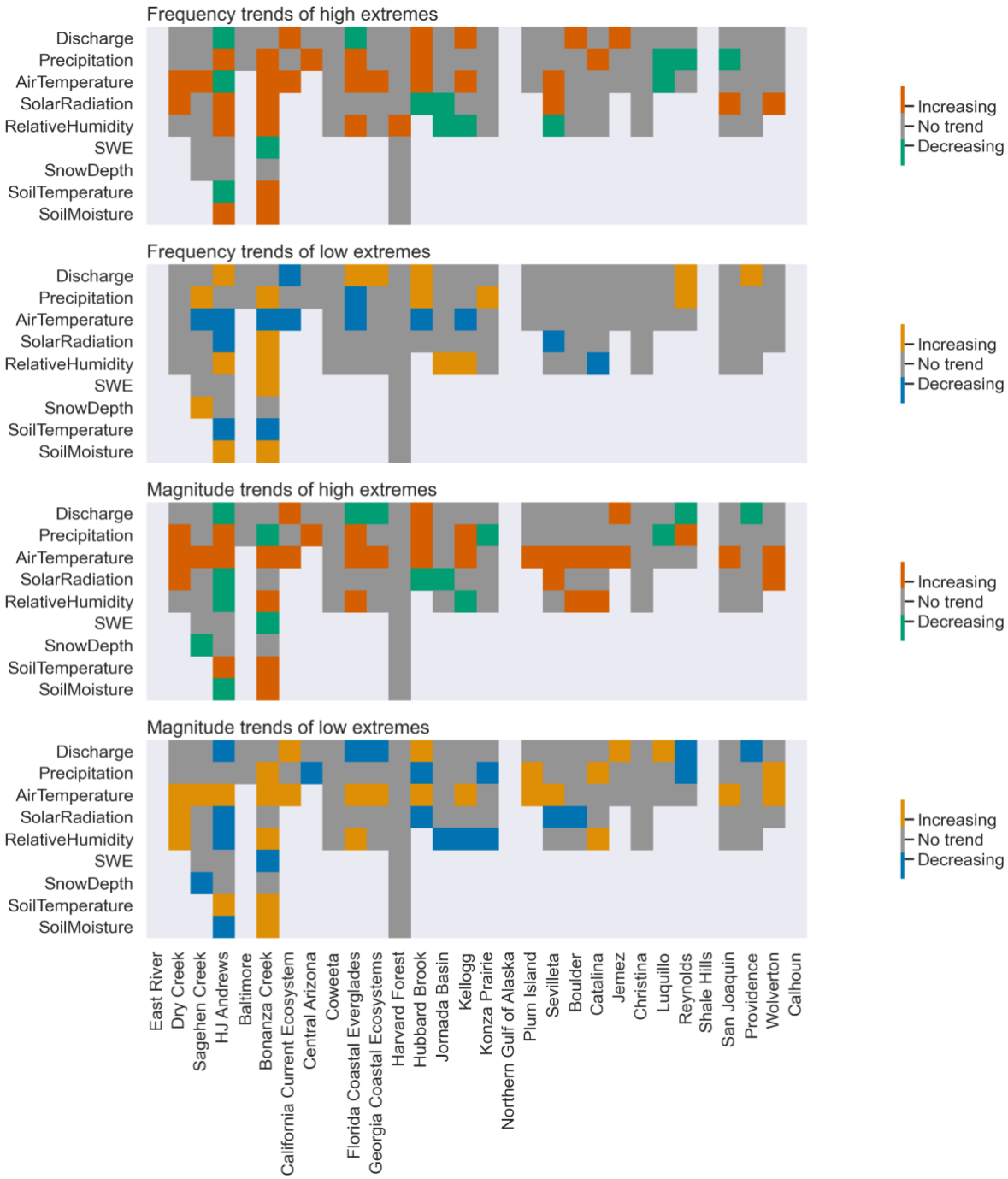


Figure B1. Significant trends of frequency and magnitude of the extreme events. Light grey color indicates that no data are available or data record is shorter than 10 years.

