



Comprehensive evaluation on the robustness of deep learning based rainfall-runoff models under climate change scenarios

Aldo A. Tapia¹ and Andrew Bennett¹

¹Department of Hydrology and Atmospheric Sciences, The University of Arizona, 1133 James E. Roger Way, Tucson, AZ 85721, United States

Correspondence: Aldo A. Tapia (aldotapia@arizona.edu)

Abstract. Deep learning rainfall-runoff models have shown the ability to reproduce the observed hydrograph exceeding the performance of physics-based models, but their reliability outside the training distribution is still an open question. Most robustness evaluation have focused on a single architecture (LSTM), and have used controlled forcing perturbations (temperature offset or precipitation scaling) to probe the non-stationarity condition. In this study, we evaluate the robustness of five deep learning architectures (LSTM, minLSTM, minGRU, MLP-Mixer, Transformer) against a re-calibrated physics-based benchmark (SAC-SMA + Snow-17) across 554 CAMELS catchments, driven by a six-Global Change Model (GCM), four-Shared Socioeconomic Pathway (SSP) NEX-GDDP-CMIP6 and three-future periods ensemble. The results show that the deep learning models outperform the physics-based benchmark in terms of accuracy, but their robustness is limited under non-stationary conditions. The LSTM and minGRU architectures show the best performance among the deep learning models, while the MLP-Mixer and minLSTM architectures show the worst performance. We found that the predictive skill does not imply robustness: the LSTM, the most accurate architecture, is the least robust. Transformer, the most robust on annual volumes, fails to reproduce the intra-annual redistribution of streamflow, diverging from the physics-based benchmark in both directions, compensating that difference at annual basis. Architecture, not GCMs, SSPs nor period, is the dominant factor of divergence (45 % of the evaluated catchments). The results highlight the need for a more comprehensive evaluation of deep learning models under non-stationary conditions, and the importance of considering both accuracy and robustness in model selection.

1 Introduction

The hydrological cycle is composed of three main processes: evaporation, precipitation and runoff. Rainfall-runoff modeling is the transformation of the precipitation into runoff at a catchment outlet, which main objective is to reproduce the observed hydrograph or its streamflow series. This is a complex problem that is still not fully understood or modeled, making it a challenging task for hydrologists and data scientists alike. The streamflow predictions that such models produce feed flood forecasting and early-warning systems, reservoir operation and hydropower scheduling, irrigation and municipal-supply planning, drought monitoring, and the allocation of water rights and environmental flows. Since that data is used for decision-making, the reliability of the model is crucial, hence the value of a rainfall-runoff model is not only in how well it reproduces the observed



hydrograph but on how well captures the variability of the streamflow series under different conditions, places, seasons, and
25 climates.

Over the last decade, the rainfall-runoff modeling field has rapidly become more interested in the development of data-driven
models, in particular deep learning (DL) models, which are able to learn the meteorology-to-streamflow mapping directly from
large-sample datasets, with the Long-Short Term Memory Network (LSTM; a type of neural network) being the most popular
architecture in this field (Kratzert et al., 2018, 2019a). These models are able to capture the non-linearity and the complex
30 interactions between the meteorological forcing variables and the streamflow, and are able to generalize even in ungauged
basins (Kratzert et al., 2019b). However, these models are not without limitations. They are often criticized for being black-box
models, which are difficult to interpret, and for being sensitive to the choice of the model architecture, the hyperparameters, the
training data, and overfitting. They are also often criticize for not being able to capture the physical processes (Beven, 2020).

The standard LSTM is not the only data-driven model in the field, this recurrent architecture has been modified to improve its
35 performance and its ability to capture the physical processes (Kratzert et al., 2019c; Frame et al., 2022; Feng et al., 2024; De la
Fuente et al., 2024), and other architectures have been proposed to improve the performance of the modeling process (Vaswani
et al., 2017; Liu et al., 2024, 2025; Rasiya Koya and Roy, 2024; Sarkar et al., 2025). These models generally outperform the
physics-based models in terms of accuracy (Kratzert et al., 2018; Newman et al., 2017; Frame et al., 2022), but the question
is whether they are able to emulate or capture the physical processes from the *in-training* data distribution in a robust manner.
40 The ability of the model to generalize to *out-of-training* data is known as model robustness.

Robustness has been central in the evaluation of rainfall-runoff models, formalized through the differential split-sample test
and its successors, which calibrate on one hydroclimatic period and validate on a contrasting one. These tests have repeatedly
exposed parameter and structural drift under non-stationary conditions (Klemeš, 1986; Coron et al., 2012; Thirel et al., 2015).
In the case of data-driven models, whose flexibility is precisely what allows them to fit the historical records effectively,
45 the ability to extrapolate outside the training distribution is of concern. Recent work supports this hypothesis: process-based
models remain competitive or superior when the validation climate departs strongly from calibration (O et al., 2020); stand-
alone LSTMs can produce physically implausible responses to warming, including unrealistic changes in the runoff coefficient
(RC) (Wi and Steinschneider, 2022); LSTM's exhibits a concave runoff response to increasing precipitation, which above a
certain threshold the RC starts to decrease when precipitation increases (Baste et al., 2025). In-training accuracy, in short, is
50 necessary but not sufficient for the out-of-training reliability that climate-impact assessment actually requires.

Assessing this reliability is complicated by the absence of a ground truth: future streamflow cannot be observed, therefore
a data-driven model's projection can only be judged against a physically based reference. Yet that reference is itself imperfect
under non-stationary forcing, for example conceptual models calibrated on a wet period systematically bias a dry period and
vice versa (Coron et al., 2012). A defensible robustness assessment must therefore hold everything but the model family fixed—
55 the same catchments, the same forcing, the same evaluation—and compare each model against its own historical baseline so
that its static bias cancels and only its response to the climate signal is judged.

