



RESEARCH ARTICLE

WILEY

On strictly enforced mass conservation constraints for modelling the Rainfall-Runoff process

Jonathan M. Frame¹  | Frederik Kratzert² | Hoshin V. Gupta³ | Paul Ullrich⁴  | Grey S. Nearing⁵

¹Department of Geological Sciences,
University of Alabama, Tuscaloosa,
Alabama, USA

²Google Research, Vienna, Austria

³Department of Hydrology and Water
Resources, The University of Arizona, Tucson,
Arizona, USA

⁴Department of Land, Air & Water Resources,
University of California Davis, Davis,
California, USA

⁵Google Research, Mountain View,
California, USA

Correspondence

Jonathan M. Frame, Department of Geological
Sciences, University of Alabama, Tuscaloosa,
AL, USA.

Email: jmframe@crimson.ua.edu

Funding information

NOAA Cooperative Agreement, Grant/Award
Number: NA19NES4320002; NASA Terrestrial
Hydrology Program, Grant/Award Number:
80NSSC18K0982

Abstract

It has been proposed that conservation laws might not be beneficial for accurate hydrological modelling due to errors in input (precipitation) and target (streamflow) data (particularly at the event time scale), and this might explain why deep learning models (which are not based on enforcing closure) can out-perform catchment-scale conceptual and process-based models at predicting streamflow. We test this hypothesis with two forcing datasets that disagree in total, long-term precipitation. We analyse the roll of strictly enforced mass conservation for matching a long-term mass balance between precipitation input and streamflow output using physics-informed (mass conserving) machine learning and find that: (1) enforcing closure in the rainfall-runoff mass balance does appear to harm the overall skill of hydrological models; (2) deep learning models learn to account for spatiotemporally variable biases in data (3) however this ‘closure’ effect accounts for only a small fraction of the difference in predictive skill between deep learning and conceptual models.

KEYWORDS

CAMELS, deep learning, large sample hydrology, LSTM, mass conservation, physics-informed machine learning, rainfall-runoff, water balance

1 | INTRODUCTION

Deep learning (DL) models are becoming the standard benchmark for predictive hydrologic modelling in the current literature because of their high accuracy relative to conceptual models (Nearing, Kratzert, et al., 2020), as well as their ability to extrapolate to new locations (Kratzert et al., 2018; Kratzert, Klotz, Hernegger, et al., 2019) and extreme events (Frame et al., 2022). There has been a recent push to combine deep learning with physical theory to (i) gain better process understanding; and (ii) improve predictive accuracy, especially under out-of-sample conditions (Jia et al., 2021; Reichstein et al., 2019; Shen et al., 2021; Willard et al., 2021). There have been several recent attempts to build hybrid DL models (sometimes referred to ‘physics-informed’ or ‘theory-guided’, e.g., Bennett & Nijssen, 2021; Daw et al., 2020; Hoedt et al., 2021; Jiang et al., 2020; Karniadakis et al., 2021; Nearing, Research, et al., 2020; Pelissier et al., 2019; Tsai

et al., 2021; Xie et al., 2021; Zhao et al., 2019). We therefore think it is important to take a step back and explore if (and what) basic components of physical theory might actually be beneficial for hydrologic prediction. The use of a single-constraint model like the MC-LSTM helps us understand hydrological processes through testing hypotheses of individual relevant processes (Nearing, Research, et al., 2020).

In this paper, we test one hypothesis in particular: we use physics-informed machine learning to explore the longstanding assumption that mass conservation should be the foundation of hydrological models. The first physical law introduced formally by Chow et al. (1988, equation 1.3.5) (a standard introductory hydrology textbook) is:

$$dS/dt = I(t) - O(t) \quad (1)$$

where the change of a system's mass storage (S) with respect to time (t) is equal to total mass input (I) minus total mass output (O). This is

the first physical constraint placed on the transfer function between inputs and output of a hydrological system (i.e., Chow et al., 1988, equation 1.3.1).

While conservation laws are considered to be a fundamental truth about (classical scale) systems in our physical world, it is not necessarily the case that this makes them a proper or useful foundation for either understanding or modelling watershed systems. This distinction is motivated by Beven (2020), who proposed that the closure problem might explain the *poor* performance of conceptual and physically-based (PB) hydrology models relative to DL:

given the epistemic uncertainties in water and energy balances, then this [conservation constraints] might not necessarily be advantageous in obtaining better DL predictions if, for example, the observational data do not themselves provide consistent mass and energy balance closure.

In other words, conceptual and PB models typically demand a degree of closure that may not necessarily be achievable given sparse and error-prone observation data, and (Beven hypothesized that) the superior performance of DL might be due to its ability to learn and account for consistent error structures present in the input-output data. In practice, PB models sometimes account for error prone data with pre- and post-processing, as well as data assimilation, however pre- and post-processing is not necessary when using DL, as these steps can be learned directly from training data (Frame et al., 2021).

The proposal explains poor rainfall-runoff model calibration and performance as being a consequence of so-called ‘disinformation’ in data (e.g., Beven et al., 2008; Beven & Westerberg, 2011; Sivapalan et al., 2003). In addition to the observational uncertainty present in data used for driving and evaluating models, there is also uncertainty regarding what actually constitute the true physical inputs and losses from a hydrologic system—for example, mass contributions to the system through natural springs and anthropogenic water resources can come from outside of the watershed ‘boundary’ and are not often directly observable or represented in the available data set. Beven’s hypothesis is that these types of effects might explain the relative accuracy of DL streamflow models, due to their not being constrained to conserve mass.

In this paper, we place a bound on this ‘closure’ effect (i.e., an upper bound on the information loss due to enforcing closure over error-prone data), and show two things:

- DL is able to learn and account for systematic (but spatiotemporally dynamic) errors in data, and
- the closure effect does *not* explain the majority of the performance gap between PB models and DL models of streamflow.

The long short-term memory (LSTM) neural network was chosen as the deep learning architecture for this study because (1) it is the best performing deep learning model for the rainfall-runoff process, and (2) because we have a directly comparable mass conserving (MC-

LSTM), and non-mass conserving version available. We test the ability and performance of the mass conserving MC-LSTM, a DL model with an architecture designed to strictly enforce mass conservation at every timestep, to performing its namesake task (mass conservation), by assessing the long-term bias of predicted runoff (Chawanda et al., 2020) in a large-sample dataset (Gupta et al., 2014). The long-term analysis of mass balance is necessary to capture hydrologic processes that span multiple events (e.g., interflow) and even multiple years (e.g., snowpack). To be clear, we are *not* questioning whether hydrologic processes in the real world are governed by the physical concept of mass conservation. What we are questioning is whether a testable, scale-relevant theory of watersheds should be based on this principle. Alternatively, it is possible that no successful scale-relevant theory of watersheds has been developed to-date because the fundamental conceptual basis for a ‘watershed’ is itself incorrect; the typical ‘fixed catchment control-volume’ represented by our watershed delineations with closed internal states cannot represent mesoscale storm-system scales and groundwater aquifer scales.

2 | DATA AND METHODS

We designed an experiment to test the hypothesis proposed by Beven (2020) that the lack of mass conservation in DL rainfall-runoff models explains the difference in skill relative to PB models that are constrained by closure. The basic experiment is as follows. We use two meteorological data sets with a large, nonlinear, and location-specific disagreement in long-term precipitation totals, assuming that at least one of these datasets is biased, to benchmark three models: (i) a standard PB model calibrated per-basin; (ii) a standard DL model trained regionally (over all basins, not per-basin), and (iii) a physics-informed DL model that is constrained to enforce mass conservation trained regionally (in the same way as the standard DL model). Our goal is to understand how much of the difference in skill between the PB and DL models can be accounted for by forcing closure on biased data.

2.1 | Data

Several recent modelling studies used open community data sets and consistent training/test procedures that allow for results to be directly comparable (Frame et al., 2021; Gauch, Kratzert, et al., 2021; Gauch, Mai, & Lin, 2021; Klotz et al., 2021; Kratzert et al., 2021; Kratzert, Klotz, Shalev, et al., 2019; Newman et al., 2017). We continue that practice here. Specifically, we used the Catchment Attributes and Meteorological Large Sample (CAMELS) data set curated by the US National Center for Atmospheric Research (NCAR; Addor et al., 2017; Newman et al., 2015). The CAMELS data set consists of daily meteorological and discharge data from 671 catchments in CONUS ranging in size from 4 to 25,000 km² that have largely natural flows and long streamflow gauge records (1980–2008). Newman et al. developed CAMELS as a data set for community model benchmarking and by

excluding basins with (i) large discrepancies between different methods of calculating catchment area; and (ii) areas larger than 2000 km². This results in the large-sample (Gupta et al., 2014) data set with 531 basins that has been used by all of the benchmarking studies cited above. In the current study, we had to omit one of these 531 basins due to a data constraint that will be explained below in Section 2.3.

CAMELS includes daily discharge data from the USGS Water Information System, which are used as training and evaluation targets. CAMELS also includes multiple daily meteorological forcing data products (Daymet, NLDAS, Maurer) that are used as model inputs, shown in Table 1. CAMELS also includes several static catchment attributes related to soils, climate, vegetation, topography, and geology (Addor et al., 2017) that are used as input features to the DL models. We used the same input features (meteorological forcings and static catchment attributes) that are listed in Table 1 by Kratzert, Klotz, Shalev, et al. (2019).

We used Daymet and NLDAS for this project because the disagree in their total precipitation amounts, as shown in Figure 1. The figure shows that the western basins are scattered about the 1 to 1 line, even though the large magnitude western basins are largely heteroscedastic, but that the basins from about longitude -100 to -70 tend to favour Daymet as the magnitude increases. This systematic difference must be either a positive Daymet bias or a negative

NLDAS bias in long-term precipitation totals, and this allows us to test hypotheses about modelling behaviour when the input data are systematically biased.

This bias can be seen clearly in some of our results presented in Section 3.2. In that section, we describe a regional analysis of the total cumulative streamflow bias grouped for different regions of CONUS. The regions were delineated according the United States Geological Survey (USGS) Water Resources Regions outlined in Water-Supply Paper 2294 (USGS, 1987). This includes 18 distinct regions, but only 17 of which have enough CAMELS basins for meaningful statistics (leaving out Souris-Red-Rainy, hydrologic unit code 09).

2.2 | Models

2.2.1 | Models inspired by physical concepts

The conceptual model that we used as a benchmark was the Sacramento Soil Moisture Accounting model (SAC-SMA) with SNOW-17 and a unit hydrograph routing function. This is the model used by (Newman et al., 2017) as a basis for standardized benchmarking with the CAMELS data set, however we re-calibrated SAC-SMA to be consistent with our training/test splits. We used the Python-based SAC-SMA code and calibration package developed by Nearing, Sampson, et al. (2020), which uses the SpotPy calibration library (Houska et al., 2019). We use the Dynamically Dimensioned Search algorithm with 10 000 model runs. SAC-SMA was calibrated separately at each of the 531 CAMELS basins using the three train/test splits outlined in Section 2.2.3, and get results comparable to Newman et al. (2017).

We also benchmarked the U.S. National Water Model (NWM) using the NOAA National Water Model CONUS Retrospective Dataset; <https://registry.opendata.aws/nwm-archive/>, accessed December 2021). The NWM is based on physics-inspired equations, but it has been argued that these types of models are still conceptual in nature, but applied to the grid scale (Beven, 1989), so we refer to non-DL models as conceptual. We present the results of the NWM benchmarks in Appendix, rather than the main body of this paper because (i) the NWM is only available for NLDAS forcing; and (ii) we are not able to calibrate the NWM to match our other models, so the NWM results are not directly comparable. A complete description of the NWM is provided in Appendix along with a complete set of figures.

2.2.2 | Deep learning models

The Long Short-Term Memory (LSTM) network is the current state-of-the-art model for predicting streamflow at the watershed scale. The LSTM is a recurrent neural network with an explicit state space, and explicit controls on input-state and state-output relationships, as well as explicit controls on memory timescales, which makes it suitable for at least many dynamical systems applications. The LSTM does not enforce conservation laws, which means that there is potential for predicted runoff to violate Equation (1).