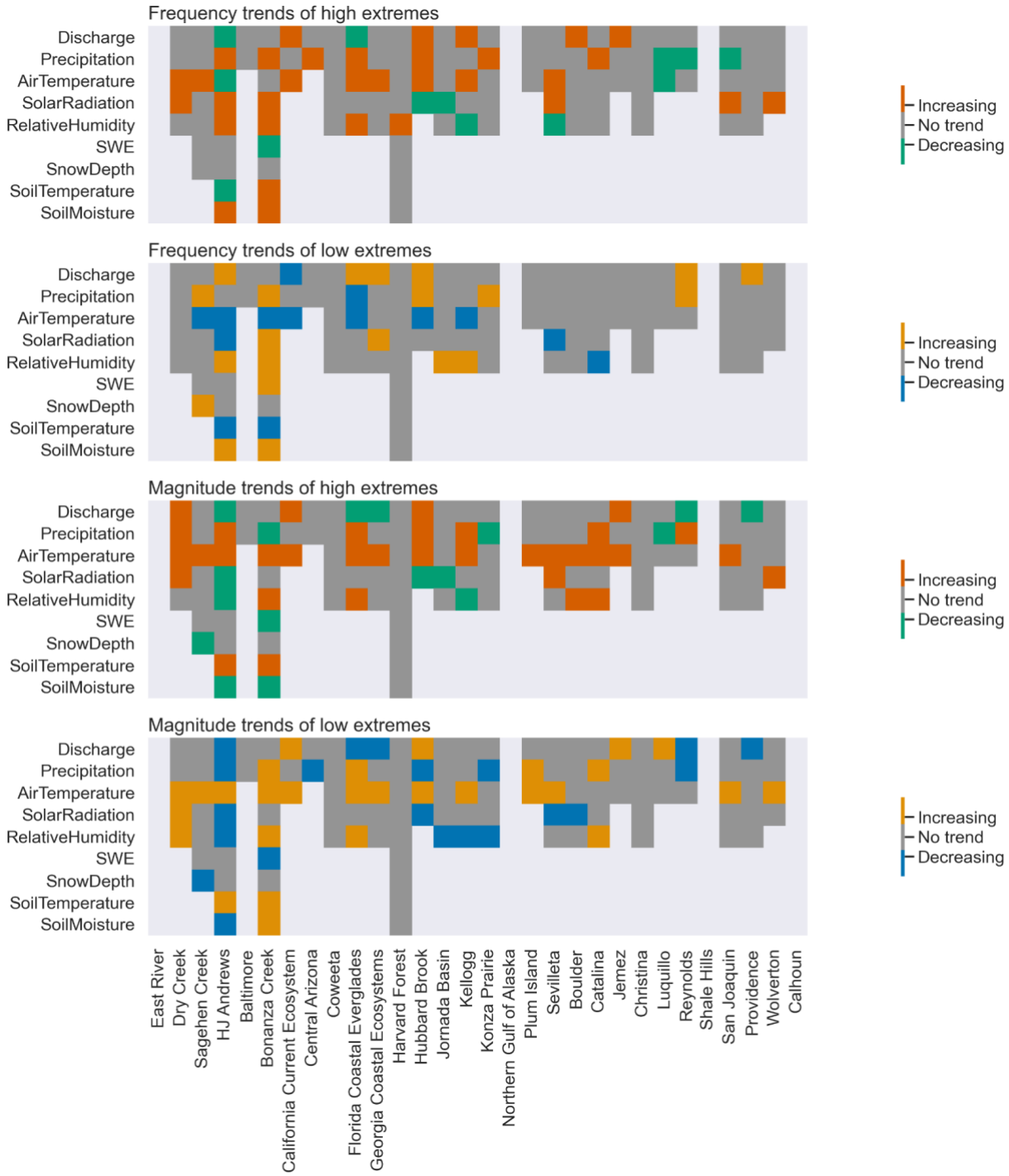


Figure B2. Significant trends of frequency and magnitude of the extreme events analyzed using data excluding those generated using the climate-catalog method. The comparisons were made in case of artifacts caused by reconstructed data using the climate catalog method.