Existing robustness assessments of deep learning hydrological models fall short on these counts in three recurring ways,
which together define the gap this study addresses. First, they almost always evaluate a single architecture—typically the



LSTM—in isolation, consequently it cannot separate robustness that is intrinsic to the data from robustness that is a property of the architecture. Second, they probe the non-stationarity condition mainly through controlled forcing perturbations (a uniform temperature offset or a precipitation scaling) rather than a spatially coherent climate-change ensemble, thus the forcing they test is internally consistent but not physically realistic. Third, they report aggregate or annual summary statistics and rarely resolve the seasonal timing of streamflow. This study fills that gap with a continental-scale, multi-architecture evaluation: five DL architectures (including different model families) are compared against Sacramento Soil Moisture Accounting Model (SAC-SMA) + Snow Accumulation and Ablation Model (SNOW-17) benchmark across 554 CAMELS catchments, driven by a six-Global Change models (GCM) and four-Shared Socioeconomic Pathway (SSP) NEX-GDDP-CMIP6 ensemble, to characterize the robustness of the models under future climate conditions. The objective of this study is to evaluate if the robustness of the data-driven models is a property of the data or of the architecture, and if the robustness extends to the intra-annual redistribution of streamflow. Taken together, these evaluations will answer the question of whether deep learning models are robust in non-stationary conditions, and whether they can be trusted to project streamflow under a climate they were never trained on.

2 Material and methods

2.1 Modeling framework

The dataset used in this study is Catchment Attributes for Large-Sample Studies, CAMELS (Newman et al., 2017; Addor et al., 2017), a benchmark dataset of 671 minimally disturbed headwater basins in the contiguous United States (CONUS). The minimal amount of human intervention is a proxy for near-natural hydrological response, thus that observed streamflow can be interpreted as a response to meteorological forcing. The catchments span a wide range of hydroclimatic and physiographic conditions, from snow-dominated catchments in the Pacific Northwest to arid basins in the Southwest. Therefore the catchments included in this dataset spans from energy-limited to water-limited Budyko regimes, and from rain-dominated to snow-dominated runoff-generation mechanisms. In addition, the dataset includes three different meteorological forcing products (Maurer, Daymet, NLDAS-2) and a rich set of static catchment attributes (climate, topography, vegetation, geology).

CAMELS is also considered a community benchmark for evaluating rainfall-runoff models, including a physically-based benchmark – Sacramento Soil Moisture Accounting model (SAC-SMA) (Burnash et al., 1973) coupled with the Snow-17 snow module (Anderson, 1973) – used commonly in hydrology, allowing a controlled comparison between data-driven and a physics-based model. Although SAC-SMA + Snow17 model results and parameter values are included in the dataset, we re-calibrated the model to match the data-driven models' training and evaluation conditions, to make a fair comparison between the two model families.

From 671 catchments availables, we selected 554 catchments based on the following criteria: (i) catchments that are present in all three meteorological forcing products (Maurer, Daymet, NLDAS-2) and in the USGS streamflow record; (ii) catchments with gap-free observed streamflow series over the study period; and (iii) catchments with drainage area below 10,000 km², excluding the largest basins for which the lumped, single-outlet modeling assumption is least appropriate. The filtering criteria



ensure that every model family is trained, calibrated, and evaluated on an identical set of basins, enabling a controlled cross-model comparison.

For the physics-based benchmark, we re-calibrated the SAC-SMA + Snow-17 coupled model on the same catchments sample
95 and training period as the deep learning models. For the data-driven models, we selected a set of different deep learning architectures (LSTM, minLSTM, minGRU, MLP-Mixer, Transformer) which are described in detail in the following sections. The training and evaluation protocol is identical for all deep learning architectures, as well as the sequence length, input/output interface, and loss function. The hyperparameters of each architecture are tuned independently via ASHA (Asynchronous Successive Halving) on the validation set. The sequence length is set to 360 days, which is long enough to capture seasonal
100 and interannual variability in streamflow response. This length differs from other studies, where a 365-day sequence is used (Kratzert et al., 2019a), however this choice makes the sequence length divisible by the patch sizes evaluated in the MLP-Mixer architecture.

The input features include 15 dynamic forcing variables: maximum and minimum daily temperature ($^{\circ}\text{C}$), daily precipitation (mm), daily downwelling shortwave radiation (W m^{-2}), and daily mean vapor pressure (Pa), each replicated across three
105 lumped forcing data sources (Maurer, Daymet, NLDAS-2). For SAC-SMA + Snow-17 only the Daymet product is used.

We used the 19 static features encode climate (mean precipitation `p_mean`, mean potential evapotranspiration `pet_mean`, aridity index, fraction of precipitation as snow `frac_snow`, high- and low-precipitation frequency and duration), topography and geometry (mean elevation, mean slope, catchment area), vegetation (forest fraction, maximum and seasonal-range leaf-area index `lai_max/lai_diff`, maximum and seasonal-range green-vegetation fraction `gvf_max/gvf_diff`), and geology
110 and hydrology (geological permeability, baseflow index, carbonate-rock fraction). The typewritten text represents the name of each variable in CAMELS dataset.

The target variable is one-day-ahead area-normalized streamflow (mm d^{-1}). The dynamic inputs and static attributes are standardized globally across the catchment ensemble (global z -scoring using training-period statistics). The same statistics are reused at validation, test, and inference time in order to avoid data leakage.

