ASSESSING THE ROBUSTNESS OF DEEP LEARNING STREAMFLOW MODELS UNDER CLIMATE CHANGE

by

LOGAN MICHELLE QUALLS

GEOFFREY R. TICK, COMMITTEE CHAIR
BO ZHANG
YONG ZHANG
GREY S. NEARING

A THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science
in the Department of Geological Sciences
in the Graduate School of
The University of Alabama

TUSCALOOSA, ALABAMA

2022

ABSTRACT

Long Short-Term Memory networks provide the most accurate rainfall-runoff predictions to-date, but their reliability under climate change is not well understood. We explore the robustness of these models under climate nonstationarity by creating train and test data splits that are designed to simulate climate bias. By training on forcing data from hydrological years of high (low) aridity and testing on data from hydrological years of low (high) aridity, we can begin to quantify the performance and relative robustness of that performance under climate nonstationarity. We benchmark against a calibrated conceptual model (the Sacramento Soil Moisture Accounting model) and a calibrated process-based model (the NOAA National Water Model) and found that LSTMs were generally more accurate than both, even when trained on climatologically biased data splits. The process-based model did not show as large of a performance gap as the conceptual and deep learning models, however (i) this model was not calibrated on a climate-biased data split and (ii) LSTMs always out-performed the process-based benchmark, even when the LSTM training data had climatological bias. We find that although all hydrologic models reported here degrade under nonstationarity, DL models demonstrate greater robustness. We also tested the hypothesis that dynamic climate attributes as inputs into the LSTM would improve performance under climate nonstationarity. We found no predictive value with the addition of dynamic, as opposed to static, climate attribute inputs

LIST OF ABBREVIATIONS

DL	Deep Learning
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
PUB	Predictions in Ungauged Basins
NCAR	National Center for Atmospheric Research
CAMELS	Catchment Attributes and Meteorology for Large-Sample Studies
NLDAS	North American Land Data Assimilation System
USGS	United States Geological Survey
NH	NeuralHydrology
NSE	Nash-Sutcliffe Efficiency
NWM	National Water Model
SAC-SMA	Sacramento Soil Moisture Accounting Model
DDS	Dynamically Dimensioned Search
MSE	Mean Squared Error
NOAA	National Oceanic and Atmospheric Administration
CONUS	Contiguous United States
CDF	Cumulative Density Function

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my academic advisor-turned-external committee member-turned-friend, Grey Nearing. His honesty serves as a mirror for self-reflection, his faith in me harvests faith in myself, and his willingness to take me on as a “Hello World!”-level coder is something I do not understand but am beyond thankful for.

I would also like to thank my academic advisor, Geoffrey Tick. With little notice (thanks, Grey), he accepted me as his own and advised me every step of the way. His role as an anchor has proven more valuable than he realizes.

Jonathan Frame is invaluable in this work and beyond. I consider him a friend and a mentor; his willingness to meet, dedicate time and energy to my smallest and biggest inquiries, and contribute to this project have been both instrumental and enjoyable. I strive to be as critical a thinker, as hard a worker, and as kind a person.

I have made many friends at the University of Alabama, and I intend to carry them with me far beyond this chapter of my life. Their loyalty, patience, and support make even the darkest moments feel light.

Finally, to my mom and dad: Thanks for not giving up on me when my first-grade teacher told you I would never make it through high school. I owe it all to you.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Deep learning (DL) hydrology models outperform their traditional counterparts in simulating streamflow both locally and regionally (Kratzert et al., 2019c). Traditional models, including conceptual and process-based models, are designed to make predictions using equations that explicitly represent the governing processes in a watershed. Although these models can (arguably) be useful for providing insight into local hydrological systems, they do not generalize well (Beven, 2000) and are less useful for making accurate predictions. DL models can learn and extrapolate from diverse, large-sample datasets, in part by learning representations of catchment similarity (Kratzert et al., 2019a; Nearing et al., 2020).

Unlike standard artificial neural networks that were used for rainfall-runoff modeling prior to Kratzert et al., (2018) (e.g., Hsu et al., 1995), recurrent neural networks (RNN) are designed explicitly to recognize patterns in time series data by maintaining temporal relationships between time steps. While RNN architectures allow information to persist in a model through time, their hidden state remains largely unconstrained. Unlike standard RNNs, Long Short-Term Memory (LSTM) models utilize a set of *gates* (gates, in this sense, are shallow neural networks) to regulate the model's memory, or cell state. LSTMs are currently state of the art in making rainfall-runoff predictions, and their success relative to other deep learning strategies is due in part to their ability to retain a working memory about meteorological inputs and hydrological responses from previous time steps and apply that memory to make future predictions.

Historically there have been concerns about the reliability of data-driven models in out-of-sample conditions (e.g., Kirchner et al., 2006; Milly et al., 2008; Sellars, 2018). Early experiments on shallow neural networks highlighted cases where data-driven models were unreliable in extrapolation (e.g., Cameron et al., 2002; Gaume and Gosset, 2003). However, recent literature suggests that LSTMs are robust under at least certain kinds of extrapolation. For example, prediction in ungauged basins (PUB) is a longstanding challenge for hydrological models (Hrachowitz et al., 2013), and Kratzert et al. (2019b) showed that LSTMs provide predictive accuracies in *ungauged* basins that are, on average, better than well-calibrated conceptual models in *gauged* basins. In addition, LSTMs have proven to be more robust than conceptual and process-based models in predicting extreme events that were not part of the training set – Frame et al. (2021a) showed that not only do LSTMs make more accurate predictions of extreme events in general, they are also more robust in extrapolating to events with increasing return periods (i.e., decreasing probability).

In addition to extrapolation to new locations and to extreme events, hydrological models should also be robust to climate nonstationarity (Milly et al., 2006). A paper in review by Wi et al. (2022) added increasing trends to temperature inputs and found that conceptual models decreased production (runoff ratio) everywhere, while LSTMs learned to account for melt-driven processes (also see Kratzert et al., 2019a) and increased streamflow in basins with glacier melt. However, to our knowledge, no study has assessed the quantitative accuracy of LSTMs under climate-biased training data.

In this paper, we test two things. First, we quantify the skill of LSTM rainfall-runoff models under test periods with climatologies that are biased relative to training data. Second, we

test the hypothesis that adding dynamic (changing) climate statistics as inputs into the LSTM will help the model to adapt to changing meteorological distributions.

The challenge of testing a model's robustness to trends in input data is that we only have past data to validate against – we can't evaluate on future change. Therefore, to introduce bias in test data our approach is to train models on forcing data from water-years characterized by one climate extreme and test on water-years characterized by the opposite climate extreme. We split the data according to an annual aridity index, which accounts for both water and energy. Model performance was measured and benchmarked against a conceptual model (the Sacramento Soil Moisture Accounting model; SAC-SMA) and a process-based model (the US National Water Model; NWM). Results were used to explore potential implications of environmental change on hydrological predictions.

METHODS

Data

The CAMELS Dataset

The National Center for Atmospheric Research (NCAR) Catchment Attributes and Meteorological data for Large-Sample Studies (CAMELS) is a publicly available, large-scale dataset detailing roughly 30 years of hydrological responses (1980-2009) from 671 watersheds within the contiguous United States (CONUS; Newman et al., 2014). Kratzert et al (2019b) extended the CAMELS data record through 2014, with data available through the HydroShare, managed by the Consortium of Universities for the Advancement of Hydrologic Science (see Code and Data Availability Section). We refer to this extended CAMELS data record as the CAMELS data record.

Three meteorological forcing datasets are available through CAMELS, including Daymet (Thorton et al., 1997), Maurer (Maurer et al., 2002), and the North American Land Data Assimilation System (NLDAS; Xia et al., 2012). Each forcing product reports daily measurements of precipitation (mm/d), minimum and maximum temperature (C), vapor pressure (Pa), and surface radiation (W/m2). Corresponding United States Geological Survey (USGS) daily streamflow data (cfs) are also provided from the USGS Water Information System.

All experiments reported in this paper use Daymet or NLDAS forcing data. These sources were chosen based on two factors: (i) Kratzert et al. (2021) found that of the three

CAMELS forcing data products, Daymet forcings generally resulted in the most accurate models, and (ii) the NWM, which is one of our benchmarks; see the Multi-Model Benchmarking Section) uses NLDAS forcings. Kratzert et al. (2020) also found that using multiple forcing data sources (i.e., all three CAMELS forcings) simultaneously provides additional information (improved model skill), however the models that we benchmark against in this paper are only able to use one precipitation forcing data source at a time, so we train DL models using only one source at a time to maintain a fair benchmark. Additionally, to be consistent with previous studies (Newman et al., 2015; Kratzert et al., 2019a,b; Gauch et al., 2021; Frame et al., 2021b; Klotz et al., 2021), we use only the 531 CAMELS basins that were selected for model benchmarking by the original CAMELS authors (Newman et al., 2015).

Supplemental work by Addor et al. (2017) provides a dataset containing static (unchanging) basin attributes related to climate, geology, land cover, soil, streamflow, and topography. The static climate indexes available through this work were derived from Daymet forcing data. The full list of static attributes used in this study is given in Table 1.