C. Changes in geographical exposure with reference to 1981 to 2005.

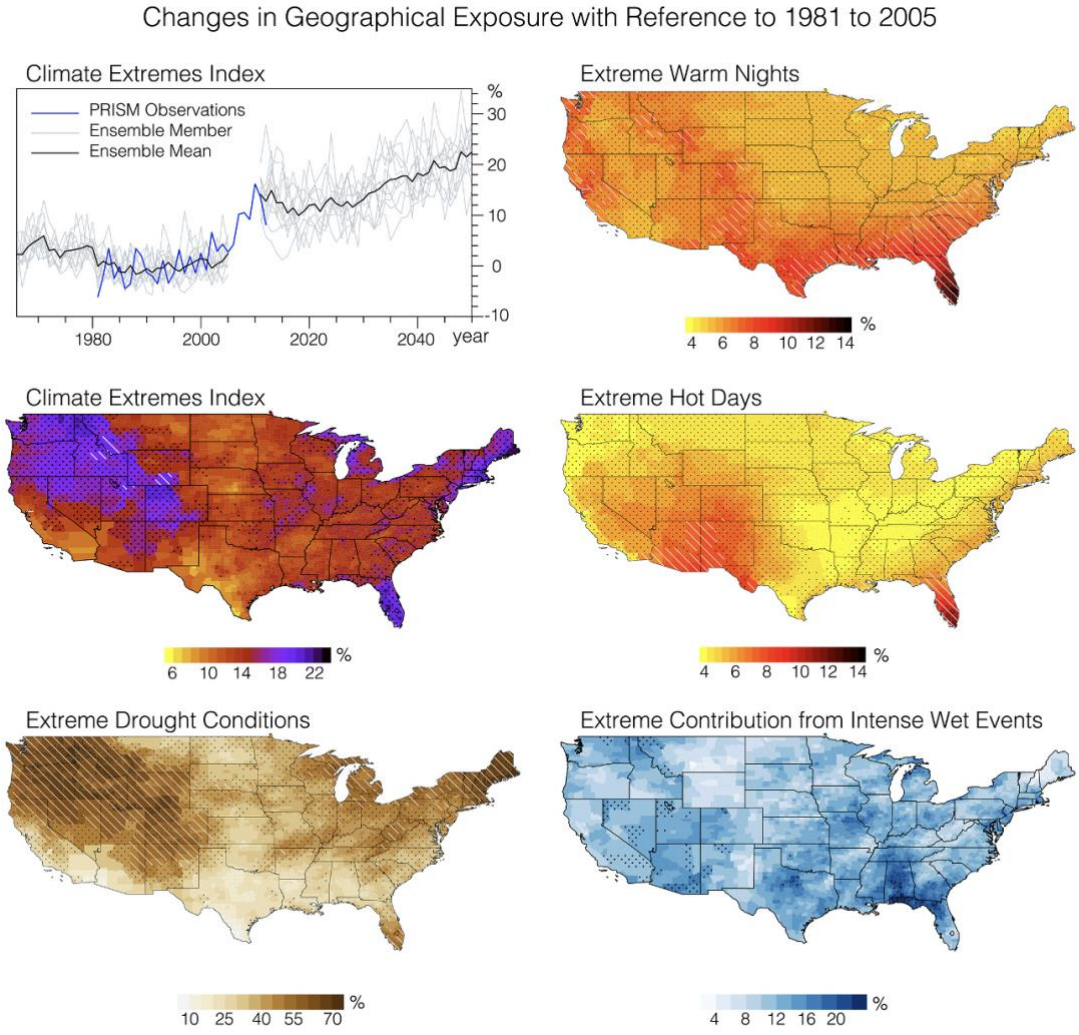


Figure C. Changes in geographical exposure with reference to 1981 to 2005. a) Continental-scale annual changes (with reference to 1981 to 2005) in geographical exposure to extremes based on the CEI in the observations (1981 to 2012), the historical period (1966 to 2005), and the future period (2011 to 2050) model simulations. County-scale changes (with reference to 1981 to 2005) in geographical exposure to b) the CEI, and extreme c) drought conditions, d) warm nights e) hot days, f) contribution from intense precipitation events. Stippling represents counties where projected changes are at least one times (black dots) or two times (diagonal lines) greater than the baseline variability (Batibeniz et al., 2020).