TABLE 1 Forcing products from the CAMELS dataset

Forcing product	Description	Citation
NLDAS	North American Land Data Assimilation System. Spatial resolution is 1/8th-degree, and the temporal resolution is hourly. The data span 1979 to present. Data can be downloaded in their native GRIB format, but CAMELS provides basin averages. This product is oriented towards land/hydrology modelling. The non-precipitation land-surface forcing fields are derived from the analysis fields of a North American Regional Reanalysis (NARR). Surface pressure, longwave radiation, air temperature and specific humidity are adjusted vertically to account for terrain height.	Xia et al. (2012)
Daymet	Daily Surface Weather Data for North America. Spatial resolution is 1-km \times 1-km in Lambert Conformal Conic projection. The data span 1980 through 2015. Data can be downloaded in their native netCDF file formats, but CAMELS provides basin averages. Several of the variables are derived from selected meteorological station data by interpolation and extrapolation algorithms. Data are assembled by parameter and year with each yearly file containing a time dimension of 365 days.	Thornton et al. (2014)

Comparing precipitation total (sum) between forcing products

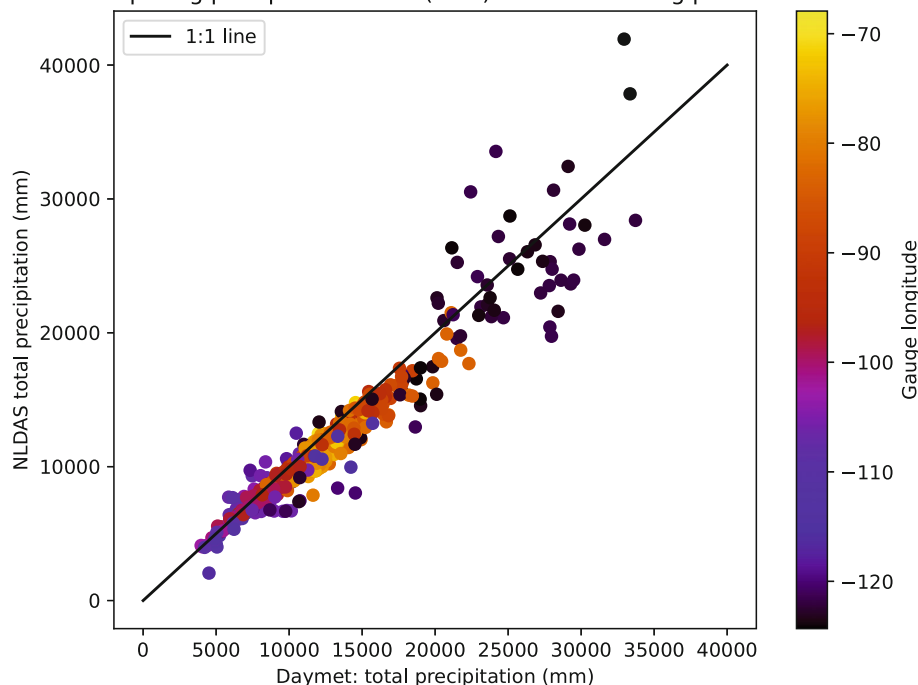


FIGURE 1 Comparison of the total precipitation from NLDAS and Daymet for each of the 530 basins, colour-coded by the longitudinal coordinate of the stream gauge

TABLE 2 Forcing variables for LSTM and MC-LSTM

Forcing variable	Role in MC-LSTM (unit)
Average daily precipitation	Mass conserving (mm)
Daily minimum air temperature	Auxiliary
Daily maximum air temperature	Auxiliary
Solar radiation	Auxiliary
Vapour pressure	Auxiliary

The Mass-Conserving LSTM (MC-LSTM) is also a recurrent neural network with an explicit state space and explicit input-state and state-output relationships. Both the LSTM and MC-LSTM use the same forcing variables, but the MC-LSTM distinguishes between mass inputs (with a specific unit of mass to conserve through the model) and auxiliary (no conservation enforced) forcing inputs, shown in Table 2.

The internal calculations of the MC-LSTM ensure mass-conservation, at every timestep, between any number of inputs (here precipitation) and outputs (here streamflow). In reality, precipitation and streamflow are not the only fluxes of water into or out of a catchment. The MC-LSTM does not account for unobserved water sources other than precipitation, and accounts for unobserved sinks (e.g., evapotranspiration, aquifer recharge and anthropogenic water resources) using a subset of cell states to accumulate mass that does not translate to streamflow. The streamflow output is the sum over the outgoing mass vector, excluding that subset of cell states representing unobserved mass sinks. Further details of this model are described in Nearing, Sampson, et al., 2020, Hoedt et al., 2021 and Frame et al., 2022. An example time series plot of predicted cell

states, including the subset of cell states representing mass sinks, is shown in Hoedt et al. (2021) in their Figure B1.

2.2.3 | Training

We used daily meteorological forcing data and static catchment attributes data as inputs features for the LSTM and MC-LSTM, and we used daily streamflow records as training targets with a normalized squared-error (NSE*) loss function that does not depend on basin-specific mean discharge (i.e., large and/or wet basins are not over-weighted in the loss function):

$$NSE^* = \frac{1}{B} \sum_{b=1}^B \sum_{n=1}^N \frac{(\hat{y}_n - y_n)^2}{(s(b) + \epsilon)^2} \quad (2)$$

where B is the number of basins, N is the number of samples (days) per basin B , \hat{y}_n is the prediction for sample n ($1 \leq n \leq N$), y_n is the corresponding observation, and $s(b)$ is the standard deviation of the discharge in basin b ($1 \leq b \leq B$), and ϵ is a small constant for numerical stability (we used 0.1), calculated from the training period (see Kratzert, Klotz, Shalev, et al., 2019).

We trained both the standard LSTM and the MC-LSTM using the same training and test procedures outlined by Kratzert, Klotz, Shalev, et al. (2019). Both models were trained for 30 epochs using sequence-to-one prediction to allow for randomized, small minibatches. We used a minibatch size of 256 and, due to sequence-to-one training, each minibatch contained (randomly selected) samples from multiple basins. The standard LSTM had 128 cell states and a 365-day sequence length. Input and target features for the standard LSTM

were pre-normalized by removing bias and scaling by variance. For the MC-LSTM the inputs were split between auxiliary, which were pre-normalized, and the mass input (in our case precipitation), which was not pre-normalized. Gradients were clipped to a global norm (per minibatch) of 1. Heteroscedastic noise was added to training targets (resampled at each minibatch) with standard deviation of 0.005 times the value of each target datum. We used an Adam optimizer with a fixed learning rate schedule; the initial learning rate of $1e-3$ was decreased to $5e-4$ after 10 epochs and $1e-4$ after 25 epochs. Biases of the LSTM forget gate were initialized to 3 so that gradient signals persisted through the sequence from early epochs. The MC-LSTM used the same hyperparameters as the LSTM except that it used only 64 cell states, which was found to perform better for this model (see Hoedt et al., 2021). Note that the memory states in an MC-LSTM are fundamentally different than those of the LSTM due to the fact that they are physical states with physical units instead of purely information states.

Both the LSTM and MC-LSTM were trained on data from 531 CAMELS catchments simultaneously. The train/test period split was the same split used in previous studies (Hoedt et al., 2021; Kratzert et al., 2021; Kratzert, Klotz, Shalev, et al., 2019). In this case, the training period included nine water years from 1 October 1999 through 30 September 2008, and the test period included 10 water years 1990–1999 (i.e., from 1 October 1989 through 30 September 1999). This train/test split was used *only* to ensure that the models trained here achieved similar performance compared with previous studies. Appendix includes an analysis of a different time period (the train period included water years 1981–1995, and the test period included water years 1996–2014), which was chosen to overlap with the NWM retrospective run.

2.3 | Performance metrics

We report two sets of performance metrics. The first set are standard benchmarking metrics that we report for two reasons: (i) to show that

the models perform similarly with previous benchmarking studies, and (ii) to allow us to demonstrate a distinction between model performance and consistency of long-term mass balance. The second set of metrics are related to long-term streamflow biases, and allow us to test our primary hypothesis. These metrics are described in the following two subsections.

2.3.1 | Standard performance metrics

We benchmarked all models using the same set of performance metrics that were used in previous CAMELS studies (Gauch, Kratzert, et al., 2021; Klotz et al., 2021; Kratzert et al., 2021; Kratzert, Klotz, Herrnegger, et al., 2019; Kratzert, Klotz, Shalev, et al., 2019). A full list of these metrics is given in Table 3. Each of the metrics was calculated for each basin separately on the whole test period for the training/test splits described in Section 2.2.3 (the test period consists of water years 1990–1999).

2.3.2 | Long-term mass balance

We conducted a long-term mass balance analysis using the absolute mass bias error for each basin:

$$\text{total absolute mass bias} = \frac{|\sum \text{obs.} Q - \sum \text{sim.} Q|}{\sum \text{obs.} Q} \quad (3)$$

where Q is the mass flux of streamflow. Note the absolute difference is taken as the numerator, so that positive and negative biases do not cancel each other out. The separated positive and negative mass bias errors for each basin are calculated as

$$\text{positive mass bias} = \begin{cases} x = \frac{\sum \text{obs.} Q - \sum \text{sim.} Q}{\sum \text{obs.} Q}, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

and

TABLE 3 Overview of performance benchmarking evaluation metrics for hydrological models

Metric	Description	Reference/equation	Details
NSE [†]	Nash-Sutcliffe efficiency	Eq. 3 in Nash and Sutcliffe (1970)	$(-\infty, 1]$, values closer to one are desirable.
KGE [‡]	Kling-Gupta efficiency Skill Score	Eq. 9 in Gupta et al. (2009)	$(-\infty, 1]$, values closer to one are desirable.
Pearson-r	Pearson correlation between observed and simulated flow		
α -NSE [§]	Ratio of standard deviations of observed and simulated flow	From Eq. 4 in Gupta et al. (2009)	$(0, \infty)$, values close to one are desirable.
β -NSE [¶]	Ratio of the means of observed and simulated flow	From Eq. 10 in Gupta et al. (2009)	$(-\infty, \infty)$, values close to zero are desirable.
Peak-Timing ^{††}	Mean peak time lag (in days) between observed and simulated peaks	Appendix B in Kratzert et al. (2021)	$(-\infty, \infty)$, values close to zero are desirable. This is a slightly different metric than described by Kratzert et al. (2021) in that we report the mean absolute peak time lag.

$$\text{negative mass bias} = \begin{cases} x = -\frac{\sum \text{obs.Q} - \sum \text{sim.Q}}{\sum \text{obs.Q}}, & \text{if } x < 0 \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

We used these metrics to provide a general measure the ability of each model to close the mass balance between precipitation and streamflow. These metrics require a continuous observation record, however this requirement is not satisfied for one of the CAMELS benchmarking basins for the test period of water years 1990–1999, leading us to discard it from our analysis. It is worth noting that these are technically volume calculations, and that we assume a constant liquid density for mass balance.

2.3.3 | Measuring information loss from modelling constraints

We used an information theoretic perspective to help investigate the hypothesis, that mass conservation inhibits the performance of hydrology models. Taking that idea one-step further, every constraint implemented on a hydrology model loses information from the inputs. Following the discussion by Nearing and Gupta (2015), we anticipate an ordering of information content like:

$$H_{\text{streamflow}} \geq I_{\text{input data}} \geq I_{\text{LSTM}} \geq I_{\text{MC-LSTM}} \geq I_{\text{SAC-SMA}} \quad (6)$$

H indicates the total entropy of whatever target data we are trying to predict (here a hydrograph in an individual basin). There is some amount of information in the input data (meteorological forcings and basin attributes), however the data processing inequality (Cover & Thomas, 2005, equation 2.122, p. 35) indicates that information is lost by any model which means that any model prediction contains less than, or equal to, information about the target data than is contained in the raw inputs (see Nearing & Gupta, 2015, for further discussion). Finally, we hypothesize that the constraints in the MC-LSTM (mass conservation) and the conceptual SAC-SMA model will mean that these two models provide less information than the LSTM. It is important to point out that the latter two terms of Equation (6) are only hypotheses—it is possible that adding constraints to a trained model (either a neural network or a calibrated conceptual model) will improve performance. We consider this unlikely, since adding constraints to a DL model serves only to restrict the space of functions that the model can emulate, however it is always possible that regularization will help avoid local minima during training, or otherwise compensate for limited information content of training data.

We quantified this (hypothesized) chain of inequalities using two difference metrics. The first metric is the standard mutual information (MI) metric calculated by histograms with 100 bins:

$$MI(U, V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} \frac{|U_i \cap V_j|}{N} \log \frac{N |U_i \cap V_j|}{|U_i| |V_j|} \quad (7)$$

where U is the observed streamflow, V is the simulated streamflow and N is the number of records. Mutual information obeys the data

processing inequality, so that the first and second terms of Equation (6) apply strictly. We calculated the MI in two ways: (1) at each basin individually for a distribution of values; and (2) using all of the flows from all basins combined for an overall MI score that does not account for distinctions between basins.