Daily Meterological Forcing Inputs	
Maximum Air Temperature	2 m daily maximum air temperature (°C)
Minimum Air Temperature	2 m daily minimum air temperature (°C)
Precipitation	Average daily precipitation (mm/day)
Radiation	Surface-incident solar radiation (W/m^2)
Vapor Pressure	Near-surface daily average (P_a)
Static Catchment Indexes	
Climate Indexes	
Precipitation	Mean daily precipitation ($mm\ day^{-1}$)
Potential Evapotranspiration (PET)	Mean daily evapotranspiration ($mm\ day^{-1}$)
Aridity Index	Ratio of mean PET and mean precipitation (~)
Precipitation Seasonality	Estimated deviation of precipitation seasonally (~)
Snow Fraction	Fraction of precipitation falling on days with temperatures < 0 ◦C (~)
High Precipitation Frequency	Frequency of days with $\leq 5\times$ mean daily precipitation ($days\ year^{-1}$)
High Precipitaion Duration	Average duration of high precipitation events (num. consecutive days with $\leq 5\times$ mean daily precipitation; days)
Low Precipitation Frequency	Frequency of dry days (< 1 mm/day; $days\ year^{-1}$)
Low Precipitation Duration	Average duration of dry periods (num. of consecutive days with precipitation < 1 mm/day; days)
Topographic Characteristics	
Elevation	Mean catchment elevation (m above sea level)
Slope	Mean catchment slope ($m\ km^{-1}$)
Area	Catchment area (km^2)
Land Cover Characteristics	
Forest Fraction	Fraction of catchment covered by forest (~)
Leaf Area Index (LAI): Maximum	Maximum monthly mean of leaf area index (~)
LAI: Difference	Difference between the max. and min. mean of the leaf area index (~)
GFV Max	Maximum monthly mean of green vegetation frequency (~)
GVF Difference	Difference between the max. and min. mean of the green vegetation frequency (~)
Soil Characteristics	
Soil Depth (Pelletier)	Depth to bedrock (max. 50 m; m)
Soil Depth (STATSGO)	Depth of soil (max. 1.5 m; m)
Soil Porosity	Volumetric Porosity (~)
Soil Conductivity	Saturated hydraulic conductivity ($cm\ h^{-1}$)
Max Water Content	Maximum water content of the soil (m)
Sand Fraction	Fraction of sand in the soil (%)
Silt Fraction	Fraction of silt in the soil (%)
Clay Fraction	Fraction of clay in the soil (%)
Geological Characteristics	
Carbonate Rocks Fraction	Fraction of catchment characterized as carbonate sedimentary rocks (~)
Geological Permeability	Surface permeability (m^2)

Table 1: List of all daily meteorological forcing inputs and static catchment attributes from the CAMELS dataset used in this study.

Static vs. Dynamic Climate Attributes

The CAMELS static attributes include ten (10) indexes characterizing long-term (1990-2009) climate in each basin (Table 1). Kratzert et al. (2019c) found that these climate indexes were some of the most important static attributes in allowing the model to learn a representation of catchment similarity and Nearing et al. (2019) showed that the LSTM can use dynamic climate indexes to learn a *dynamic* representation of catchment similarity (i.e., a representation of catchment similarity that changes over time). Nearing et al. (2019) found that using dynamic climate indexes did not improve model performance overall (relative to using static climate indexes), and one of the hypotheses that we test here is whether this is also the case under climate-biased training vs. test data.

Dynamic climate indexes were calculated separately for each day in the CAMELS data record using the preceding 365-day period and were calculated (separately) using Daymet and NLDAS meteorological data.

Train/Test Data Sets

We used two types of train/test sets: biased and random. Biased train (test) sets were designed to simulate changing distributions of precipitation and energy availability and are used to evaluate the robustness of models under climate change. For each basin, train and test sets were created by sorting water-years (365-day periods starting on October 1 and ending on September 31) by their dynamic aridity index (i.e., the daily aridity index using the 365-day period ending on September 31). Once sorted, water-years with the five (5) lowest (highest) aridity indexes were grouped for training, and water-years with the five (5) highest (lowest) aridity indexes were grouped for testing. A visualization of this biased train/test split in an example basin is shown in the lower left-hand subplot of Figure 1. We thus ran two (2) sets of

biased experiments: experiments labeled *aridity high* train on water-years with low aridity values and test on water-years with high aridity values (wet-to-dry), and experiments labeled *aridity low* train on water-years with high aridity values and test on water-years with low aridity values (dry-to-wet). This is detailed in Table 2. The *high* and *low* descriptors refer to the model's respective *test* set type.

Set Name	Train Data	Test Data	Simulation
aridity high	5 lowest aridity years	5 highest aridity years	wet-to-dry
aridity low	5 highest aridity years	5 lowest aridity years	dry-to-wet

Table 2: Naming convention and description for biased train/test sets. “High” and “low” in a set’s name refers to its test data type.

Random experiments are experiments that use randomly selected train/test splits in each basin. These were used as a benchmark for the biased climate experiments. Random train/test sets were created by grouping five (5) random years for training and five (5) different random years for testing from each basin. A visualization of how random sets were constructed in an example basin is shown in the lower right-hand subplot of Figure 1. For every biased experiment pair (aridity high and aridity low), we ran five (5) random experiments, each with a different random train/test split. This was done to account for randomness in selecting the five (5) train and five (5) test water-years. It is important to note that although the train/test splits (both biased and random) were chosen separately per basin, all training data from all basins were used to train a single LSTM model. This is critical because LSTMs perform well because they learn and generalize from large datasets. Training LSTMs on individual basins does not yield meaningful results (Nearing et al., 2021).

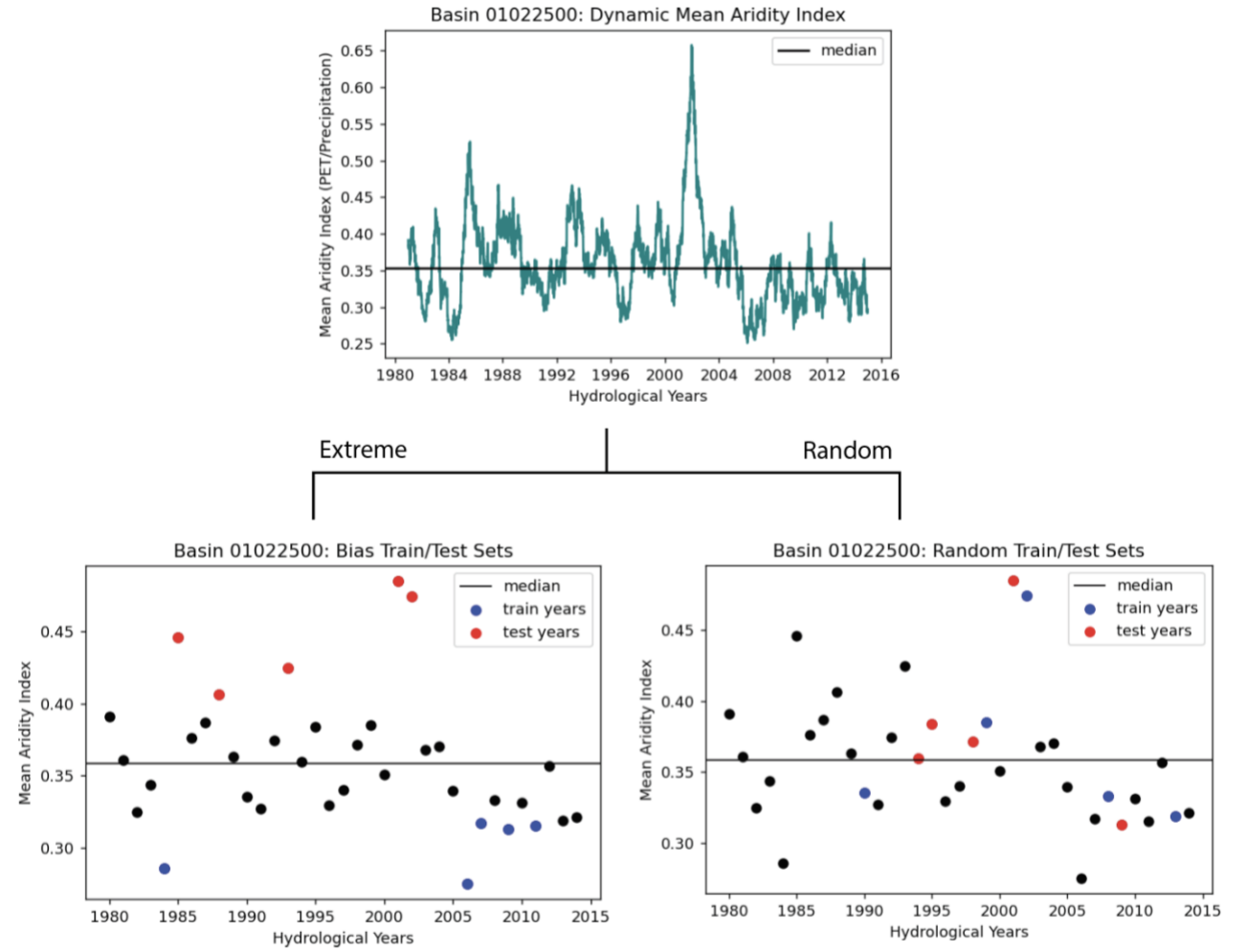

Figure 1: Visualization of the process for selecting biased and random train/test sets in an example basin. Water-years were sorted by their daily aridity index with 365-day lookback (top suplot). In this example, the lowest five water-years were chosen for training while the highest five water years were chosen for testing to create a biased train/test set, constituting an *aridity high* experiment. In the random sets (lower right subplot), five random water years were selected for training and five different random water years were selected for testing. All biased and random experiments were repeated five times to account for randomness in the train/test set.

Models

Deep Learning Models

LSTMs are a type of recurrent neural network that has an explicit memory state that persists through time (called the *cell state*). As such, the LSTM can be thought of as a state-space model, however it is not conservative - i.e., does not conserve mass or energy (however, a conservative version of the LSTM does exist; e.g., Frame et al., 2022). Gates in the LSTM allow for nonlinear transformations of information between the model inputs and the cell state, and between the cell state and model outputs. The three main gates employed in LSTMs are the input gate, which controls the input-state relationship, the forget gate, which controls the memory timescales in the model, and output gate, which controls the state-output relationship. Kratzert et al. (2018) give a detailed and accessible introduction to the LSTM architecture.