115 The chronological split for the deep learning models is 1 Oct 1999 – 30 Sep 2008 for training, 1 Oct 1980 – 30 Sep 1989 for validation, and 1 Oct 1989 – 30 Sep 1999 for testing. The training period is chosen to be long enough to capture interannual variability in streamflow response, while the validation and test periods are chosen to be of comparable length to the training period. For the case of SAC-SMA + Snow 17 model, the training period considers the training and validation periods of the deep learning models, and the test period is the same as the deep learning models.

120 The performance metric used to evaluate the predictive skill of the models is the Kling–Gupta efficiency (KGE) (Gupta et al., 2009), which is a widely used metric in hydrology that balances correlation, bias, and variability. The KGE is computed per catchment over the test period and used to rank the model families. This metric is defined as follows:

$$\text{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}, \quad (1)$$



where r is the Pearson correlation between simulated and observed streamflow, α the ratio of their standard deviations ($\frac{\sigma_{sim}}{\sigma_{obs}}$, variability bias), and β the ratio of their means ($\frac{\mu_{sim}}{\mu_{obs}}$, volume bias). KGE = 1 is a perfect simulation; the commonly cited reference value KGE ≈ -0.41 marks the skill of the mean-flow benchmark (Knoben et al., 2019).

2.2 Physics-based benchmark

As previously stated, the benchmark is the lumped Sacramento Soil Moisture Accounting model coupled with the Snow-17 temperature-index snow module (Burnash et al., 1973; Anderson, 1973). The coupled model operates at a daily time step and comprises three sequential components: (i) Snow-17 precipitation partitioning and snowmelt, (ii) SAC-SMA soil-moisture accounting, and (iii) unit hydrograph routing. Throughout, we use *rainfall-runoff model* in its conventional sense to denote the complete precipitation-to-streamflow mapping at the catchment outlet: SAC-SMA proper performs only soil-moisture accounting, but the term SAC-SMA + Snow-17 refers to this full chain: snow, soil-moisture accounting, evapotranspiration, and channel routing. The same processes are what the deep-learning rainfall-runoff models are trained to reproduce, therefore the two model families are compared on an identical input-to-streamflow task.

Since the forcings do not include the necessary dynamic variables to compute evapotranspiration using the aerodynamic forces, potential evapotranspiration (PET) is computed from temperature and net radiation via the Priestley–Taylor formulation (Priestley and Taylor, 1972), which is defined as follows:

$$PET_t = \frac{\alpha_{PT} \Delta_t}{\Delta_t + \gamma} \cdot \frac{R_t^n}{\lambda}, \quad (2)$$

where Δ_t is the slope of the saturation vapor pressure curve, γ the psychrometric constant, R_t^n net radiation (Allen et al., 1998), and λ is the latent heat of vaporization (2.45 MJ kg^{-1}). Ground heat flux is assumed negligible at the daily time step.

The calibration strategy follows the same parameters and bounds described in Addor et al. (2017), with two exceptions: (i) the objective function is Mean Square Error (MSE) to match deep learning loss functions, (ii) and we included 5 super-parameters to calibrate the Snow Areal Depletion Curve (ADC) using two Beta CDF distributions. More details about can be found in Appendix A.

These parameters are calibrated independently per catchment by using Shuffled Complex Evolution (SCE-UA) algorithm (Duan et al., 1992), implemented in SPOTpy (Houska et al., 2015). Following Addor et al. (2017), the optimization was conducted using 10 different random seeds per catchment to reduce dependence on local optima, and up to 150 catchments were calibrated in parallel using an translated and optimized version of SAC-SMA + Snow-17 in Python, where the I/O bottleneck between the wrapper and the model is removed. The initial states are obtained by iterative spin-up: the first year of forcings is recycled until the six SAC-SMA state variables (UZTWC, UZFWC, LZTWC, LZFSC, LZFPC, ADIMC) converge to within 10^{-3} of their final values or after 1000 iterations, and then the calibration continues until the maximum number of function evaluations is reached (200,000). Due to the translated and optimized implementation, the calibration procedure is significantly faster than the original implementation, allowing for a more efficient exploration of the parameter space, a more robust calibration of the model, and the addition of the new 5 super-parameters for the ADC calibration.



2.2.1 Asymmetry with the deep learning models

There is one difference between the SAC-SMA + Snow-17 and the deep learning models that is relevant to mention. In SAC-SMA + Snow-17, temperature enters to the model only through PET and Snow-17 partitioning and melt, not as a direct input. Deep learning models, by contrast, receive raw $T^{\max}, T^{\min}$ directly.

160 2.3 Deep learning architectures

Five deep learning architectures were evaluated. There is one architecture that is based only on dense layers, three that use recurrent layers, and one that uses attention layers. All architectures share the same input/output interface (360-day sequence of 34 features into 1-day-ahead, area-normalized streamflow), the same loss (MSE), the same optimizer (Adam with OneCycleLR scheduler, $\text{peak} = 1.2 \times \text{lr}$), and are tuned with ASHA via Ray Tune (max epochs of 200, grace period of 10, reduction factor
165 of 3), following the previously mentioned temporal splits.

Recurrent networks are the most common architecture used in rainfall-runoff modeling, since they handle sequential data in a way that is consistent with the temporal nature of the problem. The current state-of-the-art architecture for hydrology is the Long-Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), which is a recurrent architecture that has been widely used (Kratzert et al., 2018, 2019a). The LSTM has proven to be effective in capturing long-term dependencies in time series
170 data, making it suitable for rainfall-runoff modeling (Kratzert et al., 2018). The key core of an LSTM is the cell state, which allows information to flow across time steps without being overwritten, enabling the network to retain relevant information over long sequences. The LSTM architecture also includes gating mechanisms (input, forget, and output gates) that control the flow of information into and out of the cell state, allowing the network to selectively update and retrieve information as needed (Hochreiter and Schmidhuber, 1997).