We also report the skill score outlined by Knoben et al. (2019) based on KGE metrics:

$$KGE_{\text{skill score}} = \frac{KGE_{\text{model}} - KGE_{\text{baseline}}}{1 - KGE_{\text{baseline}}} \quad (8)$$

where the skill score compares the performance of a candidate model with a baseline. This lets us draw a connection between the benchmarking metrics in Section 2.3.1 and Equation (6). KGE does not meet the non-negativity criteria as a formal f-divergence, and thus does not obey the data processing inequality, and therefore the relationship is only intuitive. The first level of constraint that we test is the strictly enforced mass conservation in the MC-LSTM and SAC-SMA. This is analogous to the third term in Equation (6), and in this case KGE_{model} and KGE_{baseline} in Equation (8) are the MC-LSTM and the LSTM, respectively. The second level of constraint is the conceptualization of the watershed as implemented by SAC-SMA model architecture (SAC-SMA was calibrated to long data records in each basin, and we included an ensemble of several calibrated models, each of which have a unique set of parameters). This is analogous to the third term in Equation (6), and in this case KGE_{model} and KGE_{baseline} in Equation (8) are SAC-SMA and the LSTM, respectively (we use the LSTM instead of the MC-LSTM as the baseline in order to plot a direct comparison between information lost by the MC-LSTM and SAC-SMA).

2.4 | Conditionality of the modelling analysis

Uncertainty in this experiment comes from three primary sources: data, models, and training. These sources are analogous to standard sources of uncertainty in most hydrology modelling studies: data, model structure, and model parameters (training is analogous to calibration).

The hypothesis that we are testing is related to understanding relationships between data and model uncertainty. Our objective is to understand how different models deal with uncertainty in data. We do not explicitly represent uncertainty in data (e.g., probabilistically), because our experiment does not require this for testing the hypothesis. We treat training/calibration uncertainty by using an ensemble of eight models, where the only difference between the ensemble members are different random weight initializations, and we take the ensemble mean of streamflow as our final model estimate. This approach is discussed explicitly for DL models by Kratzert, Klotz, Herrnegger, et al. (2019), and for the SAC-SMA model by Newman et al. (2015). Our analysis in Section 3.2 includes a box-and-whisker plot showing the median, standard deviation and outliers of each performance metric. We also include a complete (second) set of results of the same analysis on a different time period in Appendix, with results

that are nearly identical, indicating that our results are not the result of an anomalous time period.

Our models and training are consistent with previous studies. Benchmarking results like what are reported in Section 3.1 have been repeated by several research groups using different basins and different data products. Results presented here are consistent with previous large-sample studies for all models, which provides a degree of confidence about the modelling results in general. We include 95% confidence interval of the summary statistics, and these are relatively low given our large sample size.

3 | RESULTS

3.1 | Model performance

Table 4 provides performance metrics (Section 2.3.1) for the LSTM, MC-LSTM and SAC-SMA model simulations over the test period (water years 1990–1999). Most of these scores are broadly equivalent to the metrics for the same models reported by other studies (e.g., Kratzert, Klotz, Shalev, et al., 2019). More importantly, these metrics allow us to test the hypothesis that explicit mass conservation degrades performance (as a reminder, this hypothesis was proposed by Beven, 2020). What we are looking for in these metrics is that either all the mass conserving models perform worse than the non-constrained LSTM, which would support the hypothesis that mass conservation is detrimental to models, or that the MC-LSTM with an explicit mass conserving constraint does as well or better than the LSTM, which would indicate that the problem with the conceptual model is *not* a matter of enforcing closure over erroneous data.

Results show similar average performance between the LSTM and MC-LSTM, however there were mostly small differences. With Daymet forcings, the LSTM had a higher KGE score in 323 basins, with an average difference of 0.06, and the MC-LSTM had a higher KGE score in 208 basins. With NLDAS forcings, the LSTM had a higher KGE in 249 basins, while the MC-LSTM had a higher KGE score in 282 basins, and the average difference between KGE score was 0.05. In general, the median performance metrics of the two models were broadly comparable. Notably, both models were, on average, better across all metrics than SAC-SMA. Overall, these

results suggest that enforcing closure does *not* explain the differences between data-based and process-based and conceptual models.

3.2 | Long-term cumulative discharge

Figure 2 shows the cumulative density functions (CDFs) of long-term cumulative discharge from the 530 CAMELS basins from the models during the 1989–1999 test period. The LSTM, MC-LSTM and SAC-SMA all have a similar total mass bias with the NLDAS forcing. SAC-SMA has the lowest negative mass error, but the highest positive mass error. The LSTM has the highest negative mass error, but the lowest positive mass error. The MC-LSTM is generally in between the LSTM and SAC-SMA. Overall, the LSTM and MC-LSTM predicted streamflows that result in more accurate long-term cumulative discharge than the calibrated SAC-SMA model. The LSTM and the MC-LSTM performed roughly similarly on NLDAS, and MC-LSTM slightly outperformed the LSTM on Daymet. With Daymet forcing SAC-SMA's streamflow predictions are biased toward a very high positive mass error.

Figure 3 shows the long-term positive or negative mass biases distributed across the Contiguous United States (CONUS) from for the three models with both Daymet and NLDAS forcings. The result of the SAC-SMA simulation with Daymet forcings shows a clear positive mass bias error in the eastern half of CONUS. The result of the SAC-SMA simulation with NLDAS forcings shows a mix of positive and negative mass bias throughout CONUS. The LSTM and the MC-LSTM look relatively similar, to each other and for both NLDAS and Daymet forcing. There is a clear negative mass bias error down the middle of CONUS from about Montana through East Texas. This Central CONUS (CenCon) region (i.e., Missouri, Arkansas-White-Red and Texas-Gulf) is generally tough to predict, with conceptual, physical and deep learning models. SAC-SMA also shows a negative mass bias pattern in the same central CONUS region, though to a lesser spatial extent and higher magnitude, with Daymet forcings, but not so much with NLDAS forcings.

Figure 4 shows the mass bias errors for the model runs with Daymet forcings in box and whisker plots for the U.S. Water Resources Regions. SAC-SMA shows a very high mass balance error in the eleven eastern regions, but does much better in the western regions.

TABLE 4 Median performance metrics (plus or minus the 95% confidence interval) across 530 basins calculated on the test period 1990–1999 with two separate forcing products

Metric	Daymet forcing			NLDAS forcing		
	LSTM	MC-LSTM	SAC-SMA	LSTM	MC-LSTM	SAC-SMA
NSE	0.77 ± 0.02	0.76 ± 0.01	0.65 ± 0.03	0.74 ± 0.01	0.74 ± 0.01	0.67 ± 0.02
KGE	0.76 ± 0.02	0.76 ± 0.02	0.59 ± n/a	0.74 ± 0.02	0.74 ± 0.02	0.68 ± 0.02
Pearson-r	0.89 ± 0.01	0.88 ± 0.01	0.83 ± n/a	0.88 ± 0.01	0.87 ± 0.01	0.83 ± 0.01
Alpha-NSE	0.85 ± 0.01	0.84 ± 0.01	0.76 ± 0.02	0.81 ± 0.02	0.81 ± 0.02	0.78 ± 0.02
Beta-NSE	−0.04 ± 0.01	−0.03 ± 0.01	0.06 ± 0.01	−0.03 ± 0.01	−0.02 ± 0.01	−0.01 ± 0.01
Peak-timing	0.3 ± 0.03	0.3 ± 0.03	0.38 ± 0.06	0.32 ± 0.03	0.31 ± 0.03	0.41 ± 0.06

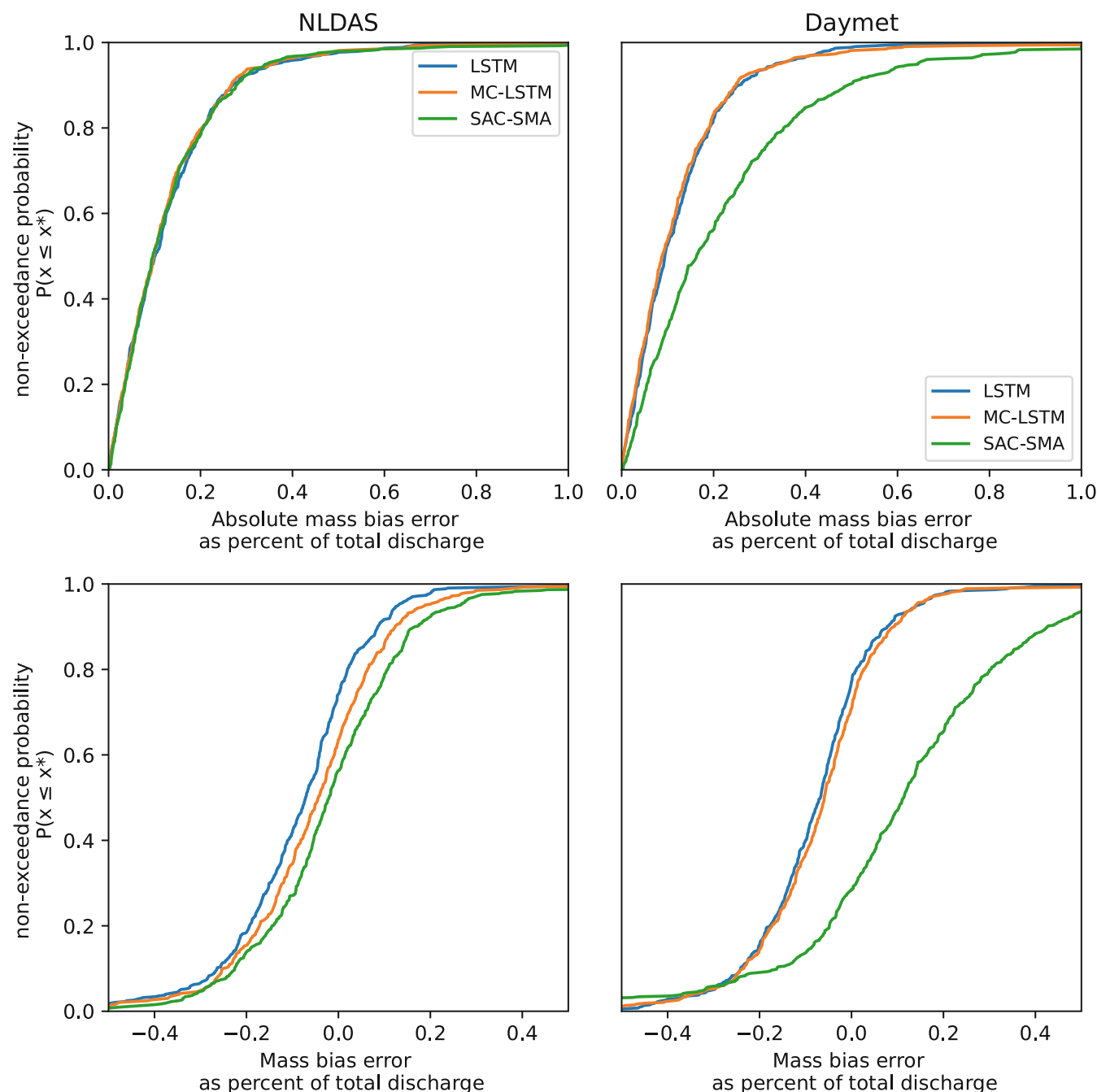


FIGURE 2 Distribution of mass balance error across the 530 basins. Top: Cumulative distribution curves of the absolute mass error from models forced with NLDAS (left) and Daymet (right). Bottom: Cumulative distributions of mass error from models forced with NLDAS (left) and Daymet (right)

The LSTM shows a high mass balance error in the Lower Colorado region, as compared to the MC-LSTM, and the MC-LSTM shows a higher mass balance error in the Rio Grande region, but the LSTM and MC-LSTM are relatively similar (more or less) in the other regions. All three models show relatively high mass bias errors in the CenCon region, which is the contribution from negative mass bias shown in Figure 3. SAC-SMA performs worse than the LSTM and the MC-LSTM in the CenCon region with Daymet forcings.

Figure 5 shows the mass bias errors for the model runs with NLDAS forcings in box and whisker plots for the U.S. Water

Resources Regions. With NLDAS forcing, SAC-SMA does not have a consistent mass bias error, as with Daymet. The pattern of SAC-SMA mass bias error in the western U.S. is generally similar between Daymet (Figure 4) and NLDAS (Figure 5). The differences between the LSTM, MC-LSTM and SAC-SMA do not show any obvious patterns. SAC-SMA and LSTM shows a high mass bias error outlier in the Lower Colorado region, but MC-LSTM does not. All three models show relatively high mass bias errors in the CenCon region, although SAC-SMA has a lower mean mass bias error than the LSTM and the MC-LSTM, but has a higher outlier in Missouri. Kratzert, Klotz, Herrnegger, et al.