NeuralHydrology (NH) is a Python package specifically developed for deep learning hydrological modeling and is designed to work seamlessly with the CAMELS dataset (Kratzert et al., 2022). NH provides several deep learning model types, including a standard LSTM, the Entity-Aware LSTM used by Kratzert et al. (2019c), and the Mass-Conserving LSTM used by Heodt et al. (2021).

We draw on prior work to choose hyperparameters for the models used in this study (Kratzert et al., 2019a,b; Kratzert et al., 2020). Our LSTMs use a cell state dimension of 128 and a linear layer on the LSTM outputs with a dropout rate of 0.4. Each model was trained for 30 epochs with a batch size of 256 using the Adam optimizer and a learning rate schedule that starts at 1e-3. The loss function is the average per-basin Nash-Sutcliffe Efficiency (NSE) described by Kratzert et al. (2019a). All data were pre-normalized to zero mean and unit variance, and predicted streamflow was clipped to zero during testing, as negative streamflow is not applicable

in most cases (predictions were not clipped during training to preserve gradients). All LSTM experiments were run with 10 ensemble members to minimize the effect of random initialization of weights and biases. All statistics reported for LSTM runs were calculated on the mean hydrograph from that 10-member ensemble (see Kratzert et al., 2019b for a discussion about the need for ensembling).

Benchmark Models

Sacramento Soil Moisture Accounting Model + Snow17

The CAMELS dataset includes output from calibrated runs of the SAC-SMA model for all basins. The SAC-SMA model calibrated by NCAR for CAMELS includes the Snow-17 (Anderson, 1973) snow module and a unit hydrograph routing model. It was necessary to calibrate SAC-SMA for benchmarking biased train/test splits - we did this with the GitHub repository developed by Nearing et al. (2020), which provides a Python interface between the original Fortran SAC-SMA source code and the Spotpy Python optimization package (Houska et al., 2019). All SAC-SMA runs reported in this study were calibrated separately for each basin using a Dynamically Dimensioned Search (DDS) optimizer (Tolson & Shoemaker, 2006) with a mean-squared error loss function and 10,000 model runs. In unreported benchmarks, this version of SAC-SMA performed similarly to the NCAR calibrated SAC-SMA runs included in the CAMELS dataset over the same train/test data splits (our calibrations were slightly better on average in the validation period).

A notable difference between SAC-SMA calibrations and LSTM training is that the LSTM is always trained regionally (on all training data from all basins at once), while SAC-SMA is always calibrated separately per basin. It should be noted that SAC-SMA does not use

static (or dynamic) catchment attribute data, and instead relies on basin-specific calibrated parameters. Like the LSTM, all SAC-SMA statistics reported in this study are calculated on the hydrograph that results from averaging ten (10) separate ensemble members, which vary only in the initial values and inherent randomness in calibration.

National Water Model

In addition to the SAC-SMA benchmarks, we benchmarked the LSTM against the NWM version 2 Reanalysis (Frame et al., 2021b). The NWM is a spatially distributed, process-based model used by the National Oceanic and Atmospheric Administration's (NOAA). The NWM is based on WRF-Hydro (Salas et al., 2018).

The NWM was calibrated by NOAA using the CONUS Retrospective Dataset Version 2.0 and with continuous (not intentionally-biased) train periods. The NWM v2 reanalysis is a single model run forced by NLDAS data for the period (1995-2014) and lacks the 10 ensemble members that the LSTM and SAC-SMA experiments use in this study. The non-biased calibration gives the NWM an advantage over the LSTM and SAC-SMA. In normal conditions, the NWM is outperformed by deep learning and conceptual models (Kratzert et al., 2019b; Frame et al., 2021b).

With these differences in mind, a benchmark was created by calculating performance metrics for the NWM on the same test years as the SAC-SMA and LSTM experiments. Note that because the NWM v2 reanalysis does not span the full CAMELS data period, we conduct separate experiments using for both the LSTM and SAC-SMA using only the years in this date range (1995-2014). These experiments are in addition to experiments (that do not include an NWM benchmark) using data from the full CAMELS time period.

Experimental Design

To reiterate, the two objectives of this study are to 1) characterize the resilience of DL models under changes in distribution of aridity index and 2) assess the value of including dynamic, rather than static, climate attributes as inputs to regionally-trained DL LSTM models, specifically when the training data is climate-biased relative to test data. We address these two objectives using the experiments described in the following subsections.

Model runs are categorized on a combination of characteristics that refer to whether it was trained (i) on a biased or random train set, (ii) with Daymet or NLDAS forcing source, (iii) on data from years extracted from all CAMELS years (1980-2014) or data extracted from years available for the NWM (1995-2014), (iv) with static or dynamic climate attribute inputs. A summary of all LSTM and SAC-SMA model runs reported in this study is illustrated in Figure 2.

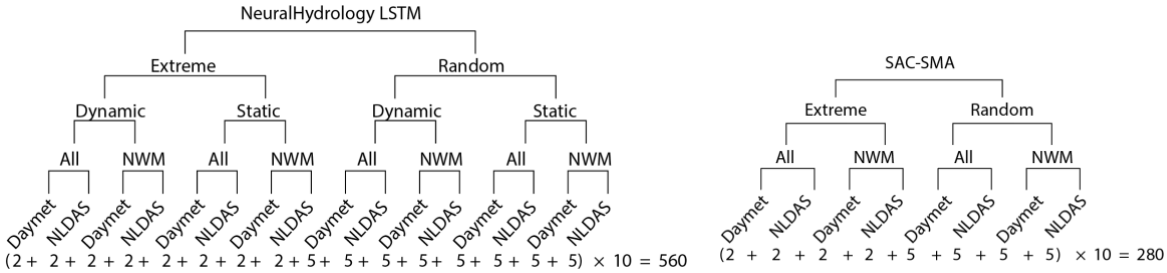

Figure 2: Visualization of all LSTM and SAC-SMA runs reported in this paper. Biased model runs include two (2) experiments, one for each biased train/test set listed in Table 2. All random runs include five (5) experiments, each with a different random train/test set. Each model reported here is an ensemble of ten (10) independently calibrated models to account for random effects in the train/test set and we report statistics over the mean hydrograph for the ensembles.

Benchmarking Climate-Biased Train/Test Sets

The first set of experiments we report benchmark the LSTM and SAC-SMA models with biased train/test data sets vs. the same models with random train/test data splits. The objective is to understand whether and how much accuracy is lost due to training on data with biased

climatology. Models reported in this section used train/test years extracted from the full CAMELS data record (1980-2014). In addition to comparing overall performance, we explore correlations between differences in model skill (between biased vs. random train/test splits) with basin attributes to try to understand whether performance differences can be understood based on local hydrology.

Multi-Model Benchmarking

The second set of experiments we report benchmark the LSTM on biased train/test sets against SAC-SMA on biased train/test sets and also against the NWM reanalysis. The purpose of these experiments is to contextualize the effect of training DL models on biased inputs, relative to other hydrological models. Because of the difference in available years of data for the NWM vs. CAMELS, and because the NWM used NLDAS forcings, we do this benchmarking in two steps.

First, to leverage the full CAMELS dataset, we benchmark the LSTM and SAC-SMA with both Daymet and NLDAS forcings. We quantify losses in accuracy due to biased training data (both wet and dry) for both models with both forcing products. Second, to allow for benchmarking against the process-based NWM, we benchmark all three models using train/test splits (for the LSTM and SAC-SMA) extracted only from the time period (1995-2014) with NWM v2 reanalysis data. In this case, we only use NLDAS forcings because this is the data product used in the NWM v2 reanalysis.

Static vs. Dynamic Climate Indexes

The final set of experiments tests the effects of using dynamic, rather than static, climate indexes, calculated over 1 year (365-day) time periods. These experiments compare the LSTM

on biased train/test sets with both Daymet and NLDAS forcing data over the full CAMELS time period.

Metrics

We calculate model performance using several standard error metrics, outlined in Table 3. Once a model is trained, tested, and ensembled (with the exception of the NWM), the metrics listed in Table 3 are calculated *only* for the years in the respective test set. It is important to remember that test sets for any given model run are different for each of the 531 basins. This set of performance metrics was chosen based on previous studies (Kratzert et al., 2019b,c; Kratzert et al, 2020; Gauch et al., 2021). Certain standard benchmarking metrics from previous machine learning hydrology studies were excluded due to the fact that these metrics are unstable and less informative, as demonstrated by (Frame et al., 2021a).

Metric	Equation
Mean Squared Error (MSE) [0,∞); values closer to 0 desirable *values close to 0 desirable* *y hat*: simulation *y*: observation	$\frac{1}{T}\sum_{t=1}^{T}\left(Q_m^t - Q_o^t\right)^2$
Root Mean Squared Error (RMSE) [0,∞); values closer to 0 desirable *y hat*: simulation *y*: observation	$\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(Q_m^t - Q_o^t\right)^2}$
Nash-Sutcliffe Efficiency (NSE) (-∞,1]; values close to 1 desirable Q: simulated (*m*) or observed (*o*) streamflow	$1 - \frac{\sum_{t=1}^{T}\left(Q_o^t - Q_m^t\right)^2}{\sum_{t=1}^{T}\left(Q_o^t - \overline{Q}_o\right)^2}$
Alpha NSE (α-NSE) *(0,∞); values close to 1 desirable* σ: STDEV of simulations (*s*) or observations (*o*)	$\sigma_s \div \sigma_o$
Beta NSE (β-NSE) *(-∞,∞); values close to 0 desirable* μ: mean of simulations (s) or observations (o)	$(\mu_s - \mu_o) \div \sigma_o$
Kling-Gupta Efficiency (KGE) *(-∞,1]; values close to 1 desirable* *s*: coefficient weights r: Pearson's r α: α-NSE β: β-KGE	$1 - \sqrt{[s_r(r-1)]^2 + [s_\alpha(\alpha-1)]^2 + [s_\beta(\beta_{KGE}-1)]^2}$
Beta KGE (β-KGE) [-∞,∞); values close to 0 desirable μ: mean of simulations (*s*) or observations (*o*)	$\mu_s - \mu_o$
Pearson's r (r) [-1,1]; 1 indicates perfect positive correlation, -1 indicates perfect negative correlation *m*: mean of vector x (*x*) or y (*y*)	$r = \frac{\sum(x-m_x)(y-m_y)}{\sqrt{\sum(x-m_x)^2\sum(y-m_y)^2}}$

Table 3: Performance metrics used for benchmarking. All metrics are reported in Appendix 1 and NSE scores are reported in the Results section.