175 However, LSTM has some limitations that have motivated the development of alternative architectures. One limitation of the LSTM is the requirement of sequential processing of the data, which can be computationally expensive for long sequences and makes the training-time a bottleneck. This limitation has motivated the development of alternative architectures that can handle sequential data directly in parallel, such as minLSTM (minimal LSTM) and minGRU (minimal Gated Recurrent Unit) (Feng et al., 2024).

180 Both minimal versions of LSTM and GRU are designed to reduce the computational complexity of the original architectures by removing the dependence of the gates on the previous hidden state, allowing for parallel processing of the sequence. This characteristic enables faster training times and can be particularly beneficial for long sequences, such as those encountered in rainfall-runoff modeling. In addition, the cell architecture is simplified, which can lead to a reduction in the number of parameters and potentially improve generalization. In the case of minLSTM, the output gate is removed, and the hidden state is
185 collapsed into the cell state, while in minGRU, the reset gate is removed, and the update gate is simplified. These changes can lead to a more efficient architecture that retains the ability to capture long-term dependencies in sequential data. The difference between the LSTM, minLSTM, and minGRU architectures is illustrated in Fig. 1.

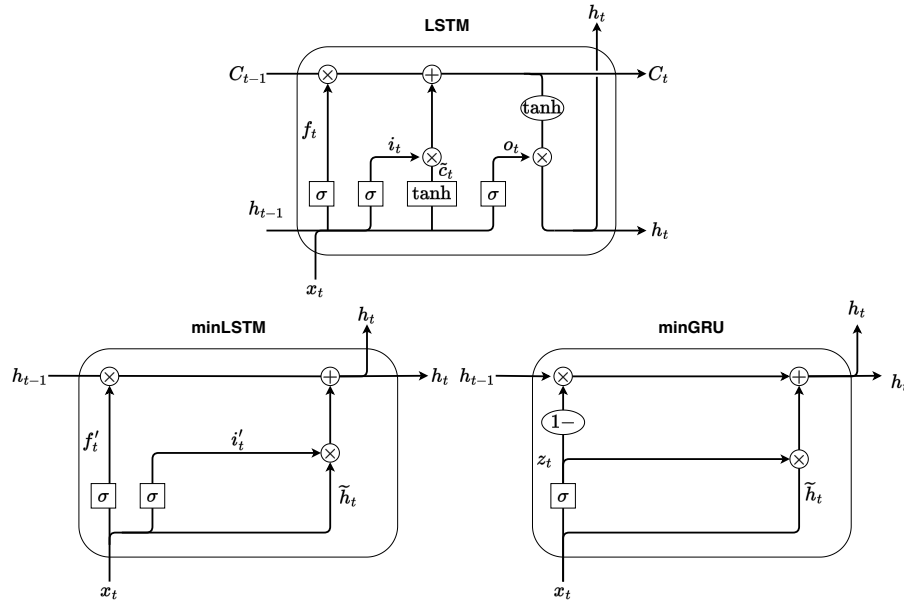


Figure 1. Cell architectures of LSTM, minLSTM, and minGRU. The LSTM architecture includes input (i_t), forget (f_t), and output gates (o_t) that control the flow of information into and out of the cell state. The minLSTM architecture removes the output gate and collapses the hidden state into the cell state, while the minGRU architecture removes the reset gate and simplifies the update gate.

These architectures have not been tested in hydrology, and it is unclear how they would perform in this context. The paper by Feng et al. (2024) provides a theoretical analysis of the minLSTM and minGRU architectures, but it does not provide empirical results on real-world datasets. Therefore, it is important to evaluate these architectures in the context of rainfall-runoff modeling to determine their effectiveness and potential advantages over traditional LSTM and GRU architectures.

We also include a Transformer architecture (Fig. 2), that consists only in the encoder structure of the original Transformer architecture (Vaswani et al., 2017), following the same setup Liu et al. (2024) tested in rainfall-runoff modeling (with a time embedding layer). The Transformer architecture has been widely used in natural language processing and computer vision, and it has shown to be effective in capturing long-term dependencies in sequential data. The characteristics of the Transformer architecture, such as the self-attention mechanism and the ability to process sequences in parallel, has motivated its application in rainfall-runoff modeling (Yin et al., 2022). Since it is an emerging architecture that has not been as widely used in hydrology as the LSTM, the inclusion of this architecture in this study allows us to evaluate its robustness and performance in the context of non-stationary conditions, and to compare it with the more established recurrent architectures.

Finally, we have included the MLP-Mixer architecture (Tolstikhin et al., 2021), which consists of a series of blocks or patches that are processed independently and then combined to produce the final output (Fig. 3). The model consists in blocks of two Multi-Layer perceptrons (MLPs) that are applied sequentially to the input data, transposing the first MLP to mix information across time steps (token-mixing MLP) to the second MLP to mix information across features (channel-mixing MLP). This is a

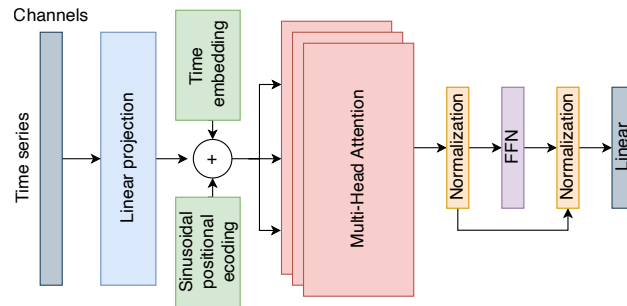


Figure 2. Transformer architecture, including layer projection, multi-head attention, and feed-forward layers.