FIGURE 3 Geospatial distribution of long-term positive or negative mass bias error. The left and right columns show the results with NLDAS and Daymet meteorological forcing data, respectively. The three rows are associated (from top to bottom) with LSTM, MC-LSTM and SAC-SMA

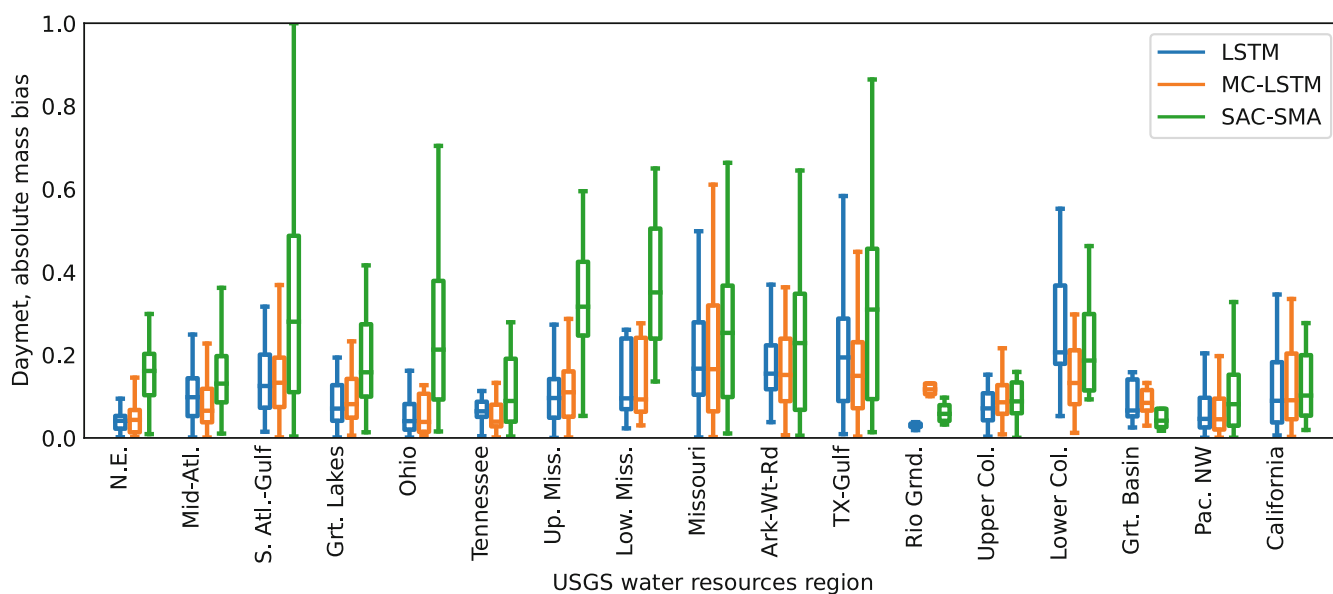
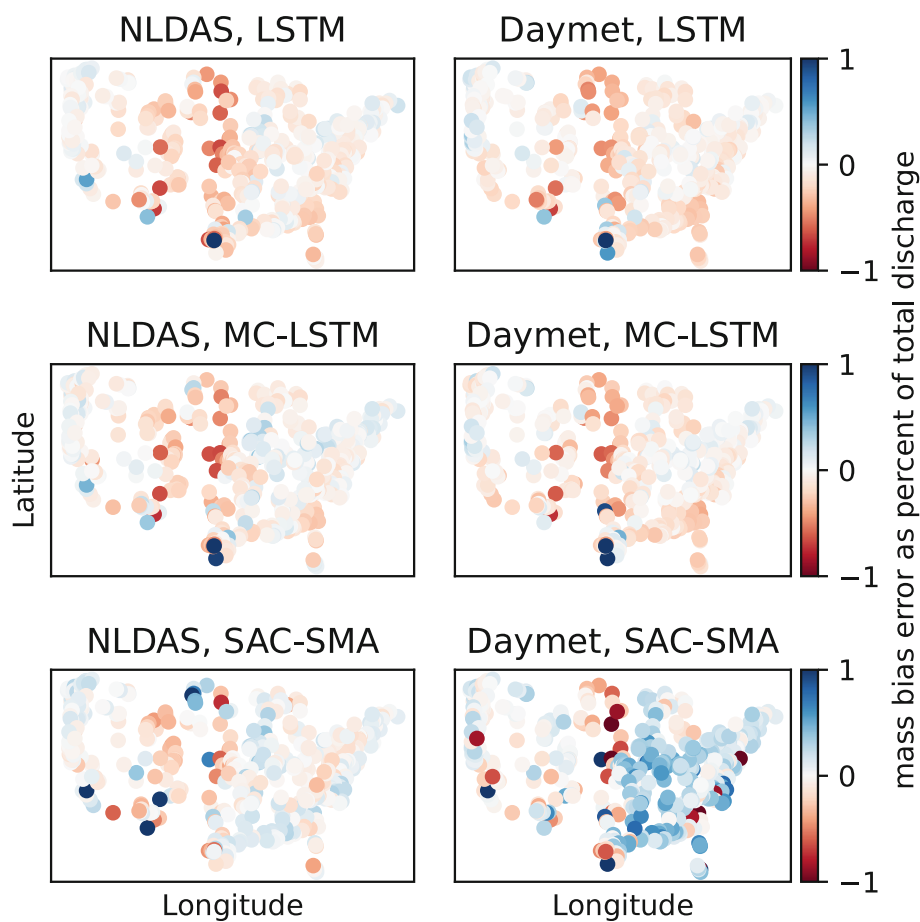


FIGURE 4 Regional mass balance errors from LSTM, MC-LSTM and SAC-SMA with Daymet forcings. The main body of each box shows the median and confidence intervals. The vertical lines extending to the most extreme, non-outlier data points. Souris-Red-Rainy region (Hydrologic Unit Code 09) is absent due to a lack of sufficient basins

(2019) show in their Figure 4 that the LSTM scores better in terms of Nash-Sutcliffe Efficiency than SAC-SMA, which seems to indicate that mass bias error in the catchment data does not explain the difference in predictive skill between deep learning and conceptual models.

Appendix includes results from a separate time period, where the NWM can be compared (with caveats of the inconsistent calibration period) on the NLDAS forcing data. The overall, spatial, and regional results are roughly similar for the LSTM, MC-LSTM and SAC-SMA.

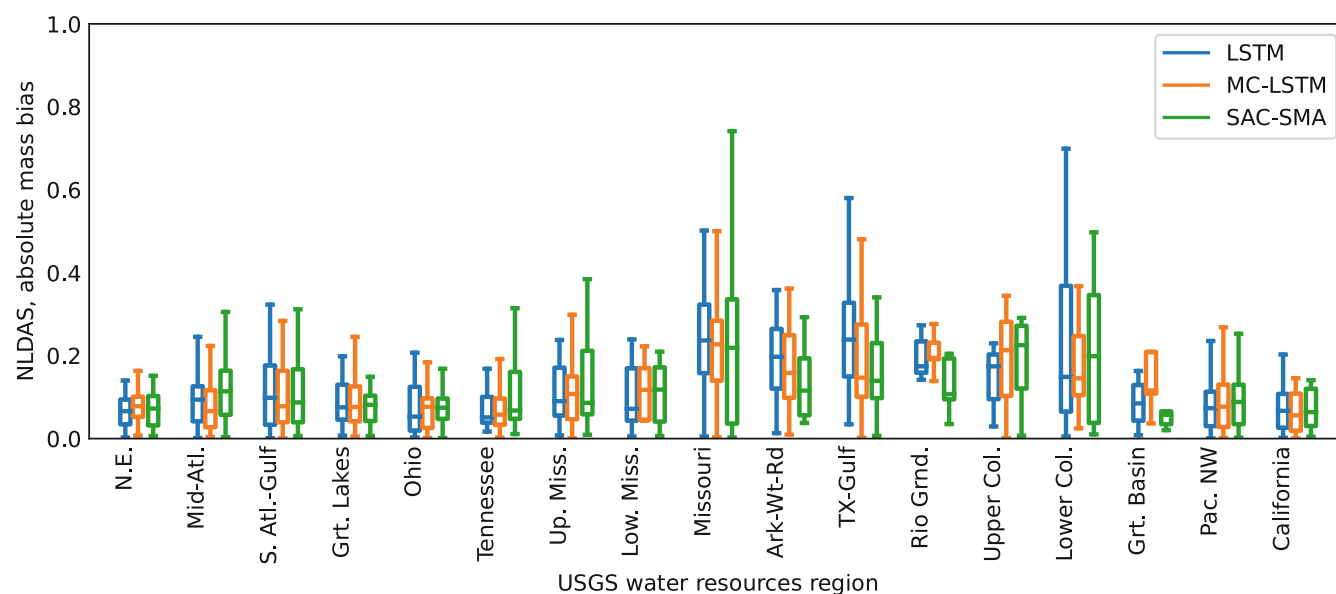


FIGURE 5 Regional mass balance errors from LSTM, MC-LSTM and SAC-SMA with NLDAS forcings. The main body of each box shows the median and confidence intervals. The vertical lines extending to the most extreme, non-outlier data points. Souris-Red-Rainy region (Hydrologic Unit Code 09) is absent due to a lack of sufficient basins

3.3 | Information loss due to modelling constraints

3.3.1 | Mutual information

The mutual information scores of the combined 530 basins (concatenated and calculated once across all basins) with NLDAS forcings are: 0.39 (LSTM), 0.37 (MC-LSTM) and 0.34 (SAC-SMA), respectively. The mutual information scores of the combined 530 basins (concatenated and calculated once across all basins) with Daymet forcings for models LSTM, MC-LSTM and SAC-SMA are 0.40, 0.37 and 0.33, respectively. Figure 6 shows the CDF plots with mutual information scores *calculated individually for each of the 530 basins*. For both Daymet and NLDAS the CDF curves show that LSTM has the most mutual information with the observed runoff, followed by MC-LSTM and then by SAC-SMA.

3.3.2 | KGE skill score

The unconstrained LSTM was used as a baseline model to measure information loss in the MC-LSTM and SAC-SMA. Results of the KGE skill score (KGE_{ss}) analysis are shown in Figure 7. The left subplot of this figure shows a clear ordering of model performance that agrees with what we hypothesized in Equation (6)—generally, model performance degrades as more constraints are added. The left subplot also shows that DL models perform better when trained and forced with Daymet data than with NLDAS data. This is somewhat counter-intuitive given the large, nonstationary bias that we saw in the previous section (Figure 2). SAC-SMA, however, performed significantly worse with the biased data. While the DL models (even those constrained to conserve mass) were able to learn to accommodate the

spatially heterogeneous biases in the input data, the PB model was not, even when trained on the biased data in each individual catchment. Daymet is the more informative precipitation product overall and a flexible DL model is able to learn and extract this information while the PB model cannot (even though the PB model is locally calibrated), however it is not the mass balance constraints that cause the problem.

The right subplot of Figure 7 plots the CDF of the skill scores (Equation 8) of the MC-LSTM and SAC-SMA relative to the unconstrained LSTM. The grey dotted vertical line represents a skill score of zero, indicating that the test model (MC-LSTM, SAC-SMA) performs equally well as the baseline (LSTM). The main takeaway from this figure is that adding mass balance constraints (both in the MC-LSTM and in SAC-SMA) helps more when using the NLDAS data, even though it was the Daymet data that showed biases.

Figure 8 plots the CDF of the difference between the MC-LSTM and the LSTM. In each basin, this difference represents an upper bound on the error introduced by mass balance constraints, relative to the LSTM. There are other possible reasons why the MC-LSTM might not perform as well as the LSTM (e.g., the way that it handles unobserved sources and sinks), however this difference (which is sometimes negative) is a conservative estimate of the error due to mass conservation in DL rainfall runoff models.