While we report all metrics in Table 3 for all model runs, we will only present primary results (figures and analyses) for the NSE scores, since this is (arguably) the most common metric for assessing hydrographs.

RESULTS

Benchmarking Climate-Biased Train/Test Sets

Table 4 shows the median NSE scores (over 531 basins) for all model runs in this paper. We report median NSE scores for biased train/test experiments (aridity high, aridity low), and we report the mean of the five (5) median NSE scores for corresponding random experiments.

		NH				SAC-SMA		NWM	
		Static		Dynamic					
		Experiment	NSE	Experiment	NSE	Experiment	NSE	Experiment	NSE
Daymet	NWM	Aridity High	0.741	Aridity High	0.737	Aridity High	0.498		
		Aridity Low	0.751	Aridity Low	0.748	Aridity Low	0.559		
		Random	0.846	Random	0.844	Random	0.661		
	All	Aridity High	0.711	Aridity High	0.694	Aridity High	0.486		
		Aridity Low	0.719	Aridity Low	0.715	Aridity Low	0.534		
		Random	0.783	Random	0.773	Random	0.594		
NLDAS	NWM	Aridity High	0.696	Aridity High	0.686	Aridity High	0.473	Aridity High	0.433
		Aridity Low	0.744	Aridity Low	0.734	Aridity Low	0.572	Aridity Low	0.593
		Random	0.863	Random	0.834	Random	0.661	Random	0.562
	All	Aridity High	0.670	Aridity High	0.652	Aridity High	0.467		
		Aridity Low	0.714	Aridity Low	0.702	Aridity Low	0.546		
		Random	0.767	Random	0.757	Random	0.648		

Table 4: Summary of median NSE scores for all model run across 531 basins. All SAC-SMA and NH runs use 10-member ensembles, and NSE scores reported for random experiments are calculated as the mean of scores from five (5) sets of ensembled runs to account for randomness in the train/test split. Figures 3, 6, 7, and 9 plot the CDFs for several of these runs and show that there is little variation in the random runs.

All LSTM and SAC-SMA models with biased train/test splits exhibit similar patterns: models tested on years characterized by low mean aridity indexes (aridity low models; dry-to-wet), outperform models tested on years characterized by high mean aridity indexes (aridity high models; wet-to-dry). Both aridity low and aridity high biased experiments show a loss of skill compared with the corresponding random (unbiased) experiments.

The leftmost plots in Figure 3 show the cumulative density functions (CDF) (over 531 basins) for random and biased LSTM models trained with static climate attributes and SAC-SMA models for both Daymet and NLDAS forcing sources. These models were trained on data extracted from all years available through the extended CAMELS dataset (1980-2014). The rightmost plots in Figure 3 shows the distribution of (x-axis) NSE scores between random and biased models.

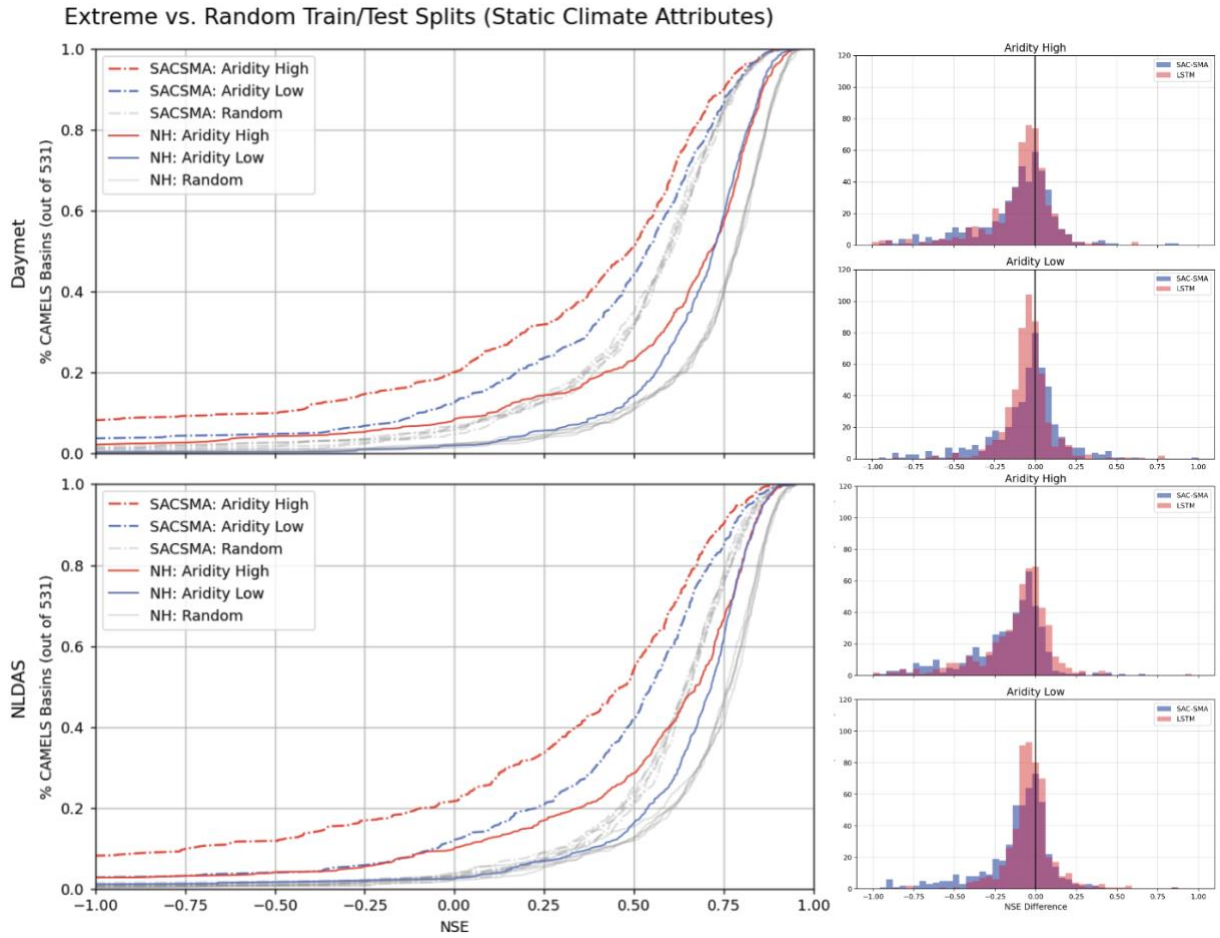

Figure 3: (Left) CDF plots for random and biased LSTM models trained with static climate attributes and SAC-SMA models for both Daymet and NLDAS forcing sources for 531 basins. (Right) Histograms showing the number of basins (y-axis) with a given difference (x-axis) in NSE score between the random and biased models in the same basin.

Table 5 reports median NSE score differences between random and biased models – we use this as a metric of predictive degradation from the random experiment benchmarks. Although Daymet and NLDAS models trained on years in the NWM date range (1995-2014) demonstrated higher overall predictive ability (visualized in Figure 3 and reported in Table 4), they were also more prone to degradation, as seen in Table 5. We will return to this table in the Multi-Model Benchmarking section below to discuss differences between different models.

		NH				SAC-SMA		NWM	
		Static		Dynamic					
		Experiment	Med. NSE Diff.	Experiment	Med. NSE Diff.	Experiment	Med. NSE Diff.	Experiment	Med. NSE Diff.
Daymet	NWM	Aridity High	-0.0869	Aridity High	-0.0916	Aridity High	-0.1401		
		Aridity Low	-0.0710	Aridity Low	-0.0732	Aridity Low	-0.0811		
	All	Aridity High	-0.0573	Aridity High	-0.0508	Aridity High	-0.0826		
		Aridity Low	-0.0425	Aridity Low	-0.0317	Aridity Low	-0.0206		
NLDAS	NWM	Aridity High	-0.1522	Aridity High	-0.1409	Aridity High	-0.1553	Aridity High	-0.0845
		Aridity Low	-0.1015	Aridity Low	-0.0797	Aridity Low	-0.0656	Aridity Low	0.0249
	All	Aridity High	-0.0700	Aridity High	-0.0860	Aridity High	-0.1445		
		Aridity Low	-0.0376	Aridity Low	-0.0308	Aridity Low	-0.0687		

Table 5: Table of the median NSE score differences between a model's random experiment and a given biased experiment for the 531 CAMELS basins. Lower values indicate more degradation. White indicates no change, while reds indicate degree of degradation.

Figure 4 shows correlations between static catchment attributes and biased model degradation (difference between NSE scores of random vs. biased models). These correlations use the LSTM model trained with Daymet forcing data with static climate attributes for all years available through the extended CAMELS dataset. Positive correlation indicates that performance tends to degrade more in basins with higher values of a given static attribute. The strongest positive correlation is with aridity and the strongest negative correlation is with maximum green vegetation fraction. In other words, basins in high aridity environments tend to degrade more than basins in low aridity environments – this is true both when the test data is more and less arid than train data. Additionally, basins with a greater fraction of precipitation falling as either high or low intensities also showed greater degradation. This is likely because water limited basins are flasher and harder to model in general, and changes in rainfall-runoff response times are basin-specific and difficult to model in a general way.