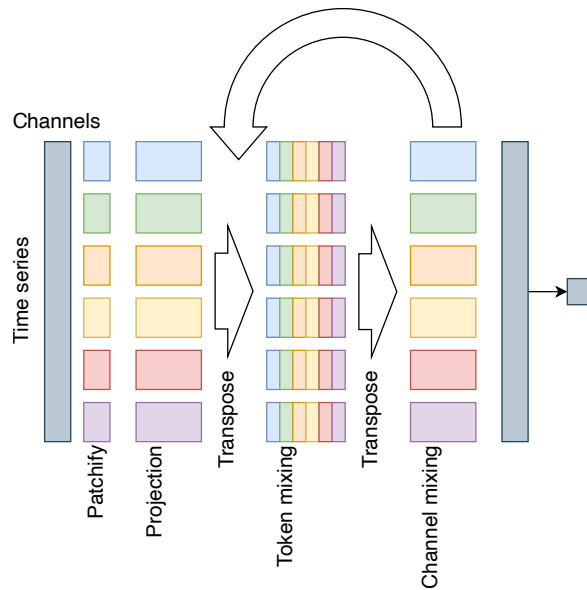


Figure 3. MLP-Mixer architecture, including token-mixing and channel-mixing MLPs.

“no attention” model, without gating mechanisms, no sequential processing, and no recurrence. It is a simple architecture that
 205 can be trained efficiently and can capture complex interactions between the input features. The MLP-Mixer architecture has
 been applied to time series forecasting tasks outside the hydrology field (Ekambaram et al., 2023; Zhong et al., 2024; Hong
 et al., 2025)

2.4 Non-stationarity conditions

Since the models were trained using CAMELS dataset, we follow the same approach as the previous studies, by using only
 210 a limited time period to evaluate the robustness of the models. In order to evaluate the robustness of the models under non-



stationary conditions, we have designed an experiment to evaluate the robustness of the models under non-stationary conditions using CMIP6 Global Change models.

The NEX-GDDP-CMIP6 dataset is a downscaled climate projection dataset that provides daily climate variables at a high spatial resolution ($0.25^\circ \times 0.25^\circ$) for an historical period (1950-2014) and future projections (2015-2100) under four Shared Socioeconomic Pathway scenarios. The dataset includes outputs from 35 CMIP6 GCM models, which have been statistically downscaled using the Bias-Correction Spatial Disaggregation (BCSD) method. The future projections are based on four SSP scenarios: SSP1-2.6, SSP2-4.5, SSP3-7.0, and SSP5-8.5, which represent different levels of greenhouse gas emissions and socio-economic development. Each SPP scenario represents a different trajectory of future climate conditions, which is a non-stationary condition by itself.

For this work, we used a subset of the available models, which includes ACCESS-CM2, CanESM5, GFDL-ESM4, GISS-E2-1-G, MPI-ESM1-2-HR, and MRI-ESM2-0. Although the use of one model would be enough to evaluate the robustness of the models under non-stationary conditions, we have used a subset of six models to extend the evaluation of the sensitivity of the models to the different climate projections.

Each model provides daily values for the following variables: maximum and minimum daily temperature, daily precipitation, daily downwelling shortwave radiation, daily mean vapor pressure, relative humidity, specific humidity, downwelling longwave radiation, and surface wind speed. Since the vapor pressure is not included in the dataset, we have computed it from the saturated vapor pressure (e_s) and the relative humidity (RH).

We applied a bias correction to the NEX-GDDP-CMIP6 projections to ensure consistency with CAMELS dataset used to train the models. For precipitation, we used Local Intensity Scaling (LOCI) for precipitation to avoid distortion of the climate-change signal (Cannon et al., 2015) with a threshold of 1 mm day^{-1} for value scaling. For the remaining variables, we followed the approach of Ballarin et al. (2023), applying Quantile Delta Mapping (QDM), which corrects systematic biases while explicitly preserving the model-projected changes in the quantiles of the distribution. The bias correction was applied to the historical period where both, the NEX-GDDP-CMIP6 dataset and the CAMELS dataset, overlap (1980–2014), and then the same correction parameters were applied to the future projections.

Model behavior was evaluated over a historical reference period (1950–2014), common to all GCMs. Each future projection was divided in 3 different periods: Period 1 (2015–2045), Period 2 (2045–2070), and Period 3 (2071–2100). The divergence of some forcings from the historical period is accounted in each period.

2.5 Robustness evaluation

2.5.1 Perturbation experiments

Before evaluating the models under the physically coherent NEX-GDDP-CMIP6 projections (Sect. 2.4), we first probe their behavior with a controlled grid of idealized precipitation and temperature perturbations. Since climate elasticity follows a non-linear behavior and is regime-dependent (Sankarasubramanian et al., 2001; Vano et al., 2012), we expect to evidence this for the physically-based model; while for the DL architectures, we could obtain an implausible streamflow increase under



warming conditions (Wi and Steinschneider, 2022), a streamflow reduction while precipitation increases (Baste et al., 2025),
 245 or any unrealistic result that comes from these architectures. Although the robustness evaluation is based on non-stationary
 conditions under global change scenarios, we opted to include this analysis to follow the same research approach that other
 authors have used to evaluate DL architectures in hydrology. Since the focus of this work is the evaluation of robustness under
 non-stationary conditions, we are limiting the analysis to a representative catchment for each macro-level geographic region in
 the United States watershed classification system (2-digit Hydrologic Unit Code or HUC-2). The selection of these catchments
 250 is based on the highest mean KGE across models, prioritizing a well-represented catchment to ensure fairness under this
 approach.