4 | CONCLUSIONS

The hypothesis tested in this paper is that errors in input/output (precipitation/streamflow) data cause apparent violations of closure that may largely explain the poor performance of conceptual models relative to deep learning. Given that the physical principle of mass balance

FIGURE 6 Left: Cumulative distribution of the mutual information for LSTM, MC-LSTM and SAC-SMA with Daymet forcing. Right: Cumulative distribution of the mutual information for LSTM, MC-LSTM and SAC-SMA with NLDAS forcing

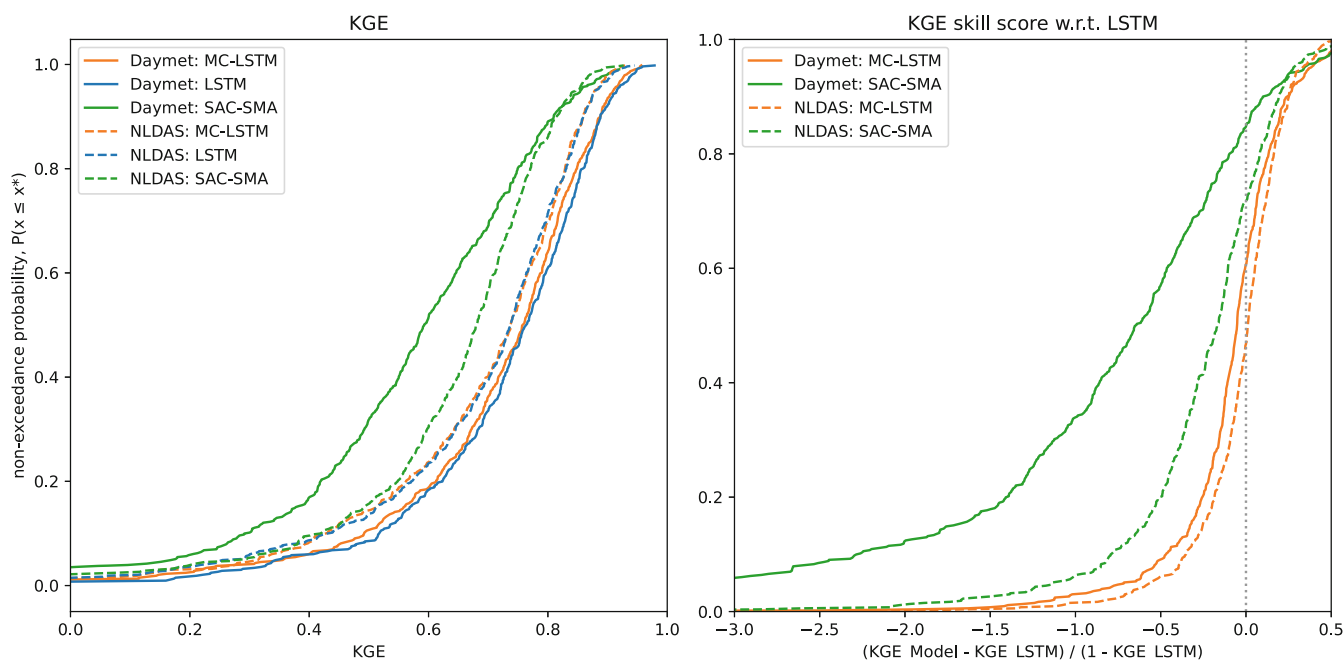
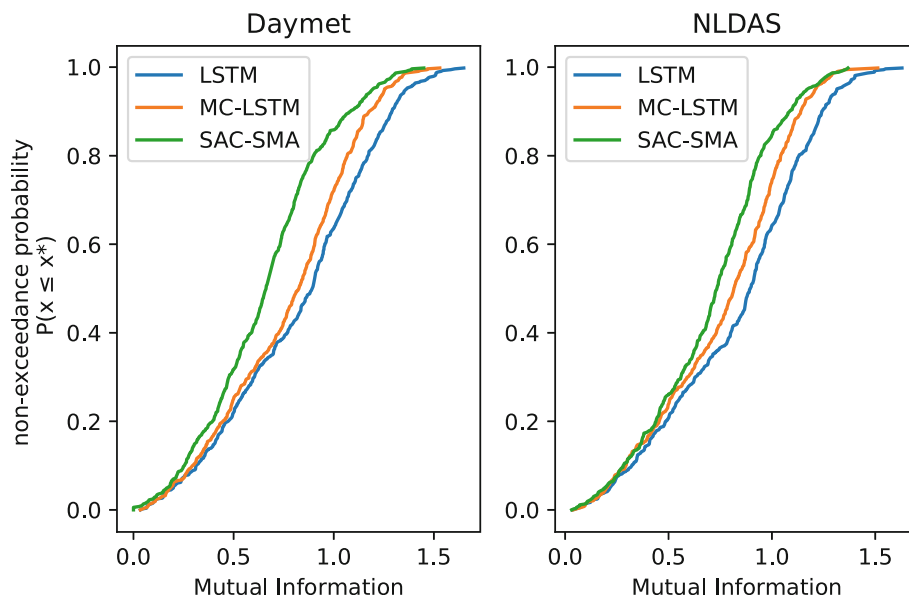


FIGURE 7 Left: Cumulative distribution of the Kling-Gupta Efficiency (KGE) for three models on two different forcing products. This subplot shows (i) that in general Daymet is more informative than NLDAS, and (ii) that the ordering of the inequality in Equation (6) is generally correct. Right: KGE skill scores (Equation 8) of SAC-SMA and MC-LSTM with respect to the unconstrained LSTM [positive values to the right of the dotted grey line mean that the MC-LSTM (SAC-SMA) performs better than the LSTM, and negative values to the left of the grey line mean that the MC-LSTM (SAC-SMA) performs worse than the LSTM]. This subplot shows that adding mass balance constraints to the LSTM has more benefit when using NLDAS inputs than when using Daymet inputs

over a control volume is one of the most fundamental components of hydrological theory, and is the first assumption we take for granted when developing theory-based hydrological models, it is arguably the first physical ‘law’ that we might test when developing physics-informed ML strategies for hydrology.

The long-term mass balance analysis (Section 3.2) shows that SAC-SMA includes a strong positive mass bias in the Eastern

U.S. with Daymet forcings. This indicates that, in regards to the discrepancy between NLDAS and Daymet long-term precipitation magnitudes, Daymet is likely positively biased, rather than NLDAS being negatively biased (i.e., Daymet overestimates precipitation magnitudes, rather than NLDAS underestimating). This propagates through the SAC-SMA predictions, and reduces the performance in terms of efficiency metrics, but not the Pearson-r nor the peak timing error.

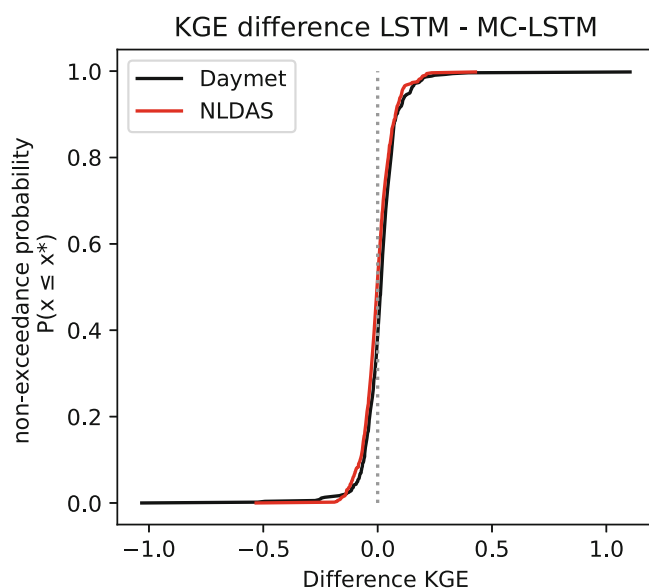


FIGURE 8 Cumulative density function of an estimated upper bound on the error introduced into deep learning streamflow predictions by adding a mass balance constraint

For Beven's closure-violation hypothesis to be true as an explanation for the general failure of traditional hydrology models, the errors in rainfall or discharge data must necessarily be systematic in a way that can be learned by a neural network, but not by a calibrated conceptual model. Our results indicate that this is indeed the case: the DL-based LSTM network was able to learn the non-uniform patterns of biases in input–output data, and thereby extract useful information from different (imperfect) precipitation products in spatially and temporally heterogeneous ways. This was true even when using heavily biased rainfall products that contribute significant so-called ‘disinformation’ when calibrating a conceptual model; for example, Daymet seems to actually contain more information about streamflow than the less biased NLDAS precipitation product. It is important to note that these data biases are not simple additive shifts in the mean – instead they are complex and heterogeneous throughout our large-sample dataset, and the DL models are able largely to learn this heterogeneity.

While Beven (2020) was correct that the imposition of conservation laws is generally harmful for hydrologic prediction, this fact does not help to explain most of the significantly better skill provided by DL over traditional (theory-based) rainfall-runoff models. These findings demonstrate two things:

1. that conservation laws may not be a good foundation for scale-relevant hydrological theory,
2. data that are supposedly ‘disinformative’ when used in the context of calibrating poorly conceived models might actually contain significant amounts of useful information that is accessible when used in the context of better conceived models.

In other words, for catchment-scale rainfall-runoff prediction it is arguably the current hydrological theory that is (more) disinformative,

not the hydrological data. In summary, model performance degrades as constraints are added, which causes loss of information between the inputs (atmospheric forcings) and the target (streamflow), as shown in 2.3.3.

5 | DISCUSSION

There is some subtlety to this conclusion due to the fact that the MC-LSTM includes a flux term that accounts for unobserved sinks (e.g., evapotranspiration, sublimation, aquifer recharge). Like all mass balance models, however, the MC-LSTM *explicitly* accounts for all water in and across the boundaries of the system. Even with this strong constraint, the MC-LSTM performs significantly better than the mass-conserving benchmark conceptual model. This result indicates that classical hydrology model structures (conceptual architectures and flux equations) actually cause prediction errors that are *larger* than can be explained as being due to errors in the forcing and observation data.

Our ability to properly conduct a more rigorous and detailed analysis of long-term water balances is limited by the fact that accurate evapotranspiration and percolation data (etc.) are not readily available at watershed scales. Nonetheless, what our analysis based on examining cumulative discharge shows is that an LSTM architecture *not* constrained to conserve mass is able to extract information from the available data that enables it to learn ‘effective’ water balances that are similar to those learned by a similar model architecture (MC-LSTM) that is explicitly constrained to enforce such closure, and that this effective water balance is in general better than that achieved by traditional conceptual and PB model architectures. Results from Lees et al. (2022) suggest that LSTM learns to reproduce stores of water, such as soil moisture and snow cover.

A likely reason for this is that the current body of hydrological theory does not aggregate well to the scale of unorganized complex watershed systems (Nearing, Kratzert, et al., 2020). While it is true that hydrological theory can enable a modeller to ‘interpret’ a watershed response (assuming a proper accounting for uncertainty), such theory does not currently translate into accurate predictions of catchment-scale behaviours using available data. Meanwhile, the most accurate way to generate a predictive model is to impose as few ‘physical constraints’ as possible on its ability to extract information from the available data, and consequently any model that is constrained to obey some ‘deeper’ physical understanding of the system must be *less* accurate in a predictive sense, unless that physical understanding actually contributes predictively-useful information that cannot be otherwise extracted directly from the data.

Looking forward, a particular application of rainfall-runoff modelling that necessarily requires the imposition of strict mass-balance constraints is ‘Earth-system-scale’ modelling. In this context, any model that seeks to explain components of long-term climate variability (for instance) cannot allow for any significant amount of residual mass to go unexplained. To use a dramatic example, unaccounted for losses at the catchment-scale could potentially result in the removal of all water mass from the global water cycle, which would render a

long-term simulation useless. Global-scale modelling of land-surface dynamics could be a potentially powerful application of the MC-LSTM network approach, and could be implemented by training additional model targets of mass-loss representations of 'losses' (transfers) to the sub-surface and 'losses' (transfers) to the atmosphere.

ACKNOWLEDGEMENTS

This research was supported by NOAA cooperative agreement (grant number NA19NES4320002), and the NASA Terrestrial Hydrology Program (grant number 80NSSC18K0982).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

All LSTMs and MC-LSTMs were trained using the NeuralHydrology Python library available at <https://github.com/neuralhydrology/neuralhydrology>. A snapshot of the exact version that we used is available at https://github.com/jmframe/mclstm_2021_extrapolate/neuralhydrology and under DOI number 10.5281/zenodo.5051961. Code for calibrating SAC-SMA is from <https://github.com/Upstream-Tech/SACSMA-SNOW17>, which includes the SpotPy calibration library <https://pypi.org/project/spotpy/>. Input data for all model runs except the NWM came from the public NCAR CAMELS repository <https://ral.ucar.edu/solutions/products/camels> and were used according to instructions outlined in the NeuralHydrology readme. NWM data are available publicly from <https://registry.opendata.aws/nwm-archive/>. All model output data generated by this project is available on the CUAHSI HydroShare platform under a DOI number <https://doi.org/10.4211/hs.d750278db868447dbd252a8c5431affd>. Interactive Python scripts for all post-hoc analysis reported in this paper, including calculating metrics and generating tables and figures, are available at https://github.com/jmframe/mclstm_2021_mass_balance and under DOI number 10.5281/zenodo.7641775.