Basins with higher green vegetation fraction tend to be more robust to performance degradation than basins with low green vegetation fractions. This was likewise true for all of the vegetation indexes, including root depth. This is possibly due to the fact that vegetation can compensate for changes in water and energy availability, either by drawing from deeper soil water reserves in dry periods and transpiring more in wet periods.

Figure 5 illustrates three of these correlations for (i) aridity, (ii) maximum green vegetation fraction, and (iii) snow fraction. Although snow fraction had a lower total correlation with NSE degradation than aridity or green vegetation fraction, the performance scatterplot in Figure 5 reveals that model performance degrades in almost every basin with any appreciable snow cover. Both the timing and magnitude of melt driven runoff is heavily dependent on energy availability in a catchment, and although the LSTM is able to initiate melt-driven processes based on temperature (i.e., learns to melt stored snowpack when temperatures are above freezing; Kratzert et al., 2019a), our results apparently indicate that this depends on climatologically relevant training data.

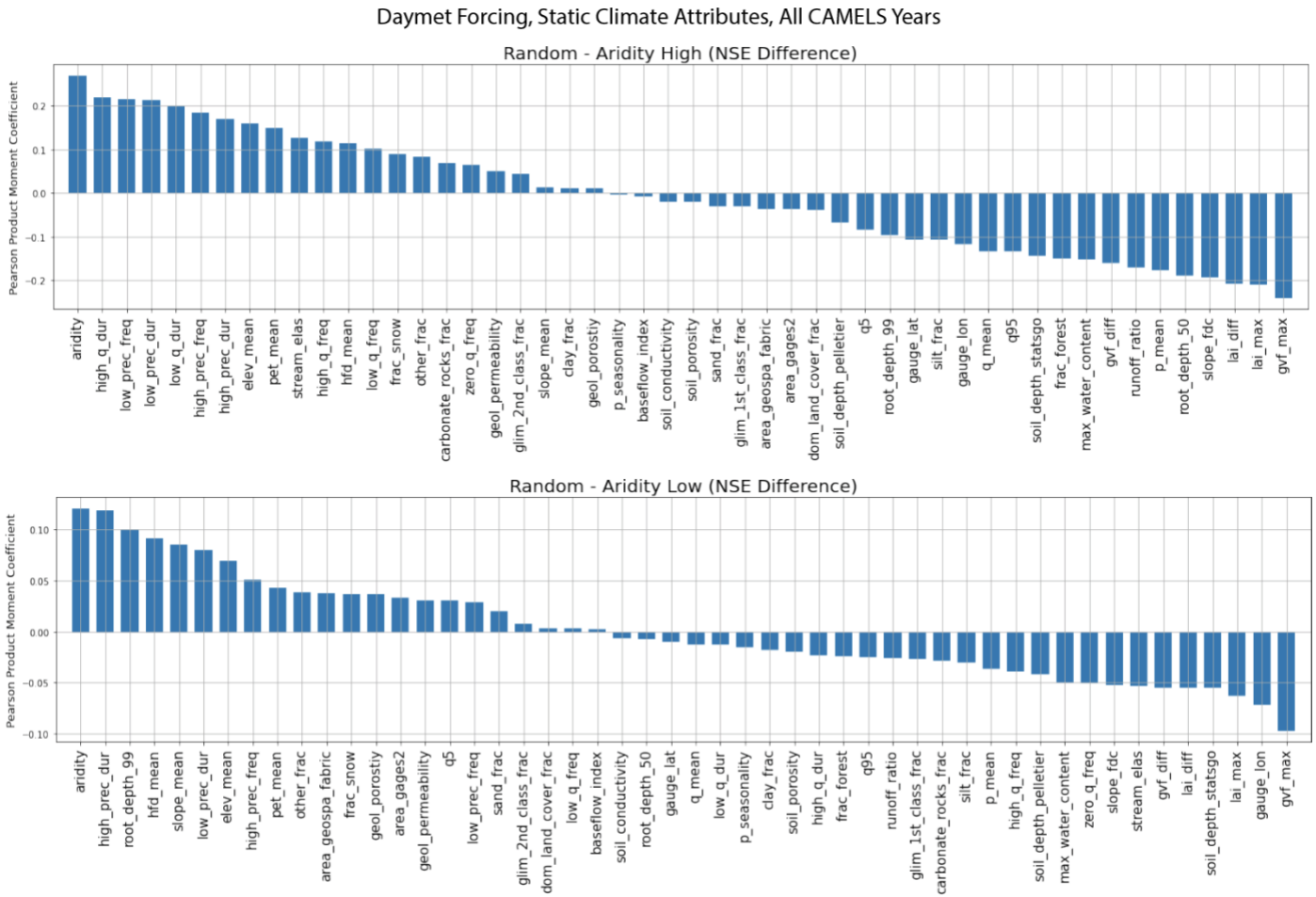

Figure 4: Visualization of the Pearson product moment correlation coefficient for the models trained with Daymet forcing data with static climate attributes for all available forcing years. This coefficient represents the strength of the linear relationship between the basins' attributes and basins' degradation from random for a given experiment with a biased train/test set. *positive correlation = more degradation as variable increases

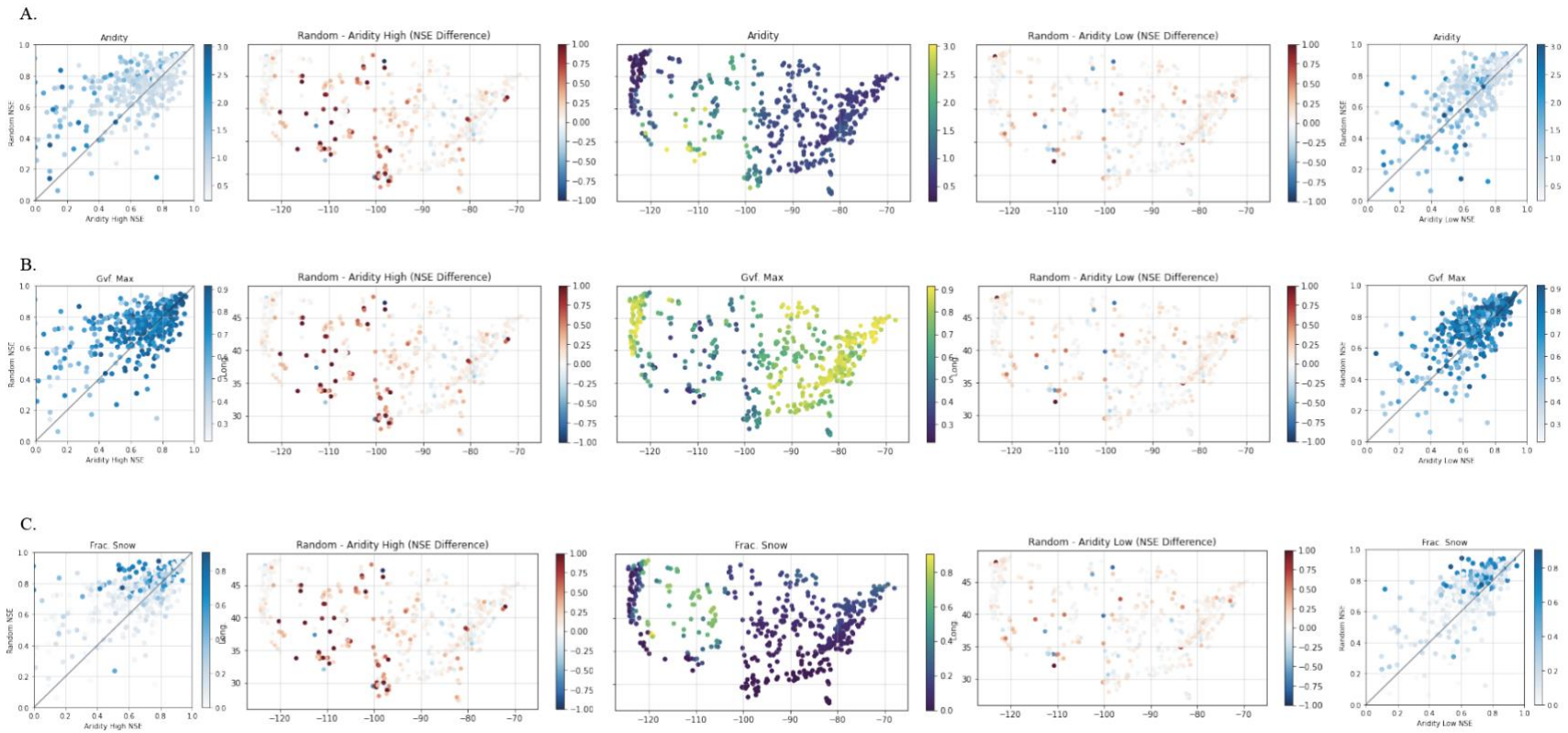

Figure 5: Scatterplots of 531 CAMELS basins that compare random vs. biased models (right = aridity high, left = aridity low) with color representing the value of a given basin attribute (top = aridity, middle = maximum green vegetation fraction, bottom = fraction of precipitation falling as snow). All results are from models with Daymet forcings with data from water-years extracted from the full extended CAMELS period. Spatial plots show the spatial distribution of the static attribute (middle) as well as the differences between random and biased NSE values. Red indicates degradation (from random to biased) and blue indicates improvement.