The perturbation experiment keeps every model fixed and alters only its inputs, so the response reflects the learned (for the
 deep learning models) or encoded (for SAC-SMA) input-output, anticipating what we could obtain from the non-stationary
 conditions.

255 Formally, let $\mathbf{X} = (\mathbf{p}, \mathbf{t}_{\min}, \mathbf{t}_{\max}, \dots)$ be the matrix of input forcings over the evaluation window, where $\mathbf{p}$ is precipitation,
 $\mathbf{t}_{\min}$ and $\mathbf{t}_{\max}$ are the minimum and maximum temperature, and the ellipsis represents the remaining forcings, which are left
 unchanged. A perturbation vector δ applied to the inputs produces a change in any scalar streamflow signature ϕ :

$$\Delta\phi(\delta) = \phi(f_{\theta}(\mathbf{X}(\delta))) - \phi(f_{\theta}(\mathbf{X})), \quad (3)$$

where f_{θ} is a fixed (already trained or calibrated) model, ϕ maps the simulated streamflow to a scalar (mean, median, or
 260 a parameter of interest), and $\mathbf{X}(\delta)$ is the perturbed input forcing matrix. We perturb precipitation, temperature, and both
 variables jointly, keeping all other forcings at their observed values so that the response is attributable to P and T alone.

Precipitation is perturbed multiplicatively,

$$\mathbf{p}' = \delta_P \mathbf{p}, \quad (4)$$

where δ_P takes 11 values from 0.0 to 2.0 in increments of 0.2, and temperature is perturbed additively on both the minimum
 265 and the maximum,

$$\mathbf{t}'_{\min} = \mathbf{t}_{\min} + \delta_T, \quad \mathbf{t}'_{\max} = \mathbf{t}_{\max} + \delta_T, \quad (5)$$

with δ_T ranging in 2°C steps from -10 to 10°C . The combined perturbation applies both simultaneously, yielding $\mathbf{X}' =$
 $(\mathbf{p}', \mathbf{t}'_{\min}, \mathbf{t}'_{\max}, \dots)$ over the full 11×11 grid of (δ_P, δ_T) pairs. As noted in Sect. 2.2.1, the deep learning models receive
 the perturbed $\mathbf{t}'_{\min}$ and $\mathbf{t}'_{\max}$ directly, whereas in SAC-SMA + Snow-17 temperature is not a direct input: the perturbation
 270 propagates through the Snow-17 rain/snow partitioning and melt, and through the Priestley–Taylor potential evapotranspiration,
 $\text{PET}' = \text{PET}(\mathbf{t}'_{\max}, \mathbf{t}'_{\min}, \dots)$, keeping radiation and the remaining PET inputs fixed.

To make the response comparable across models and catchments, we characterize it through the fractional change in the
 streamflow signature relative to the unperturbed reference,

$$\Lambda_{\delta_P, \delta_T}^{\phi} = \frac{\phi(\hat{\mathbf{q}}_{\delta_P, \delta_T})}{\phi(\hat{\mathbf{q}}_0)}, \quad (6)$$



275 where $\hat{q}_0$ is the simulated streamflow at reference conditions ($\delta_P = 1, \delta_T = 0$) and $\hat{q}_{\delta_P, \delta_T}$ the streamflow under the perturbation. This is the same within-model normalization used later for the CMIP6 experiments (Eq. (9)), so that each model's systematic bias cancels and only its response is compared. We report one scalar signature for this analysis: the mean, $\Lambda_{\delta_P, \delta_T}^\mu$, which summarizes the change in water volume expecting to produce a two-dimensional response surface exhibiting a non-linear behavior.

280 Because the elasticity structure depends on the intra-annual catchment processes, a snow-dominated catchment responds to a temperature perturbation very differently in winter than during the summer, and differently against a rain-dominated catchment. Using the test period (1 Oct 1989 – 30 Sep 1999) and 360-day perturbed input sequences, the response is evaluated year-round, during summer and winter (July-September and January-March, respectively) together with the SAC-SMA + Snow-17 benchmark. As in the CMIP6 experiments, the fractional change Λ of each deep learning model is contrasted against
285 that of SAC-SMA + Snow-17, therefore any unexpected behavior can be traced back to a specific perturbation regime. Because the deep learning models have no explicit representation of snow accumulation and melt, we expect their response to the temperature perturbation to diverge most from the physics-based benchmark in snow-dominated catchments and during the cold-season subsets.

2.5.2 Evaluation over testing period

290 The testing period is the same for all models, and spans from 1 Oct 1989 to 30 Sep 1999. Although it is a relatively short period, it is long enough to capture interannual variability in streamflow response. Following the Differential Split-Sample Test (DSST) used in previous studies, the available period is organized by Hydrological Year (1 Oct y through 30 Sep $y + 1$) and categorized into wet, normal, and dry years based on the annual precipitation of that period, and cold, normal, and warm years based on the annual mean temperature of that period. For this period, only 8 years are available to evaluate the robustness of the
295 models for the same dataset (1989-1999). The procedure is to select the 3 wettest and 3 driest years, and keep the remaining 2 years as normal conditions. The evaluation is performed by computing the PBIAS of the models on the testing period, and then comparing the PBIAS of the models on the wet and dry years against the PBIAS on the normal years. The PBIAS is computed as follows:

$$\text{PBIAS} = 100 \cdot \frac{\sum_{t=1}^n (\hat{q}_t - q_t)}{\sum_{t=1}^n q_t}, \quad (7)$$

300 where $\hat{q}_t$ is the simulated streamflow at time t , q_t is the observed streamflow at time t , and n is the number of time steps in the evaluation period. This section is the only one than compares the models against the observed streamflow, the following sections compare the models against the SAC-SMA benchmark.