ORCID

Jonathan M. Frame  <https://orcid.org/0000-0002-2533-3843>

Paul Ullrich  <https://orcid.org/0000-0003-4118-4590>

REFERENCES

- 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 (HESS)*, 21, 5293–5313.
- Bennett, A., & Nijssen, B. (2021). Deep learned process parameterizations provide better representations of turbulent heat fluxes in hydrologic models. *Water Resources Research*, 57, 1–14. <https://doi.org/10.1029/2020WR029328>
- Beven, K. (1989). Changing ideas in hydrology—The case of physically-based models. *Journal of Hydrology*, 105, 157–172. [https://doi.org/10.1016/0022-1694\(89\)90101-7](https://doi.org/10.1016/0022-1694(89)90101-7)
- Beven, K. (2020). Deep learning, hydrological processes and the uniqueness of place. *Hydrological Processes*, 34, 3608–3613.
- Beven, K., & Westerberg, I. (2011). On red herrings and real herrings: Disinformation and information in hydrological inference. *Hydrological Processes*, 25, 1676–1680. <https://doi.org/10.1002/hyp.7963>
- Beven, K. J., Smith, P. J., & Freer, J. E. (2008). So just why would a modeller choose to be incoherent? *Journal of Hydrology*, 354, 15–32. <https://doi.org/10.1016/j.jhydrol.2008.02.007>
- Chawanda, C. J., Arnold, J., Thiery, W., & van Griensve, A. (2020). Mass balance calibration and reservoir representations for large-scale hydrological impact studies using SWAT+. *Climatic Change*, 163, 1307–1327.
- Chow, V. T., Maidment, D. R., & Mays, L. W. (1988). *Applied hydrology* Chow 1988.pdf. McGraw-Hill, http://ponce.sdsu.edu/Applied_Hydrology_Chow_1988.pdf
- Cover, T. M., & Thomas, J. A. (2005). *Elements of information theory*. John Wiley & Sons, Inc. <https://doi.org/10.1002/047174882X>.
- Daw, A., Thomas, R. Q., Carey, C. C., Read, J. S., Appling, A. P., & Karpatne, A. (2020). Physics-guided architecture (PGA) of neural networks for quantifying uncertainty in lake temperature modeling. *Proceedings of the 2020 SIAM International Conference on Data Mining* (pp. 532–540). <https://doi.org/10.1137/1.9781611976236.60>
- Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shelev, G., Gilon, O., Qualls, L. M., Gupta, H. V., & Nearing, G. S. (2022). Deep learning rainfall-runoff predictions of extreme events. *Hydrology and Earth System Sciences*, 26, 3377–3392. <https://doi.org/10.5194/hess-26-3377-2022>
- Frame, J. M., Kratzert, F., Raney, A., Rahman, M., Salas, F. R., & Nearing, G. (2021). Post-processing the national water model with long short-term memory networks for streamflow predictions and model diagnostics. *Journal of the American Water Resources Association*, 57, 1–21. <https://doi.org/10.1111/1752-1688.12964>
- Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., & Hochreiter, S. (2021). Rainfall-runoff prediction at multiple timescales with a single long short-term memory network. *Hydrology and Earth System Sciences*, 25, 2045–2062.
- 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, 104.
- Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377, 80–91.
- 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, 463–477. <https://doi.org/10.5194/hess-18-463-2014>
- Hoedt, P.-J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G. S., Hochreiter, S., & Klambauer, G. (2021). MC-LSTM: Mass-conserving LSTM. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 4275–4286). PMLR, <http://proceedings.mlr.press/v139/hoedt21a.html>
- 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. *Geophysical Research Abstracts*, 21, 7878.
- Jia, X., Willard, J., Karpatne, A., Read, J. S., Zwart, J. A., Steinbach, M., & Kumar, V. (2021). Physics-guided machine learning for scientific discovery: An application in simulating lake temperature profiles. *ACM/IMS Transactions In Data Science*, 2, 1–26. <https://doi.org/10.1145/3447814>
- Jiang, S., Zheng, Y., & Solomatine, D. (2020). Improving AI system awareness of geoscience knowledge: Symbiotic integration of physical approaches and deep learning. *Geophysical Research Letters*, 47, e2020GL088.
- Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, 3, 422–440. <https://doi.org/10.1038/s42254-021-00314-5>
- Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., & Nearing, G. (2021). Uncertainty

- estimation with deep learning for Rainfall–Runoff modelling. *Hydrology and Earth System Sciences Discussions*, 26, 1–32.
- Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, 23, 4323–4331.
- 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, 6005–6022.
- Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. (2019). Toward improved Predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, 55, 11344–11354. <https://doi.org/10.1029/2019WR026065>
- Kratzert, F., Klotz, D., Hochreiter, S., & Nearing, G. S. (2021). A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling. *Hydrology and Earth System Sciences*, 25, 2685–2703. <https://doi.org/10.5194/hess-25-2685-2021>
- 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, 5089–5110. <https://doi.org/10.5194/hess-23-5089-2019>
- Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., Kumar Sahu, R., Greve, P., Slater, L., & Dadson, S. J. (2022). Hydro-logical concept formation inside long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*, 26, 3079–3101. <https://doi.org/10.5194/hess-26-3079-2022>
- Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology*, 10, 282–290.
- Nearing, G., Research, G., Kratzert, F., Klotz, D., Hoedt, P.-J., Klambauer, G., Hochreiter, S., Gupta, H., Nevo, S., & Matias, Y. (2020). A Deep Learning Architecture for Conservative Dynamical Systems: Application to Rainfall–Runoff Modeling, AI for Earth Sciences Workshop NEURIPS 2020.
- Nearing, G., Sampson, A. K., Kratzert, F., & Frame, J. (2020). Post-processing a conceptual rainfall–runoff model with an LSTM.
- Nearing, G. S., & Gupta, H. V. (2015). The quantity and quality of information in hydrologic models. *Water Resources Research*, 51, 524–538.
- Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., & Gupta, H. V. (2020). What role does hydrological science play in the age of machine learning? *Water Resources Research*, 57, e2020WR028091.
- Newman, A., Clark, M., Sampson, K., Wood, A., Hay, L., Bock, A., Viger, R., Blodgett, D., Brekke, L., Arnold, J., & 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, 209.
- Newman, A. J., Mizukami, N., Clark, M. P., Wood, A. W., Nijssen, B., & Nearing, G. (2017). Benchmarking of a physically based hydrologic model. *Journal of Hydrometeorology*, 18, 2215–2225.
- Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., Kumar, A., Manning, K., Niyogi, D., Rosero, E., Tewari, M., & Xia, Y. (2011). The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *Journal of Geophysical Research: Atmospheres*, 116. <https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2010JD015139>
- Pelissier, C., Frame, J., & Nearing, G. (2019). Combining parametric land surface models with machine learning. <http://arxiv.org/abs/2002.06141>
- Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven earth system science. *Nature*, 566, 195–204.
- 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, 7–27.
- Shen, C., Chen, X., & Laloy, E. (2021). Editorial: Broadening the use of machine learning in hydrology. *Frontiers in Water*, 3, 1–4. <https://doi.org/10.3389/frwa.2021.681023>
- Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J. J., Mendiola, E. M., O'Connell, P. E., Oki, T., Pomeroy, J. W., Schertzer, D., Uhlenbrook, S., & Zehe, E. (2003). IAHS decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. *Hydrological Sciences Journal*, 48, 857–880. <https://doi.org/10.1623/hysj.48.6.857.51421>
- Thornton, P., Thornton, M., Mayer, B., Wilhelmi, N., Wei, Y., Devarakonda, R., & Cook, R. (2014). Daymet: Daily surface weather data on a 1-km Grid for North America, Version 2. <https://doi.org/10.3334/ORNLDAC/1219>
- Tsai, W.-P., Pan, M., Lawson, K., Liu, J., Feng, D., & Shen, C. (2021). From parameter calibration to parameter learning: Revolutionizing large-scale geoscientific modeling with big data. *Nature Communications*, 12, Article 5988.
- USGS. (1987). Hydrologic unit maps: U.S. geological survey water-supply paper 2294. http://pubs.usgs.gov/wsp/wsp2294/pdf/wsp_2294.pdf
- Willard, J., Jia, X., Xu, S., Steinbach, M., & Kumar, V. (2021). Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Computing Surveys*, 1, 1–35.
- Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., Luo, L., Alonge, C., Wei, H., Meng, J., Lin, H., 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*, 117, D03109. <https://doi.org/10.1029/2011JD016048>
- 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>
- Zhao, W. L., Gentile, P., Reichstein, M., Zhang, Y., Zhou, S., Wen, Y., Lin, C., Li, X., & Qiu, G. (2019). Physics-constrained machine learning of evapotranspiration. *Geophysical Research Letters*, 46, 14496–14507. <https://doi.org/10.1029/2019GL085291>

How to cite this article: Frame, J. M., Kratzert, F., Gupta, H. V., Ullrich, P., & Nearing, G. S. (2023). On strictly enforced mass conservation constraints for modelling the Rainfall–Runoff process. *Hydrological Processes*, 37(3), e14847. <https://doi.org/10.1002/hyp.14847>

APPENDIX A: : COMPARISON WITH THE U.S. NATIONAL WATER MODEL

The NOAA National Water Model (NWM) retrospective run version 2 is used as an additional benchmark because of its wide scale use and availability. We can only use the NWM to compare with NLDAS data, as it has not been run with Daymet. The NWM is based on WRF-Hydro (Salas et al., 2018), which is a model that includes Noah-MP (Niu et al., 2011) as a land surface component, kinematic wave overland flow, and Muskingum-Cunge channel routing. NWM was previously used as a benchmark for LSTM simulations in CAMELS by Kratzert, Klotz, Herrnegger, et al. (2019), Gauch, Kratzert, et al. (2021) and Frame et al. (2021). Public data from NWM is hourly and CONUS-wide—we pulled hourly flow estimates from the USGS gauges in the CAMELS data set and averaged these hourly data to daily over the time period 1 October 1980 through 30 September 2008. As a point of comparison, Gauch, Kratzert, et al. (2021) compared hourly and daily LSTM predictions against the NWM and found that the NWM was significantly more accurate at the daily timescale than at the hourly timescale, whereas the LSTM did not lose accuracy at the hourly timescale versus the daily timescale. All experiments in the present study were done at the daily timescale.

The NWM is also susceptible to the kinds of mass bias error propagation from the forcings. We cannot, however, test the same hypothesis with the NWM because we do not have the capability to re-calibrate and run the NWM with Daymet forcing, as the complete set of data to run the NWM are not publicly available. The National Oceanic and Atmospheric Association (NOAA) has made publicly available a NWM retrospective run using NLDAS forcing data. This allows us to directly compare the mass balance errors with the LSTM, MC-LSTM and SAC-SMA. The NWM retrospective run (NWM) does not completely overlap with our test period (1989–1999). We performed the same experiment on a test period that can be compared with the NWM, which includes training/calibrated the LSTM, MC-LSTM and SAC-SMA. The train/test period split used a test period that aligns with the availability of benchmark data from the US National Water Model. The train period included water years 1981–1995, and the test period included water years 1996–2014 (i.e., from 1 October 1995 through 30 September 2014). This was the same training period used by Newman et al. (2017) and Kratzert, Klotz, Herrnegger, et al. (2019), but with an extended test period. This train/test split was used because the NWM data record is not long enough to accommodate the train/test split used by previous studies (item above in this list).

The NWM was calibrated by NOAA personnel on about 1400 basins with NLDAS forcing data and includes a regionalization strategy that attempts to use the calibrated parameters across basins not included in the calibration set, however most of the CAMELS basins are included in that calibration set. The NWM calibration time period is on water years 2009–2013. Because of the inconsistencies in the time period and basins included in the calibration, we cannot directly compare the NWM to the other models. But we include the NWM

here as [Appendix](#) because it is relevant to the hydrologic community, even if not directly comparable, because of the ongoing development and growing user-base of the NWM (Table A1).

The time period split used in this [Appendix](#) has a similar discrepancy between the total precipitation for NLDAS and Daymet, shown in Figure A1, which is a positive Daymet bias or a negative NLDAS bias in basins with large precipitation totals from about –100 to –70 degrees longitude.

Figure A2 shows the cumulative density functions (CDFs) of long-term mass biases from the 484 CAMELS basins from the models during the 1996–2014 test period. Note that we excluded basins that did not have a complete observation time series throughout the entire test period. The LSTM and MC-LSTM both predicted streamflows that result in more accurate long-term cumulative discharge than the calibrated SAC-SMA model and the NWM. The LSTM and the MC-LSTM performed roughly similarly on both NLDAS and Daymet on absolute mass bias, but LSTM does slightly better on negative mass bias. The MC-LSTM does slightly better on positive mass bias with NLDAS forcings, but are roughly similar with Daymet forcings.

Figure A3 shows the Mass balance results from for the three models with both Daymet and NLDAS forcings. The result of the SAC-SMA simulation with Daymet forcings shows a clear positive mass bias error in the eastern half of CONUS. The result of the simulation with NLDAS forcings shows a mix of positive and negative mass bias throughout CONUS.