Multi-Model Benchmarking

SAC-SMA

The LSTM outperformed the SAC-SMA benchmark models with respect to overall predictive ability in all experiments; in fact, the LSTM's worst models, the biased aridity high models, had better predictive abilities than SAC-SMA's unbiased random models, as reported in Table 4 and visualized in Figure 6.

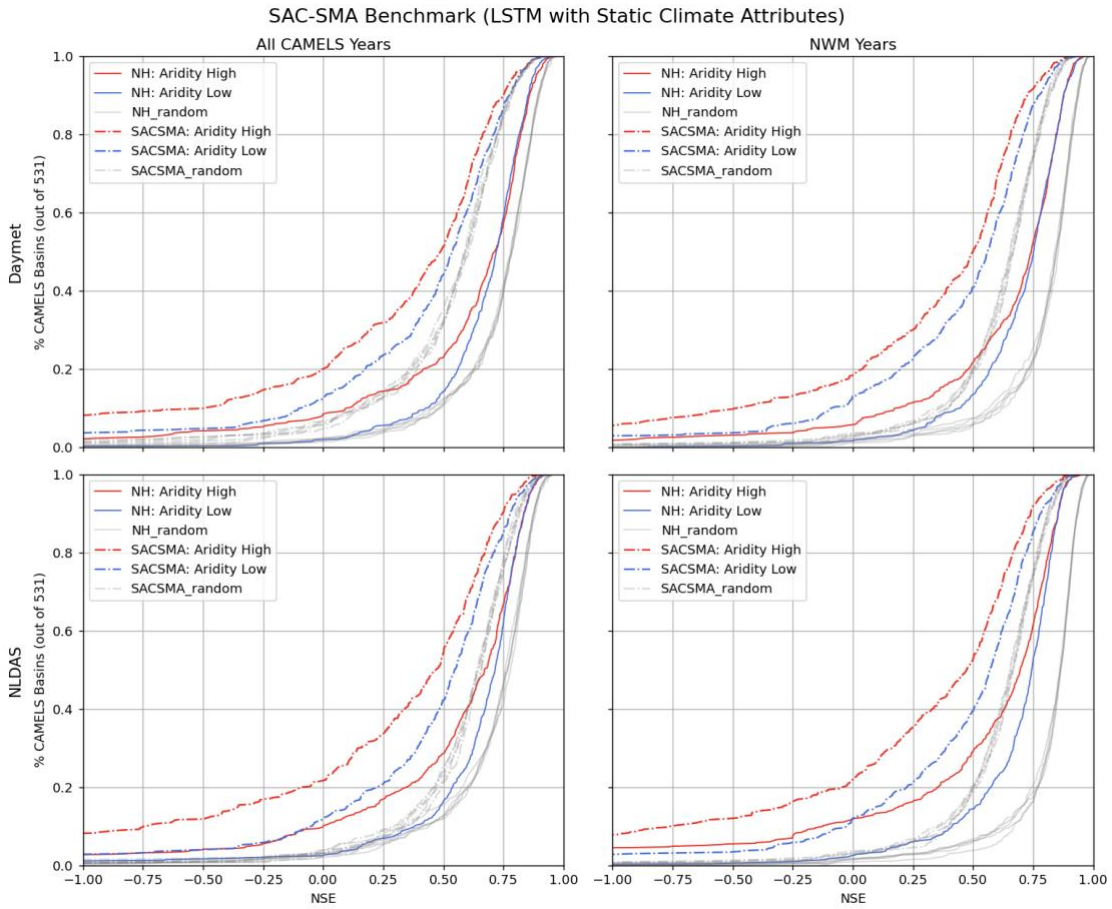

Figure 6: CDF plots for LSTM and SAC-SMA models trained on NLDAS or Daymet forcing data for 531 basins. The left column corresponds to models trained with train/test sets derived from all available CAMELS years (1980-2014) and the right column corresponds to models trained with train/test sets derived from the NWM subset years (1995-2014). The LSTM models shown here were trained with static climate attributes.

The LSTM not only showed higher predictive ability than all of the corresponding SAC-SMA benchmark models, but also demonstrated a higher degree of robustness, i.e., NSE values generally degraded less from the random experiments, as shown in Table 5. Figure 6 shows Daymet and NLDAS model performance for both the LSTM (trained with static climate attributes) and SAC-SMA models trained with (leftmost plots) data from years extracted from the extended CAMELS dataset (1980-2014) and with (rightmost plots) data from years extracted from the NWM date range (1995-2014).

NWM

The models trained and tested on data extracted from the NWM date range (1995-2014) allow us to benchmark against the NWM. Note that, as shown in Table 4, while the models trained on train/test sets derived from the NWM date range outperform models trained on train/test sets derived from all years available through the extended CAMELS dataset (1980-2014), the latter demonstrates less degradation from the random experiments (Table 5).

The left subplot in Figure 7 compares the overall performance of LSTM (trained with static attributes) and NWM models trained on NLDAS data with train/test sets derived from the NWM date range. The right subplot in Figure 7 visualizes the number of basins with a given difference in median NSE score between the random and biased models in the same basin for both the LSTM and NWM model.

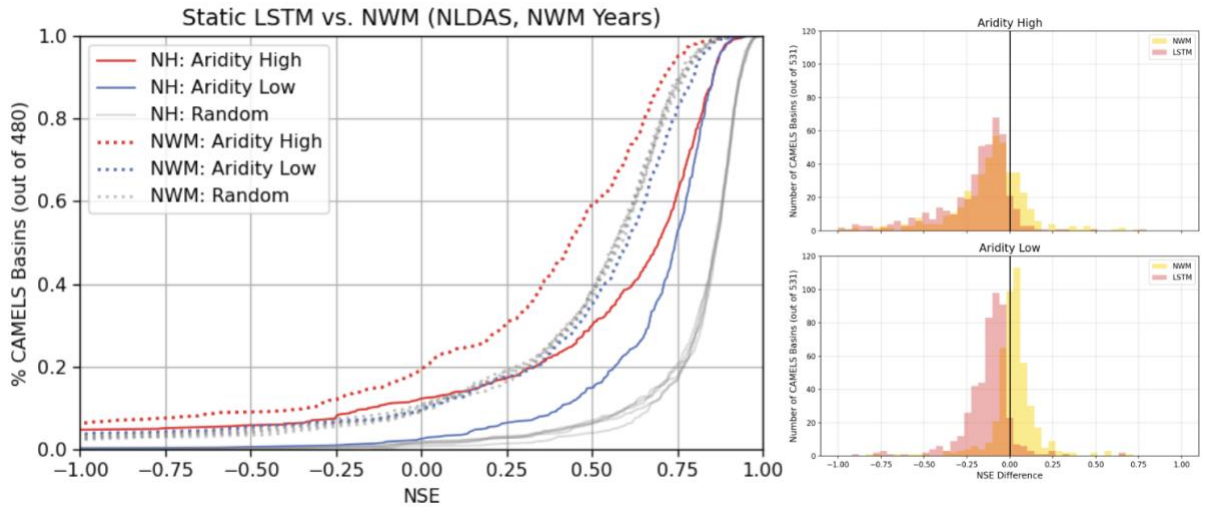

Figure 7: (Left) CDF plot for the LSTM (trained with static climate attributes) and NWM models. All models shown here were trained with NLDAS forcing data on a subset of NWM years (1995-2014). Only the 480 CAMELS basins where the NWM was calibrated (by NOAA) are used in all results shown in this figure. (Right) Histograms showing the number of basins (y-axis) with a given difference (x-axis) in median NSE score between the random and biased models.

Unlike all the LSTM and SAC-SMA model experiment sets, the NWM demonstrated a different random-biased experiment relationship. While the aridity high model (wet-to-dry) still demonstrated the most degradation from random, the aridity low model outperformed the random experiment. These results are qualitatively different from the LSTM and SAC-SMA results. This is because, in general, it is easier to model humid basins. Not only was the NWM not calibrated using a biased train/test split, the statistics reported here are from the calibration period instead of from a combination of train and test periods (the calibration period for the NWM was 2009-2013). This means that the NWM has two very strong advantages over the other two models, but still performs significantly worse in almost all basins. Figure 8 plots the spatial distribution of biased LSTM vs NWM results.

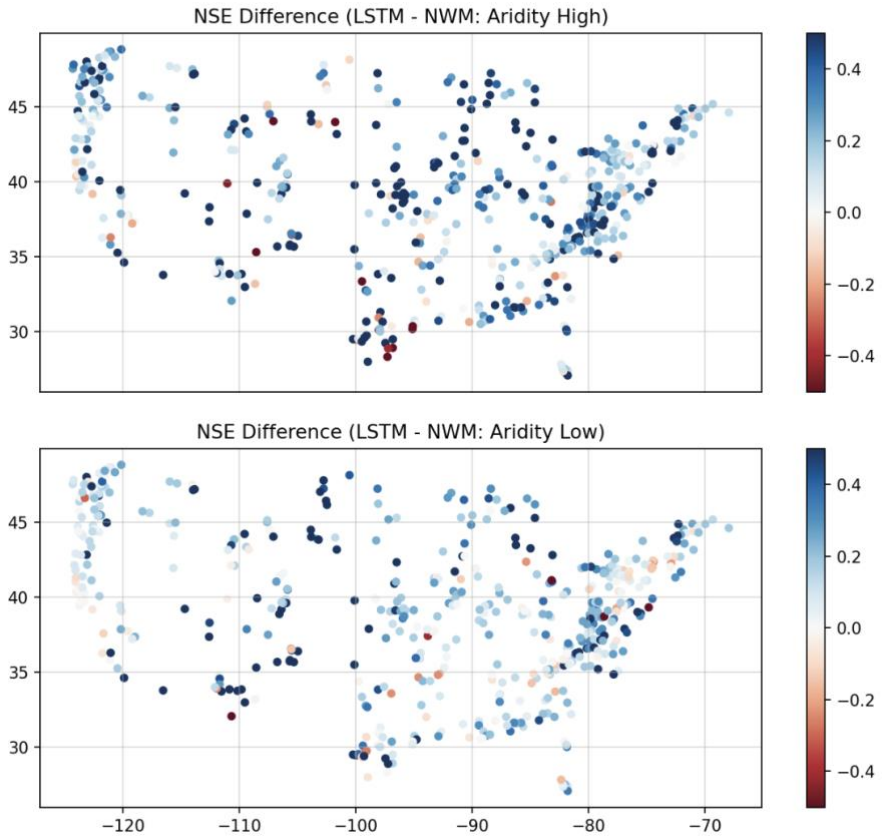

Figure 8: Spatial plot of the per-basin differences in NSE scores between the LSTM and NWM benchmark experiments. These differences are calculated from the NSE scores for the LSTM and NWM models trained with NLDAS data on train/test sets from the NWM date range.