2.5.3 Water-balance: the runoff ratio

PBIAS measures *relative* divergence between two models but is blind to whether either model's projection is physically ad-
305 missible in absolute terms. As an independent plausibility check we compute the annual runoff ratio:

$$RC_y^c = \frac{\bar{q}_y^c}{\bar{p}_y^c}, \quad RC^c = \sum_{i=1}^y RC_i^c, \quad (8)$$

where $\bar{q}_y^c$ is the mean annual streamflow and $\bar{p}_y^c$ the mean annual precipitation (forcing bias-corrected with Daymet) in catch-
ment c for year y . RC^c represents the runoff coefficient of the analyzed period. The runoff ratio is affected not only by the
model's internal water balance, but also by the climate forcing, since p^c is an input to the model. Also, other factors such as
310 PET and soil moisture storage can affect the runoff ratio. The RC^c should not be constant through time, since it is affected by
the climate forcing, but it should be consistent year to year, and consistent with the physics-based benchmark. The runoff ratio
was computed over the historical period of the bias-corrected GCM datasets to evaluate the effect of the correction over the
models. A physically consistent model must satisfy $RC \in [0, 1]$, unless there are inter-basin transfers or multi-decadal storage
trends, since runoff cannot exceed precipitation.

315 Although the runoff ratio varies from year to year, the mean runoff ratio over the historical period is a useful reference
point. As a way to quantify the difference between the ML models and SAC-SMA over the historical period of the CMIP6
GMCM, we computed the mean runoff ratio over the historical period for each model and catchment, and compared it against
the SAC-SMA benchmark.

2.5.4 Fractional change relative to historical

320 Because each model carries its own biases, historical and future signatures are computed with the same model, so that model-
specific bias cancels in the comparison. Robustness is then assessed by comparing each model's fractional change in signatures
against that of SAC-SMA, taken as the weak benchmark.

For GCM (m), scenario (sce), period (per), and model architecture (DL or SAC-SMA), the fractional change is computed
as:

$$325 \quad \Lambda_{m,sce,per,model}^\phi = \frac{\phi(\hat{q}_{m,sce,per,model})}{\phi(\hat{q}_{m,hist,model})}, \quad (9)$$

where ϕ is a streamflow parameter, $\hat{q}_{m,sce,model}$ the future-period simulated streamflow vector, and $\hat{q}_{m,hist,model}$ the historical-
period simulated streamflow vector. The parameter ϕ can represent any signature, and for this study the period-mean streamflow
or the day-of-year climatology are evaluated, so that Λ can be evaluated on both. Hence, the fractional change Λ is a *within-*
model comparison between the historical and each SSP scenario and period. The fractional change Λ is therefore a measure of
330 the model's response to non-stationary forcing, independent of its calibration skill.



2.5.5 Divergence from SAC-SMA: PBIAS-based robustness rating

To assess robustness, the DL-model response is compared against the SAC-SMA benchmark. The divergence is quantified by the percent bias (PBIAS) of the DL-model fractional change relative to SAC-SMA,

$$\text{PBIAS}_{m,\text{sce,per,DL}}^{\phi} = 100 \cdot \left(\Lambda_{m,\text{sce,per,DL}}^{\phi} - \Lambda_{m,\text{sce,per,SAC-SMA}}^{\phi} \right), \quad (10)$$

expressed in percentage points. Note that this PBIAS is on fractional changes (Λ), not on flow magnitudes themselves; it measures how differently the DL model and SAC-SMA respond to the same climate forcing. Hence, the PBIAS is normalized across catchments, independent of the flow magnitude.

There is no established threshold for the divergence of a climate-change signal between two models, so we define four robustness bands ourselves. The cut points are chosen for interpretability and following the same vocabulary of Moriasi et al. (2007) and Moriasi et al. (2015), but they are not taken from those studies as a validated criterion. The four robustness ratings are assigned as follows:

Table 1. Robustness rating based on PBIAS^{ϕ} of the fractional change (Λ) relative to SAC-SMA.

$\text{PBIAS}^{\phi} (\%)$	Robustness rating
$ x < 5\%$	very robust
$5\% \leq x < 10\%$	robust
$10\% \leq x < 15\%$	marginally robust
$ x \geq 15\%$	non-robust

These bands are a qualitative, interpretable scale, not a validated criterion. The Moriasi thresholds resemble were derived for simulated-versus-observed streamflow during calibration, whereas here PBIAS measures a model-versus-model difference in a fractional change signal ($\Lambda_{\text{DL}} - \Lambda_{\text{SAC-SMA}}$, in percentage points), for which no established threshold exists. The specific values are therefore our own choice; the upper 15 % limit denotes a reasonable threshold to define if a model is robust or not, and the remainder limits add a qualitative interpretation to distinguish between these arbitrary levels. The thresholds are applied to the absolute value of PBIAS, so that both over- and under-prediction relative to SAC-SMA are treated symmetrically.

A general robustness evaluation framework is agnostic to the choice of streamflow signature ϕ . In this study we focus on two complementary signatures: (i) the period-mean streamflow, which collapses each future window to a single scalar per catchment, and (ii) the day-of-year climatology, which resolves intra-annual timing and seasonality. The latter is more demanding, daily timing and hydrograph-shape differences accumulate over the year, and could be partially canceled in the annual mean. We can expect that long term shifts are captured by the period-mean signature, but that the day-of-year signature will reveal additional divergence in the timing and shape of the hydrograph. The two signatures are therefore complementary, and both are evaluated in this study.