Figure A4 shows the mass bias errors for the model runs with NLDAS forcings in box and whisker plots for the U.S. Water Resources Regions. A mass bias error is clearly shows for SAC-SMA in the Easter CONUS regions, while the LSTM and MC-LSTM do not express this pattern. There is generally a correlation between the three models, where the regions with high mass bias error are expressed by all three models. For instance, the Upper Colorado region shows low mass bias error for all three models, and the Lower Colorado shows a relatively high mass bias error by all three models.

Figure A5 shows the mass bias errors for the model runs with Daymet forcings in box and whisker plots for the U.S. Water Resources Regions. The NWM has high outliers in the Central CONUS regions. There is a correlation of mass bias error across all four models, where when the conceptual models (SAC-SMA and NWM), the physics informed ML Model (MC-LSTM) and the pure data driven model (LSTM) all show relatively small to moderate mass bias error in the Northeastern CONUS, high mass bias error in the central CONUS and moderate mass bias in west coast regions. The exception to this trend is that the Great Basin has a low mass bias error from LSTM, MC-LSTM and SAC-SMA, but a high mass bias error from NWM.

Results of the KGE skill score (KGE_{ss}) analysis are shown in Figure A6. The left subplot of this figure shows a clear ordering of model performance that agrees with what we hypothesized in Equation (6)—generally, model performance degrades as more constraints are added. The left subplot also shows that DL models perform better when trained and forced with Daymet data than with NLDAS data. The right subplot of Figure A7 plots the CDF of the skill

scores (Equation 8) of the MC-LSTM and SAC-SMA relative to the unconstrained LSTM. The grey dotted vertical line represents a skill score of zero, indicating that the test models (NWM, MC-LSTM, SAC-SMA) performs equally well as the baseline (LSTM). NWM-Rv2 does worse than SAC-SMA on the low end of the distributions, but slightly better on the high ends.

Mutual information of the models are shown in Figure A7. NWM shows more information loss than SAC-SMA in the worst performing

basins, but less in the better performing basins. This could be because the NWM has more constraints in the form of a multi-layered modelling chain. The NWM starts with a land surface model, which causes runoff across a terrain routing model, which is also two-way coupled with the land model, and finally the terrain model feeds into the channel routing model, which provides an estimate of streamflow. There are multiple steps along that modelling chain that cause different amounts of information loss.

TABLE A1 Median performance metrics (plus or minus the 95% confidence interval) across 484 basins calculated on the test period 1996–2014 with two separate forcing products

Metric	Daymet forcing			NLDAS forcing			
	LSTM	MC-LSTM	SAC-SMA	LSTM	MC-LSTM	SAC-SMA	NWM*
NSE	0.74 ± −0.02	0.74 ± −0.02	0.59 ± −0.08	0.71 ± −0.05	0.72 ± −0.02	0.63 ± −0.05	0.63 ± −0.05
KGE	0.78 ± −0.02	0.77 ± −0.02	0.56 ± n/a	0.77 ± −0.02	0.74 ± −0.02	0.68 ± −0.02	0.67 ± −0.05
Pearson-r	0.88 ± −0.01	0.88 ± −0.01	0.81 ± n/a	0.86 ± −0.01	0.86 ± −0.01	0.81 ± −0.01	0.82 ± −0.01
Alpha-NSE	0.96 ± −0.02	0.91 ± −0.01	0.88 ± −0.02	0.94 ± −0.02	0.87 ± −0.02	0.83 ± −0.02	0.85 ± −0.03
Beta-NSE	0.03 ± −0.01	0.03 ± −0.01	0.13 ± −0.02	0.01 ± −0.01	−0.01 ± −0.01	−0.01 ± n/a	−0.01 ± n/a
Peak-timing	0.34 ± −0.03	0.33 ± −0.03	0.45 ± −0.06	0.38 ± −0.03	0.4 ± −0.03	0.53 ± −0.06	0.54 ± −0.05

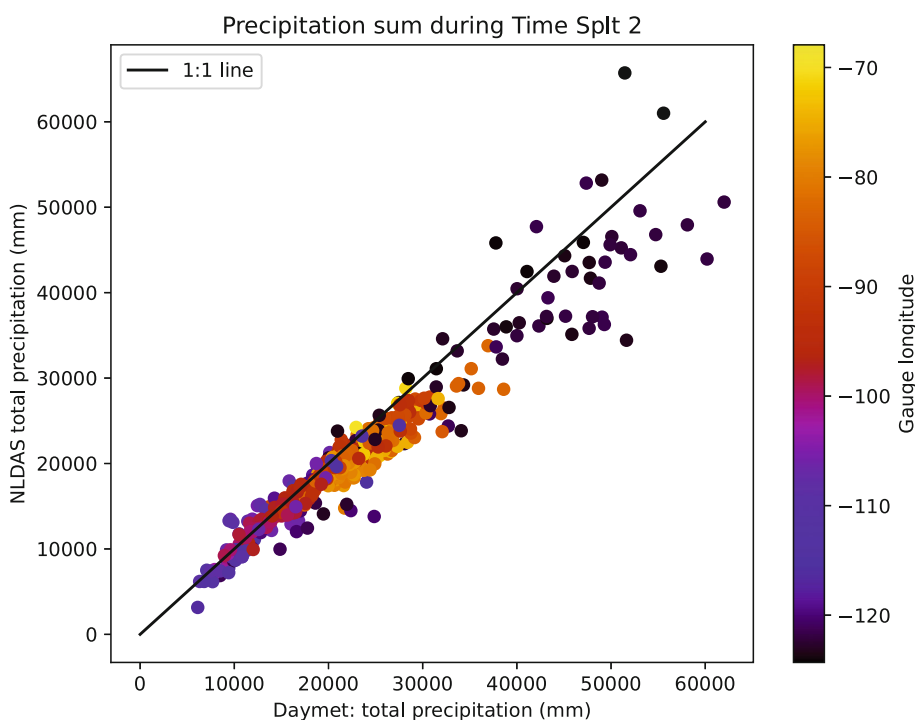


FIGURE A1 Comparison of the total precipitation from NLDAS and Daymet for each of the CAMELS basins, colour-coded by the longitudinal coordinate of the stream gauge

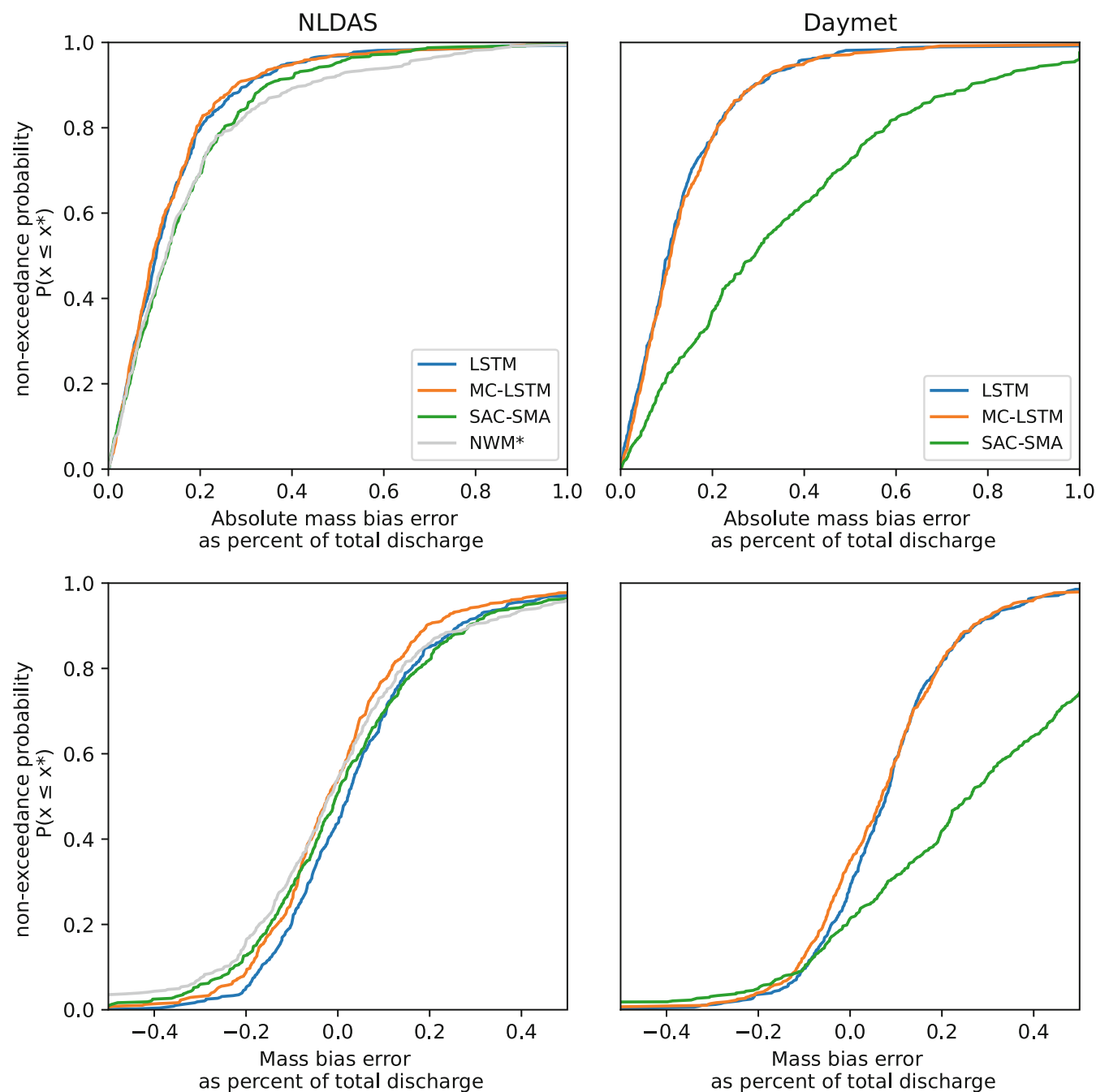


FIGURE A2 Distribution of mass balance error across the 484 basins. Top: Cumulative distribution curves of the absolute mass error for models with NLDAS (left) and Daymet (right). Bottom: Cumulative distributions of mass error from models forced with NLDAS (left) and Daymet (right)

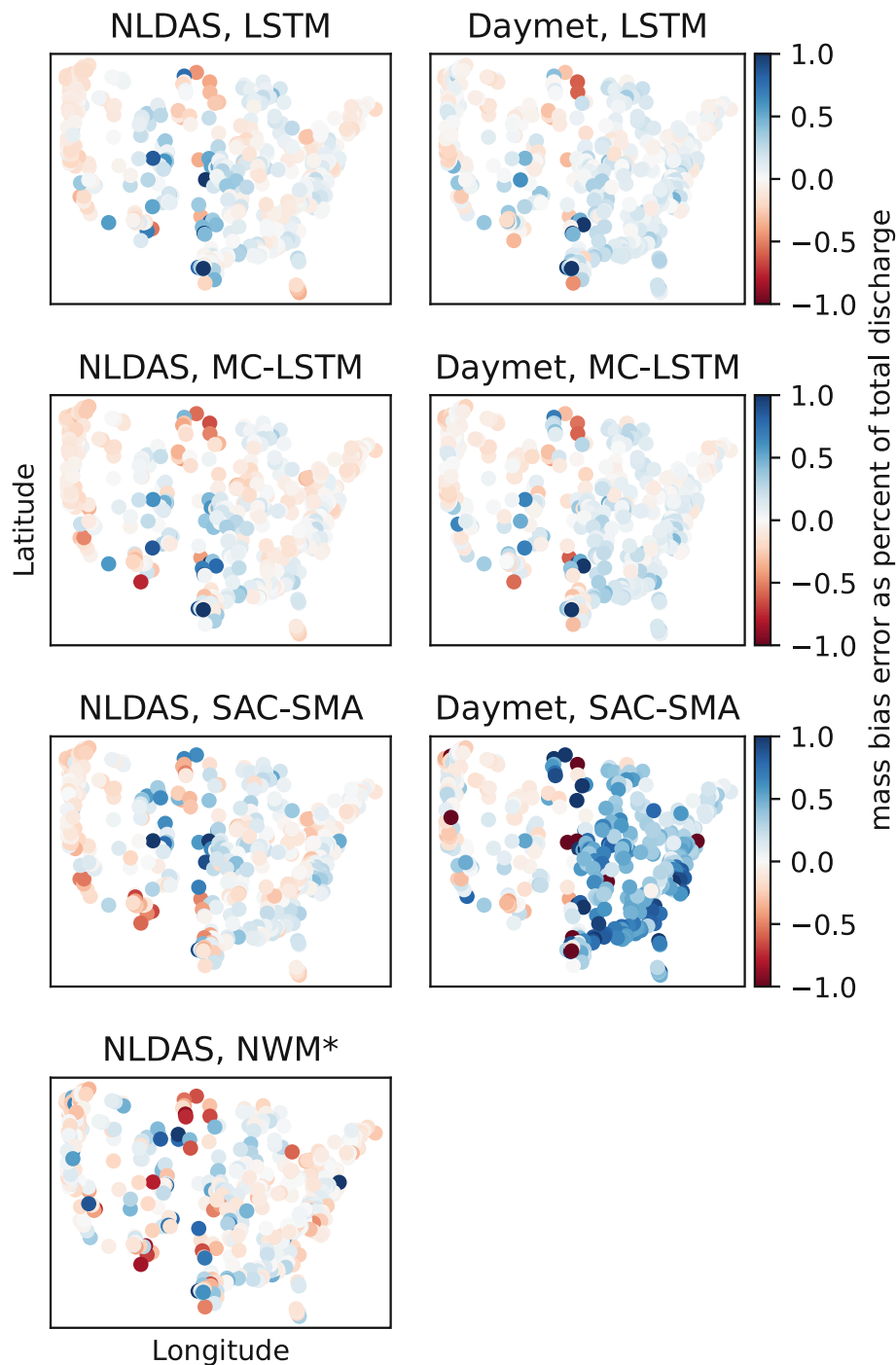


FIGURE A3 Geospatial distribution of long-term positive or negative mass bias error. The left and right columns show the results with NLDAS and Daymet meteorological forcing data, respectively. The four rows are associated (from top to bottom) with LSTM, MC-LSTM, SAC-SMA and NWM. The asterisk (*) on the bottom left sub-plot label indicates that the NWM was not calibrated on the same time period as the LSTM, MC-LSTM and SAC-SMA models