D. The LSTM model

LSTMs are a subclass of recurrent neural networks that consider antecedent conditions by using lagged observations as input. This generates models that have "memory" of a user-defined length. Cell states (c^t) in the LSTM store "long-term memory" while hidden states (h^t) store "short-term memory". With each time step, new data are fed into the model and the cell states and hidden states are updated. The current cell state (c^t) consists of the former cell state (c^{t-1})modified by data added via the input gate (i^t) and removed via the forget gate (f^t). The adjustments made by the input and forget gates are both determined by the former hidden state (h^{t-1}) value and new input data (x^t). The current hidden state (h^t) value is generated by multiplying the output state (o^t) and the current updated cell state (c^t). The output state is a function of the former hidden state and new input data. Finally, the prediction at each time step is a linear transformation of the current hidden state.

$$i^t = \sigma(W_{ii}x^t + b_{ii} + W_{hi}h^{t-1} + b_{hi})$$
$$f^t = \sigma(W_{if}x^t + b_{if} + W_{hf}h^{t-1} + b_{hf})$$
$$g^t = tanh(W_{ig}x^t + b_{ig} + W_{hg}h^{t-1} + b_{hg})$$
$$o^t = \sigma(W_{io}x^t + b_{io} + W_{ho}h^{t-1} + b_{ho})$$
$$c^t = f^t \odot c^{t-1} + i^t \odot g^t$$
$$h^t = o^t \odot tanh(c^t)$$

At each time step, the hidden state and cell state value have the same length as the input sequence. For example, if each input sequence is 270 days long, each individual cell and hidden state will also be 270 days long. Although both the cell state and the hidden state are responsive to input data, the hidden state, which represents the short-term memory, is more sensitive to new data at each time step.

E. Evaluation metrics

(1) Nash-Sutcliffe model efficiency coefficient (NSE)

$$NSE = 1 - \frac{\sum_{t=1}^{T}(Q_{obs}^t - Q_{pred}^t)^2}{\sum_{t=1}^{T}(Q_{obs}^t - \overline{Q_{obs}^t})^2}$$

(2) Pearson correlation coefficient (r)

$$r = \frac{\sum_{t=1}^{T}(Q_{obs}^t - \overline{Q_{obs}^t})(Q_{pred}^t - \overline{Q_{pred}^t})}{\sqrt{\sum_{t=1}^{T}(Q_{obs}^t - \overline{Q_{obs}^t})^2}\sqrt{\sum_{t=1}^{T}(Q_{pred}^t - \overline{Q_{pred}^t})^2}}$$

(3) Bias in the Kling-Gupta Efficiency (KGE)

$$Bias = \frac{\sum_{t=1}^{T} Q_{pred}^{t}}{\sum_{t=1}^{T} Q_{obs}^{t}} \times 100\%$$

(4) Bias in FDC high-segment volume (FHV)

$$FHV = \frac{\sum_{h=1}^{H}(Q_{pred}^{h} - Q_{obs}^{h})}{\sum_{h=1}^{H} Q_{obs}^{h}} \times 100\%$$

where all the discharges in summation have the frequency < 0.02 (H is the total number of high discharge values).

(5) Bias in FDC mid-segment slope (FMS)

$$FMS = \frac{[\log(Q_{pred}^{0.2}) - \log(Q_{pred}^{0.7})] - [\log(Q_{obs}^{0.2}) - \log(Q_{obs}^{0.7})]}{[\log(Q_{obs}^{0.2}) - \log(Q_{obs}^{0.7})]} \times 100\%$$

Q0.2 and Q0.7 represent the discharge values with a frequency of 0.2 and 0.7 in the flow curve respectively.

(6) Bias in FDC low-segment volume (FLV)

$$FLV = -1 \cdot \frac{\sum_{l=1}^{L}[\log(Q_{pred}^{l}) - \log(Q_{pred}^{L})] - \sum_{l=1}^{L}[\log(Q_{obs}^{l}) - \log(Q_{obs}^{L})]}{\sum_{l=1}^{L}[\log(Q_{obs}^{l}) - \log(Q_{obs}^{L})]} \times 100\%$$

where all the discharges in summation have the frequency >0.7. L is the total number of low discharge values. Q^L is the minimum value of discharge except for zero values. Note that we filter out the zero discharge values before calculating the frequency of discharge to avoid zero division.