Static vs. Dynamic Climate Inputs

Figure 8 shows CDF plots of LSTM models trained and testd with static vs. dynamic climate indexes. The addition of daily dynamic climate attributes provided no value in streamflow prediction, even under climate-baised train/test splits. In fact, all of the LSTMs trained with dynamic climate attributes were outperformed by their counterparts trained on static climate attributes. These results are similar to what was reported by Nearing et al. (2019) for unbiased train/test splits.

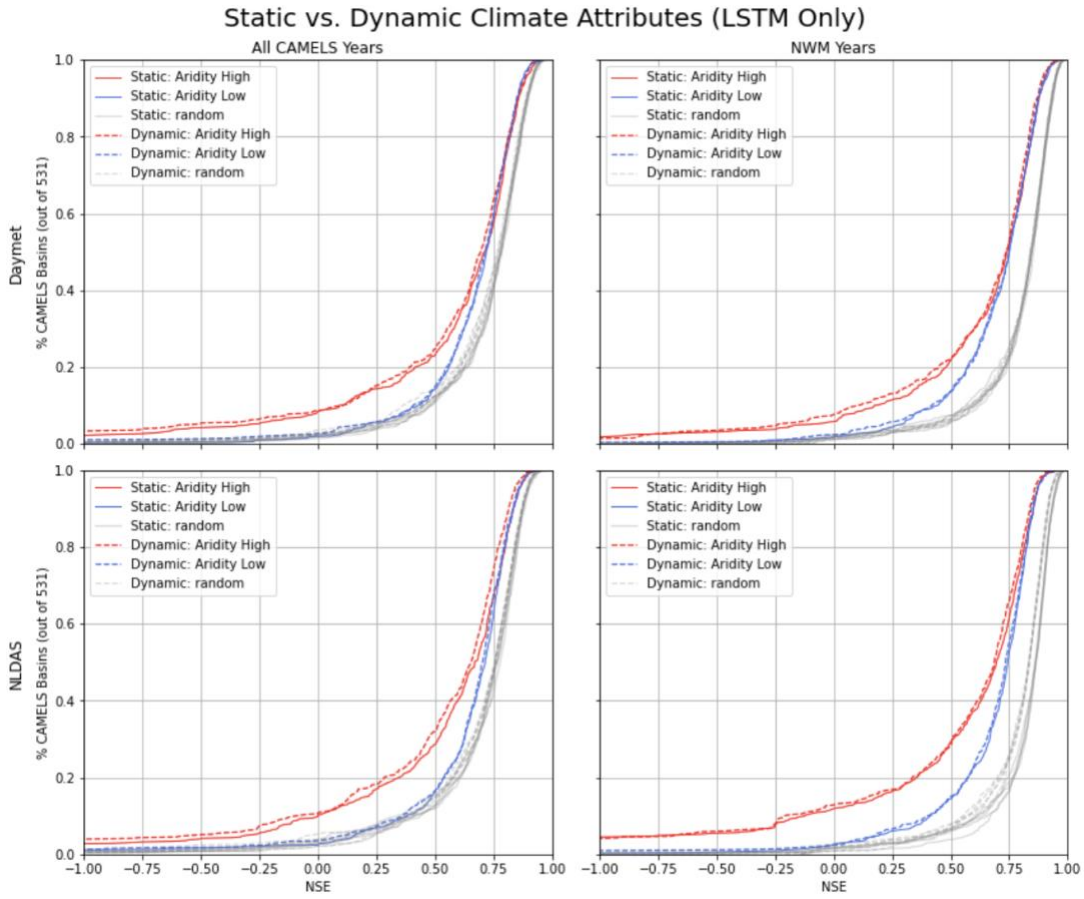

Figure 9: CDF plots of LSTM models trained with static or dynamic climate attributes.

DISCUSSION

The main findings of this work are as follows:

1. The accuracy of DL and calibrated conceptual models degrade under climate nonstationarity. Results show that model performance under climatologically-biased train/test data (simulating climate nonstationarity) degrades for both deep learning (LSTM) models and the calibrated conceptual benchmark (SAC-SMA). Unbiased random train/test splits outperform biased train/test splits, and the performance degrades more when catchments move from wetter to drier regimes.
2. LSTMs are more robust than the calibrated conceptual benchmark under climate nonstationarity. LSTMs perform better in all cases and in almost all basins than SAC-SMA, even when the LSTM was trained on climatologically biased data and SAC-SMA was calibrated on unbiased data.
3. We did not test the process-based model under biased calibration/validation data splits, however the process-based model performed worse than both other benchmarks, even though the test statistics for the process-based model included a combination of training and test period data.
4. LSTMs do not benefit from daily dynamic climate attributes as inputs, even under climatologically biased train/test splits.

We feel that the fourth point deserves further study. Nearing et al. (2019) reported that the LSTM was able to use dynamic climate indexes to learn dynamic embedding representations of basin similarity and showed that it produced different hydrographs with the same forcing inputs under different climate scenarios. This indicates that the model is learning to use dynamic climate indexes, however we have not found a way to translate this to improved model skill. Daily dynamic climate attribute inputs provided no overall benefit in the LSTM models' predictive ability.

CAMELS basins were chosen for their lack of human interference, and therefore are less prone to significant and sudden changes in the local hydrological cycle. While this was a design choice made by the origins CAMELS authors (Newman et al 2015), it is perhaps less well-suited for assessing the benefit of dynamic inputs. We would expect the effect of this basin selection on climate-related nonstationarity to be small, however it is important to understand that there is this particular bias in the data.

One result worth noting (that was not explained above) is that models calibrated with train/test sets derived from the NWM date range (1995-2014) have better predictive abilities than the models calibrated on train/test sets derived from the full available date range (1980-2014). The NWM date range divides the water-years available for training and testing in half, effectively reducing the total climatological change in a basin. In other words, water-years closer in time are more similar to years farther away in time, and therefore require less extrapolation and therefore less error in the model.

CODE AND DATA AVAILABILITY

Source code for this work was written by Logan Qualls and is available through GitHub (https://github.com/loganmqualls/NeuralHydrology-Climate-Experiments).

The CAMELS dataset, including all forcing and static attribute datasets, is available through NCAR (https://ral.ucar.edu/solutions/products/camels). The extended dataset for NLDAS is available through Hydroshare (https://www.hydroshare.org/resource/0a68bfd7ddf642a8be9041d60f40868c/), alongside an extended dataset for Maurer (not used in this work; https://www.hydroshare.org/resource/17c896843cf940339c3c3496d0c1c077/)

The NeuralHydrology modeling package can be found on GitHub (https://github.com/neuralhydrology/neuralhydrology) Documentation for this codebased is located at (https://neuralhydrology.readthedocs.io/en/latest/index.html).

Code to calibrate SAC-SMA-Snow17 using the CAMELS dataset with a Python interface comes from (https://github.com/Upstream-Tech/SACSMA-SNOW17). This code uses the SpotPy optimization package, which can be downloaded through the SpotPy project website (https://pypi.org/project/spotpy/).

Source code for processing the NWM v2 Reanalysis as a CAMELS benchmark dataset came from (https://github.com/jmframe/nwm-post-processing-with-lstm).

REFERENCES

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrol. Earth Syst. Sci.*, 21, 5293-5313, doi:10.5194/hess-21-5293-2017

Anderson, E. A. (1973). National Weather Service river forecast system: Snow accumulation and ablation model / Eric A. Anderson. Washington, D.C.: U.S. Dept. of Commerce, National Oceanic and Atmospheric Administration, National Weather Service. http://archive.org/details/nationalweathers00ande

Beven, K. J. (2000). Uniqueness of place and process representations in hydrological modelling. *Hydrology and Earth System Sciences*, *4*(2), 203–213. https://doi.org/10.5194/hess-4-203-2000

Cameron, D., Kneale, P., and See, L.: An evaluation of a traditional and a neural net modelling approach to flood forecasting for an upland 365 catchment, Hydrological Processes, 16, 1033–1046, https://doi.org/10.1002/hyp.317, 2002.

Frame, J., Kratzert, F., Klotz, D., Gauch, M., Shelev, G., Gilon, O., Qualls, L. M., Gupta, H. V., & Nearing, G. S. (2021a). Deep learning rainfall-runoff predictions of extreme events. *Hydrology and Earth System Sciences Discussions*, 1–20. https://doi.org/10.5194/hess-2021-423

Frame, J. M., Kratzert, F., Raney II, A., Rahman, M., Salas, F. R., & Nearing, G. S. (2021b). 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

Gauch, M., Mai, J., & Lin, J. (2021). The proper care and feeding of CAMELS: How limited training data affects streamflow prediction. *Environmental Modelling & Software*, *135*, 104926. https://doi.org/10.1016/j.envsoft.2020.104926

Gaume, E. and Gosset, R.: Over-parameterisation, a major obstacle to the use of artificial neural networks in hydrology?, Hydrology and Earth System Sciences, 7, 693–706, https://doi.org/10.5194/hess-7-693-2003, 2003.