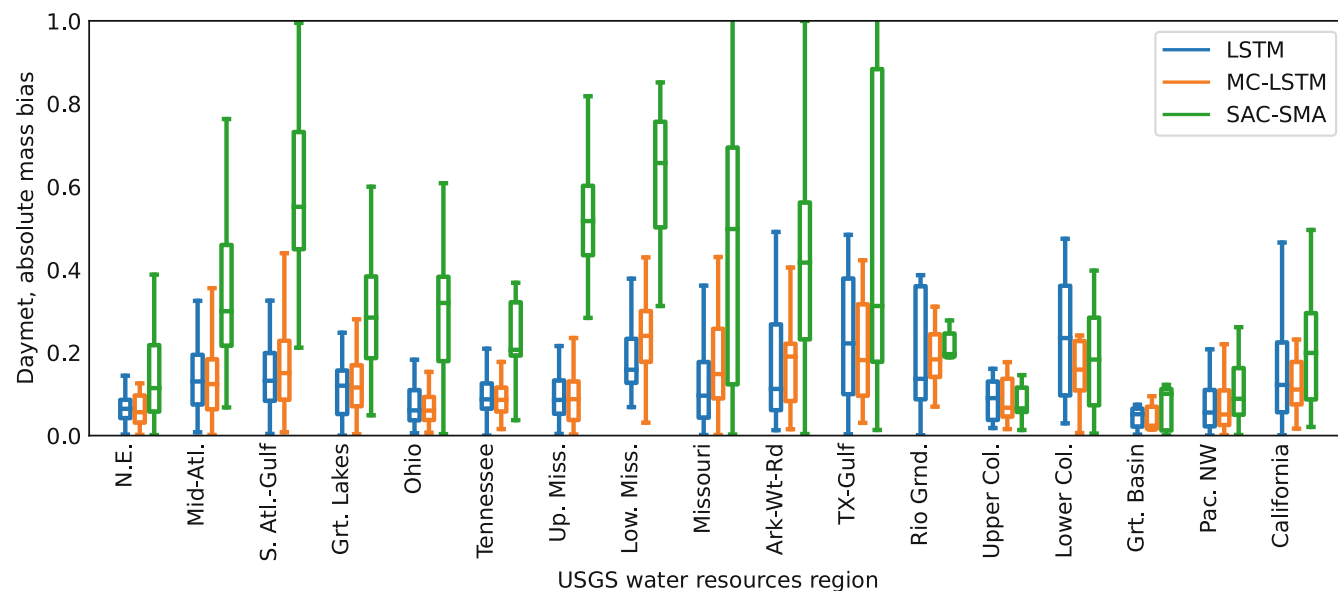


FIGURE A4 Regional mass balance errors from LSTM, MC-LSTM and SAC-SMA with Daymet forcings. Souris-Red-Rainy region (Hydro-logic Unit Code 09) is absent due to a lack of sufficient basins

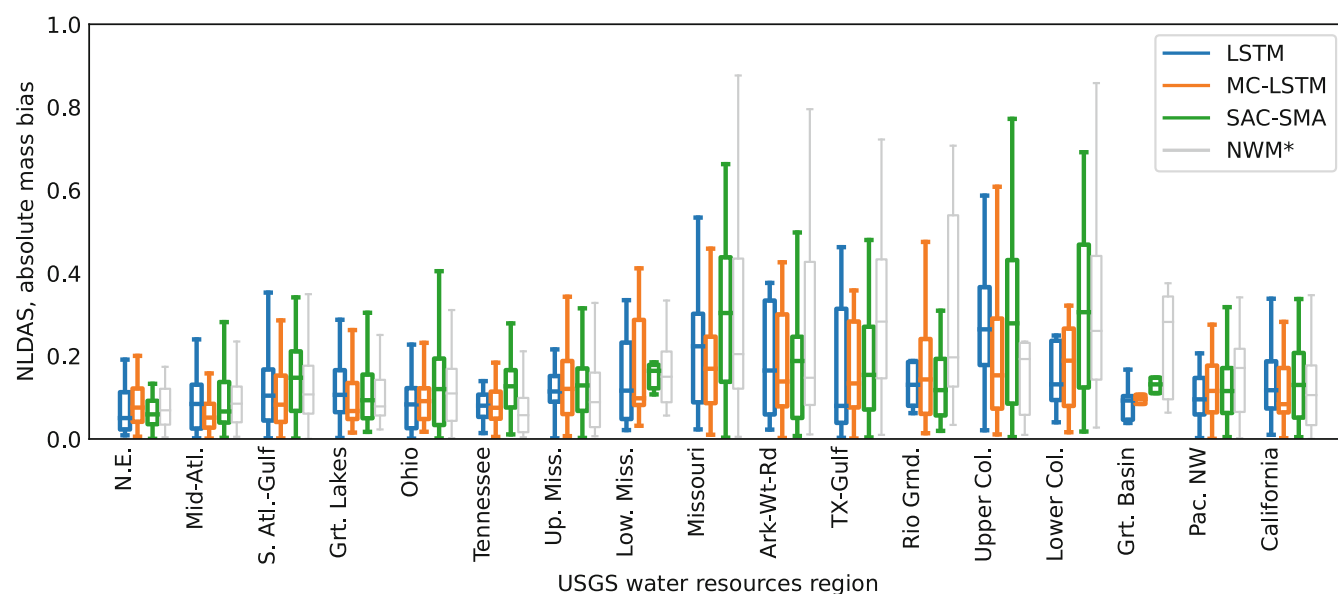


FIGURE A5 Regional mass balance errors from LSTM, MC-LSTM and SAC-SMA with NLDAS forcings. Souris-Red-Rainy region (Hydro-logic Unit Code 09) is absent due to a lack of sufficient basins. The asterisk (*) on the NWM label indicates that the model was calibrated on a separate time period than the other three models, and is thus not directly comparable

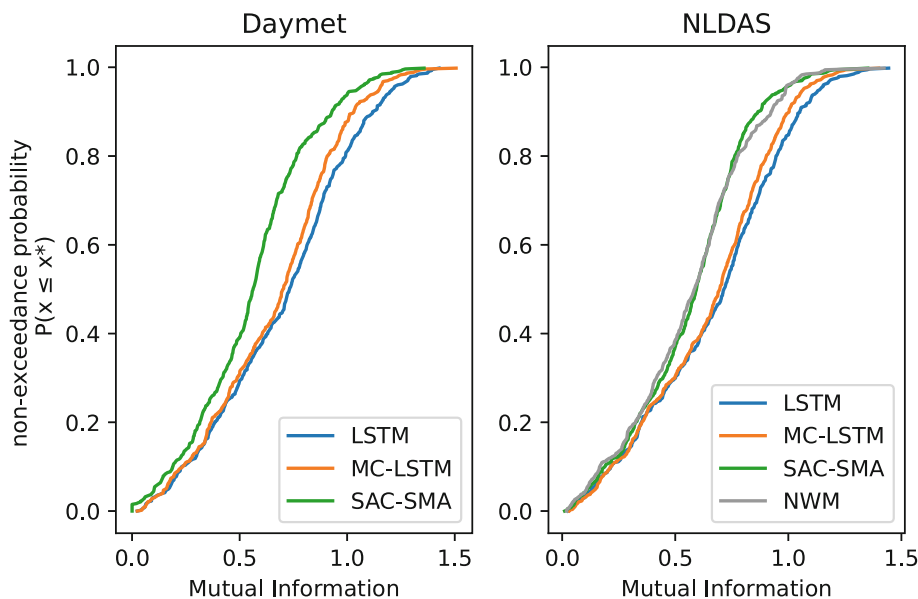


FIGURE A6 Left: Cumulative distribution of the mutual information for LSTM, MC-LSTM and SAC-SMA with Daymet forcing. Right: Cumulative distribution of the mutual information for LSTM, MC-LSTM, SAC-SMA and NWM with NLDAS forcing

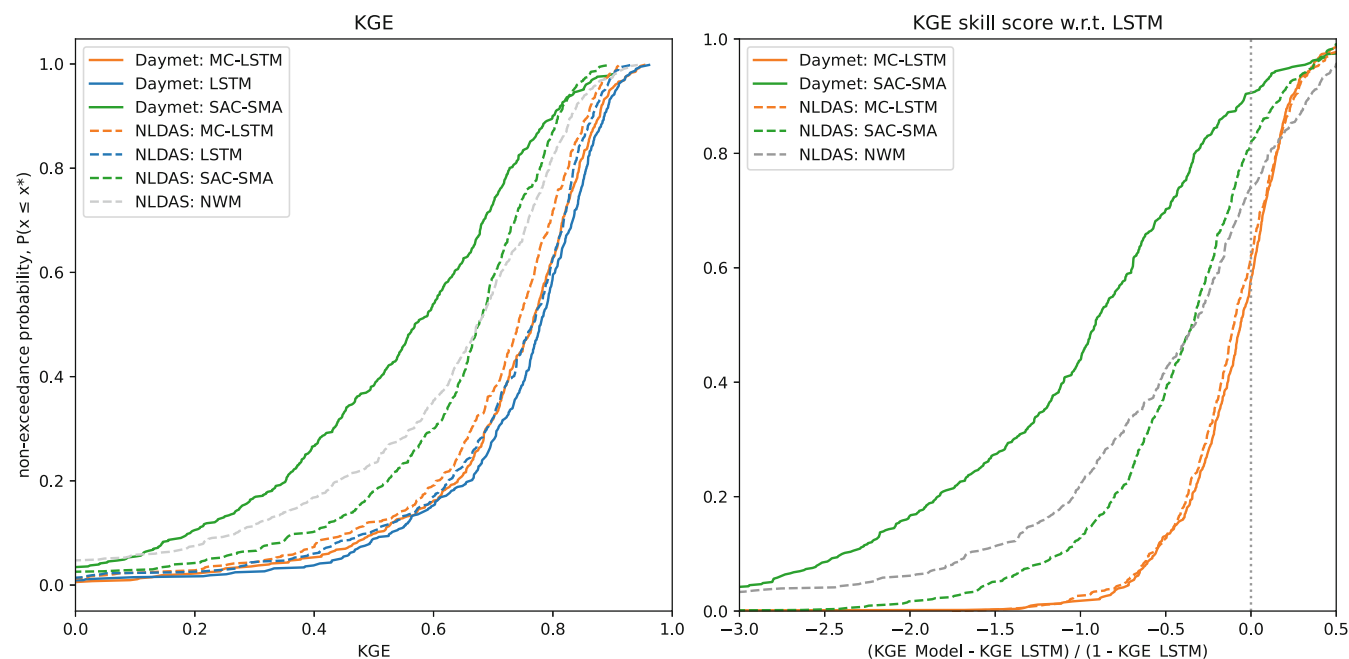


FIGURE A7 Left: Cumulative distribution of the Kling-Gupta Efficiency (KGE) for four models on two different forcing products. Right: KGE skill score with respect to the unconstrained LSTM