F. Feature importance in site-specific model

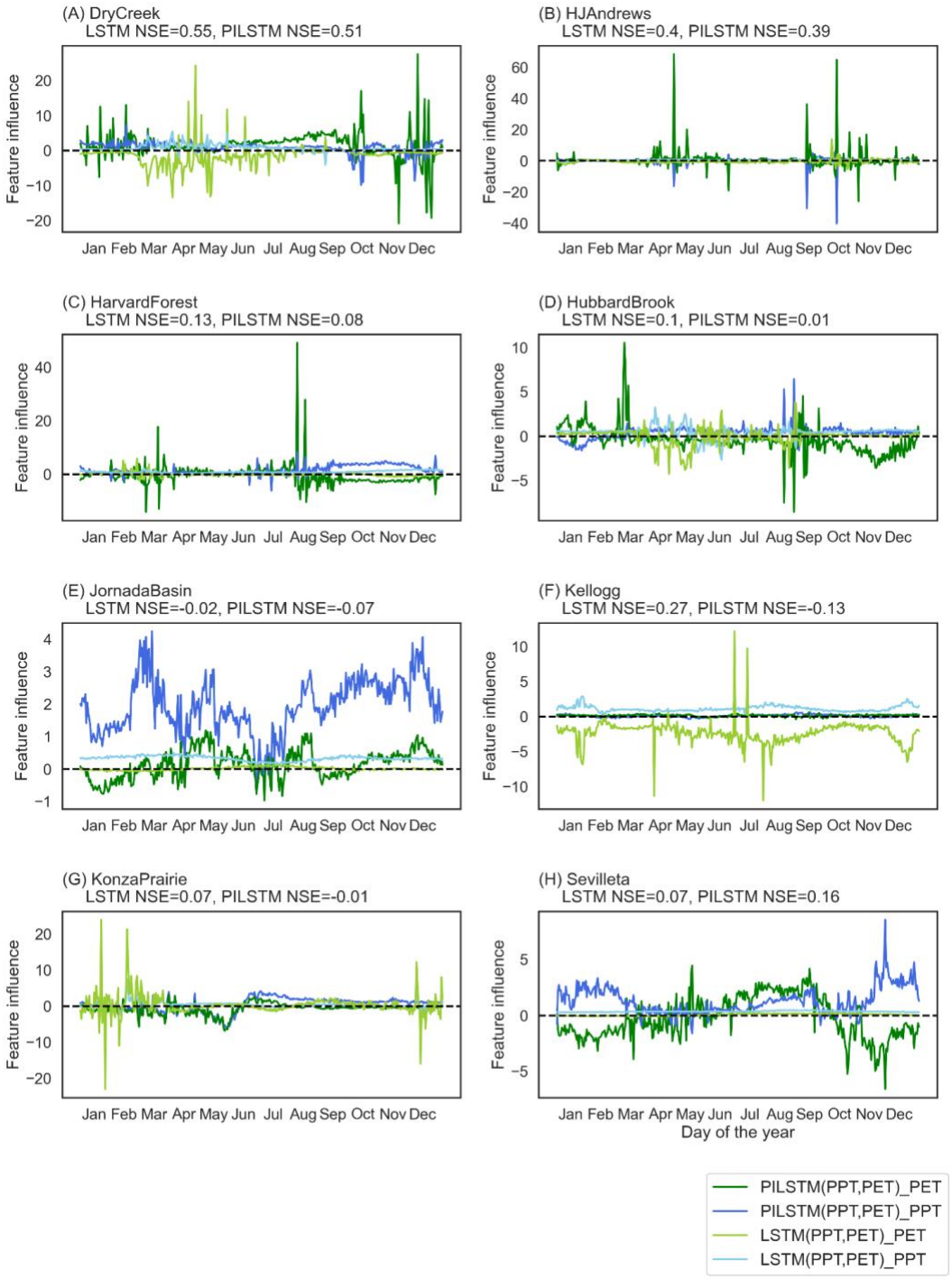


Figure F. Impact of input features using the integrated gradients method for site-specific PILSTM models. All PILSTM models employ a lambda value of one.

G. Features used for watershed clustering

G1. Site attributes of CAMELS sites

Name of the attribute	Description	Type
area_gages2	Catchment area	Topography
elev_mean	Catchment mean elevation	
slope_mean	Catchment mean slope	
aridity	Aridity (ratio of mean PET to mean precipitation)	Climate
frac_snow	Fraction of precipitation falling as snow	
high_prec_dur	Average duration of high precipitation events (number of consecutive days >= 5 times mean daily precipitation)	
high_prec_freq	Frequency of high precipitation days (>= 5 times mean daily precipitation)	
low_prec_dur	Average duration of dry periods (number of consecutive days <1 mm/day)	
low_prec_freq	Frequency of dry days (<1 mm/day)	
p_mean	Mean daily precipitation	
p_seasonality	Seasonality and timing of precipitation (estimated using sine curves to represent the annual temperature and precipitation cycles, positive [negative] values indicate that precipitation peaks in summer [winter], and values close to 0 indicate uniform precipitation throughout the year)	
pet_mean	Mean daily PET [estimated by N15 using Priestley-Taylor formulation calibrated for each catchment]	
frac_forest	Forest fraction	