Houska, T., Kraft, P., Chamorro-Chavez, A., & Breuer, L. (2019). *SPOTPY: A Python library for the calibration, sensitivity- and uncertainty analysis of Earth System Models*. 7878.

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., Arheimer, B., Blume, T., Clark, M. P., Ehret, U., Fenicia, F., Freer, J. E., Gelfan, A., Gupta, H. V., Hughes, D. A., Hut, R. W., Montanari, A., Pande, S., Tetzlaff, D., … Cudennec, C. (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

Hsu, K., Gupta, H. V., & Sorooshian, S. (1995). Artificial Neural Network Modeling of the Rainfall-Runoff Process. *Water Resources Research*, *31*(10), 2517–2530. https://doi.org/10.1029/95WR01955

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, *42*(3). https://doi.org/10.1029/2005WR004362

Klotz, D., Kratzert, F., Gauch, M., Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., & Nearing, G. (2021). *Uncertainty Estimation with Deep Learning for Rainfall–Runoff Modelling*. https://doi.org/10.5194/hess-2021-154

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, *22*(11), 6005–6022. https://doi.org/10.5194/hess-22-6005-2018

Kratzert, F., Herrnegger, M., Klotz, D., Hochreiter, S., & Klambauer, G. (2019a). NeuralHydrology—Interpreting LSTMs in Hydrology. *ArXiv:1903.07903 [Physics, Stat]*, *11700*, 347–362. https://doi.org/10.1007/978-3-030-28954-6_19

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019b). 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. (2019c). Towards Learning Universal, Regional, and Local Hydrological Behaviors via Machine-Learning Applied to Large-Sample Datasets. *ArXiv:1907.08456 [Cs, Stat]*. http://arxiv.org/abs/1907.08456

Kratzert, F., Gauch, M., Nearing, G., & Klotz, D. (2022). NeuralHydrology—A Python library for Deep Learning research in hydrology. *Journal of Open Source Software*, *7*(71), 4050. https://doi.org/10.21105/joss.04050

Maurer, E., Wood, A., Adam, J., Lettenmaier, D., & Nijssen, B. (2002). A Long-Term Hydrologically-Based Data Set of Land Surface Fluxes and States for the Conterminous United States. *Journal of Climate*, *15*, 3237. https://doi.org/10.1175/1520-0442(2002)015<3237:ALTHBD>2.0.CO;2

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., & Stouffer, R. J. (2008). Stationarity Is Dead: Whither Water Management? *Science*, *319*(5863), 573–574. https://doi.org/10.1126/science.1151915

Nearing, G., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J., Prieto, C., & Gupta, H. (2020). *What Role Does Hydrological Science Play in the Age of Machine Learning?* https://eartharxiv.org/repository/view/422/

Nearing, G. S., Pelissier, C. S., Kratzert, F., Klotz, D., Gupta, H. V., Frame, J. M., & Sampson, A. K. (n.d.). *Physically Informed Machine Learning for Hydrological Modeling Under Climate Nonstationarity*. 5.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., & Duan, Q. (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

Salas, F. R., Somos-Valenzuela, M. A., Dugger, A., Maidment, D. R., Gochis, D. J., David, C. H., Yu, W., Ding, D., Clark, E. P., & Noman, N. (2018). Towards Real-Time Continental Scale Streamflow Simulation in Continuous and Discrete Space. *JAWRA Journal of the American Water Resources Association*, *54*(1), 7–27. https://doi.org/10.1111/1752-1688.12586

Sellars, S. L. (2018). "Grand Challenges" in Big Data and the Earth Sciences. *Bulletin of the American Meteorological Society*, *99*(6), ES95–ES98. https://doi.org/10.1175/BAMS-D-17-0304.1

Thornton, P. E., Running, S. W., & White, M. A. (1997). Generating surfaces of daily meteorological variables over large regions of complex terrain. *Journal of Hydrology*, *190*(3–4), 214–251. https://doi.org/10.1016/S0022-1694(96)03128-9

Tolson, B. A., & Shoemaker, C. A. (2006). *The Dynamically Dimensioned Search (DDS) Algorithm as a Robust Optimization Tool in Hydrologic Modeling*. *2006*, H41I-07.

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Lettenmaier, D., Koren, V., Duan, Q., Mo, K., Fan, Y., & Mocko, D. (2012). Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *Journal of Geophysical Research: Atmospheres*, *117*(D3). https://doi.org/10.1029/2011JD016048

APPENDIX 1

Median performance metrics for all ensembled model runs are given in Table A1. Descriptions of the metrics in this table are given in Table 3.

Experimental Design					Metrics							
Model	Forcing	Climate Attributes	Year Range	Experiment	alpha-NSE	beta-KGE	beta-NSE	KGE	MSE	NSE	RMSE	Pearson r
NH LSTM	Daymet	Static	NWM	Aridity High	0.878733	1.022305	0.015431	0.746649	0.443082	0.741288	0.665644	0.879823
				Aridity Low	0.856995	0.937718	-0.03082	0.75456	2.082557	0.750663	1.443107	0.880899
				Random	0.917127	0.990028	-0.004023	0.855282	0.797778	0.845708	0.893057	0.924292
			All	Aridity High	0.856041	1.044905	0.026248	0.680178	0.432945	0.710859	0.657986	0.87075
				Aridity Low	0.830512	0.944697	-0.032531	0.728444	2.867281	0.71887	1.693305	0.870274
				Random	0.891557	0.988062	-0.005727	0.79685	1.062771	0.78263	1.030707	0.897106
		Dynamic	NWM	Aridity High	0.868716	0.987979	-0.004954	0.727424	0.494415	0.736661	0.703147	0.876347
				Aridity Low	0.870547	0.973161	-0.012992	0.758301	2.088403	0.748234	1.445131	0.876207
				Random	0.920788	0.994517	-0.001912	0.853988	0.773893	0.84424	0.879584	0.924546
			All	Aridity High	0.854783	1.012186	0.007774	0.676715	0.446159	0.693863	0.667951	0.864374
				Aridity Low	0.851551	0.991949	-0.00366	0.734125	2.861269	0.715365	1.691529	0.86808
				Random	0.893626	0.988217	-0.005372	0.781075	1.128091	0.773443	1.061908	0.893876
	NLDAS	Static	NWM	Aridity High	0.839508	0.995585	0.004634	0.690095	0.521884	0.695652	0.722415	0.86257
				Aridity Low	0.852301	0.976854	-0.008603	0.750757	2.288009	0.743618	1.512617	0.881981
				Random	0.9171	0.994189	-0.002159	0.865654	0.713633	0.862731	0.844622	0.932401
			All	Aridity High	0.812359	0.997395	0.003128	0.666491	0.510666	0.669943	0.714609	0.844281
				Aridity Low	0.848864	0.978076	-0.01274	0.729693	3.12878	0.714495	1.768836	0.864039
				Random	0.888286	0.98444	-0.006865	0.779183	1.158393	0.767198	1.076099	0.889655
		Dynamic	NWM	Aridity High	0.834456	0.962952	-0.01509	0.680836	0.569767	0.685834	0.754829	0.851913
				Aridity Low	0.858816	0.993235	-0.001951	0.739194	2.397472	0.734311	1.548377	0.876369
				Random	0.908706	0.993338	-0.002495	0.838795	0.841304	0.834425	0.916966	0.919115
			All	Aridity High	0.812208	0.953166	-0.026251	0.639179	0.533852	0.652347	0.730651	0.839798
				Aridity Low	0.840401	1.024379	0.014339	0.723598	3.171721	0.702389	1.780933	0.855486
				Random	0.89126	0.993757	-0.0017	0.769088	1.186219	0.757388	1.088953	0.884685
SAC-SMA	Daymet		NWM	Aridity High	0.818521	1.316392	0.166807	0.435189	0.97274	0.497586	0.986276	0.784501
				Aridity Low	0.791821	1.143222	0.084958	0.558894	3.976123	0.558634	1.994022	0.80219
				Random	0.809224	1.170463	0.092949	0.59549	1.773378	0.660777	1.331648	0.841864
			All	Aridity High	0.797808	1.292733	0.163788	0.398831	0.822808	0.485594	0.907088	0.781154
				Aridity Low	0.764989	1.121742	0.074366	0.554993	5.10249	0.534149	2.258869	0.795654
				Random	0.809232	1.14624	0.081949	0.579378	2.003844	0.59443	1.415194	0.812554
	NLDAS		NWM	Aridity High	0.838278	1.339763	0.202517	0.379773	1.035809	0.472769	1.017747	0.788658
				Aridity Low	0.776111	1.126517	0.073962	0.554618	3.798215	0.572229	1.948901	0.807787
				Random	0.809224	1.170463	0.092949	0.59549	1.773378	0.660777	1.331648	0.841864
			All	Aridity High	0.842018	1.318418	0.203882	0.389674	0.879295	0.466817	0.937707	0.783524
				Aridity Low	0.756517	1.091747	0.055769	0.567145	4.870168	0.545659	2.206846	0.797236
				Random	0.782333	1.127061	0.073829	0.596884	1.769856	0.647629	1.330065	0.836486
NWM	NLDAS		NWM	Aridity High	0.81815	0.880851	-0.054392	0.507279	1.047335	0.432865	1.023394	0.737371
				Aridity Low	0.874872	0.950034	-0.021599	0.631153	3.667469	0.593248	1.915064	0.81443
				Random	0.847215	0.921538	-0.034607	0.606132	2.288901	0.562139	1.512785	0.794495

Table A1: Median performance metrics for all model runs for 531 basins.