gvf_diff	Difference between the maximum and minimum monthly mean of the green vegetation fraction (based on 12 monthly means)	Land cover
gvf_max	Maximum monthly mean of the green vegetation fraction (based on 12 monthly means)	
lai_diff	Difference between the maximum and minimum monthly mean of the leaf area index (based on 12 monthly means)	
lai_max	Maximum monthly mean of the leaf area index (based on 12 monthly means)	
clay_frac	Clay fraction (of the soil material smaller than 2 mm, layers marked as organic material, water, bedrock, and "other" were excluded)	Soil
max_water_content	Maximum water content (combination of porosity and soil_depth_statgso, layers marked as water, bedrock, and "other" were excluded)	
sand_frac	Sand fraction	
silt_frac	Silt fraction	
soil_conductivity	Saturated hydraulic conductivity	
soil_depth_pelletier	Depth to bedrock (maximum 50m)	
soil_depth_statsgo	Soil depth (maximum 1.5m, layers marked as water and bedrock were excluded)	
soil_porosity	Volumetric porosity	
carbonate_rocks_frac	Fraction of the catchment area characterized as "Carbonate sedimentary rocks"	Geology
geol_permeability	Subsurface permeability (log10)	

The information about characteristics is referenced from the CAMELS dataset.

G2. Hydrological signatures of CAMELS sites

Label	Definition	Process implication

R[Pint,RC]	The Spearman correlation coefficients between event runoff coefficients and event rainfall intensity	Stormflow processes which are sensitive to rainfall intensity, for example, HOF
R[Pvol,RC]	The Spearman correlation coefficients between event runoff coefficients and event rainfall volume	Stormflow processes which are sensitive to rainfall volume, for example, SSF1, SOF, SSF1, and GWF
R[S,RC]	The Spearman correlation coefficients between event runoff coefficients and pre-event storage	Stormflow processes which are sensitive to pre-event catchment storage, for example, SOF, SSF1, GWF, and SSF2
TS	The characteristic time scale of event runoff response, estimated based on a linear-reservoir-based net-rainfall-runoff model	The timing of stormflow response: low TS is related to HOF, SOF, SSF1, and SSF2 and high TS is related to GWF
BFI	The ratio between base flow and total streamflow	The contribution of base flow on total streamflow
VARb	The standard deviation of log-scale base flow time series	The variability of base flow, low variability implies large groundwater storage

The information about characteristics is referenced from table 1 in (Wu et al., 2021).

H. Additional results using different weights in training local models

1. Local exps have no static input features (k means method, inverse of distance-> weight)

Model name	NSE mean	NSE median	Number of sites with NSE values<0	Number of sites where this model performs optimally
global_benchmark	0.48	0.58	25	14
global_static_attr	0.51	0.67	18	89
global_static_hydro	**0.59**	**0.69**	**13**	**155**
global_static_TE	0.49	0.64	15	114
local_attr	0.50	0.59	26	19
local_hydro	0.50	0.59	25	12
local_TE	0.49	0.59	25	12

2. Local exps also use static input features (k means method, inverse of distance-> weight)

Model name	NSE mean	NSE median	Number of sites with NSE values<0	Number of sites where this model performs optimally
global_benchmark	0.48	0.58	25	18
global_static_attr	0.51	0.67	18	56
global_static_hydro	**0.59**	**0.69**	**13**	85
global_static_TE	0.49	0.64	15	51
local_attr_w_static	0.50	0.66	18	64
local_hydro_w_static	0.58	**0.69**	15	**88**
local_TE_w_static	-0.32	0.65	23	53

I. Results using the exponential of probability of watershed belonging to each cluster as the weights in local LSTM model trained for each cluster

Model name	NSE median	NSE mean	Number of sites with NSE values<0	Number of sites where this model performs optimally
global_benchmark	0.58	0.48	25	16
global_attr	0.67^{a}	0.51	18	65
global_hydro	0.69^{b}	0.59	13	92
global_TE	0.64^{a}	0.49	15	47
local_attr	0.66^{a}	0.52	20	61
local_hydro	0.69^{b}	0.59	16	82
local_TE	0.65^{a}	0.39	18	52