355 2.6 Statistical assessment

The fractional change (Δ) is a ratio, therefore it can take very large values when the historical flow in the denominator is close to zero, as happens in arid catchments and on low-flow days of the day-of-year climatology. Its distribution is therefore far from normal, which is a common feature of relative-change metrics in dry catchments (Anderson et al., 2024) rather than a problem specific to this design. For this reason, we keep PBIAS for the robustness rating, but use rank-based statistics for the analysis and report medians instead of means to avoid parameter inflation due to high-value outliers.

To test whether the divergence from SAC-SMA depends on the deep learning architecture, the GCM, the future period (Periods 1, 2 and 3), or the SSP scenario, we use the Kruskal–Wallis test (Kruskal and Wallis, 1952), a rank-based alternative to ANOVA that does not assume the data are normally distributed or that the groups have equal spread (Brown and Forsythe, 1974). Because the dataset is very large, almost any difference comes out as statistically significant, consequently we focus on the size of the effect rather than on p -values, using the rank-based effect size $\eta_H^2 = (H - k + 1)/(n - k)$ of Tomczak and Tomczak–Łukaszewska (2014), where H is the Kruskal–Wallis statistic, k the number of groups, and n the total number of observations. Differences between individual groups are then examined with Dunn’s test (Dunn, 1964; Dinno, 2015), using a Benjamini–Hochberg correction to control the false discovery rate across the multiple pairwise comparisons (Benjamini and Hochberg, 1995).

For each catchment, we run a Kruskal–Wallis test for each of the four factors in isolation and use the effect size η_H^2 to measure how much of the divergence that factor explains, following the same idea used to separate the sources of uncertainty in climate-projection ensembles (Chegwidden et al., 2019). The factor with the largest η_H^2 is taken as the dominant one for that catchment. This gives whether the architecture, the GCM, the period, or the scenario drives most of the divergence, which we then map across the country.

375 3 Results

3.1 Model performance on the CAMELS dataset

The performance of the evaluated models on the CAMELS dataset is summarized in Fig. 4, which shows the distribution of the Kling–Gupta efficiency (KGE) scores for each model over the test period, allowing a direct comparison of their predictive skill. Since the KGE jointly balances correlation, variability and volume bias (Eq. (1)), a higher score reflects an overall better reproduction of the observed hydrograph rather than skill in a single component. The recurrent architectures achieved the best cumulative performance, with the LSTM, minGRU and Transformer reaching the highest median KGE (0.785, 0.778 and 0.778, respectively), followed by the minLSTM (0.762). MLP-Mixer was the worst-performing model, with the lowest median KGE (0.759). The SAC-SMA model, which serves as the physically-based benchmark, reached a median KGE (0.749) comparable to the MLP-Mixer performance. The performance obtained for SAC-SMA in this study is higher than that previously reported for the CAMELS dataset (Newman et al., 2015, 2017), which is likely due to the re-calibration of the model with the inclusion of the 5 super-parameters for the Snow Areal Depletion Curve (ADC).

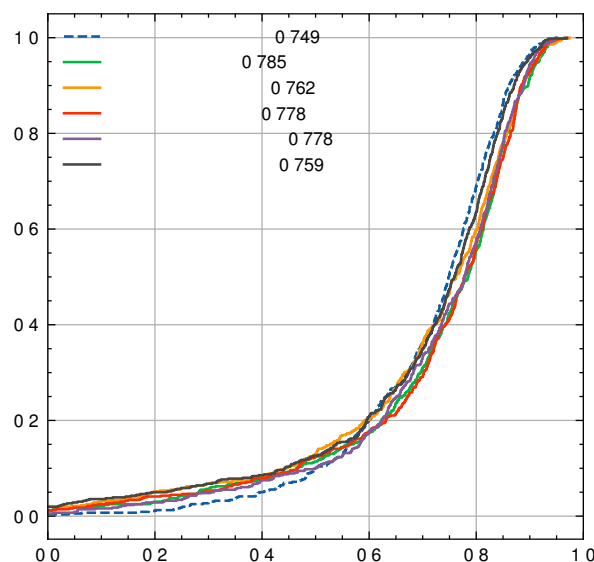


Figure 4. KGE distribution of the evaluated models on the CAMELS dataset. Median KGE is included in the legend.

3.2 Robustness evaluation

3.2.1 Perturbation experiments

Based on the established criteria, a set of 18 catchments were selected with a mean KGE from 0.72 to 0.92. The perturbation was applied to the evaluation dataset, which has 3 272 batches of 360-day periods, generating an output database of 395K modeled streamflow records for each catchment and model. Since the results are extensive, we are only including two catchments of different regime in the main text to evidence the trend among the catchments and the importance of the snow process in the streamflow response. The results of the remaining catchments are included in the Supplementary Material.

The two selected catchments are: gauge ID **10316500** (Lamoille CK NR Lamoille, NV. HUC-2 Great Basin Region) and **8025500** (Bayou Toro near Toro, LA. HUC-2 Texas-Gulf Region). The first catchment is a snow-dominated catchment, with a mean annual precipitation of 800 mm and a mean annual temperature of 3.5 °C. The second catchment is a rain-dominated catchment, with a mean annual precipitation of 1 450 mm and a mean annual temperature of 17.5 °C.

Figure 5 shows the mean snow water equivalent, mean potential evapotranspiration, and mean fractional change for catchments 10316500 and 8025500. In general, there are two trends: (i) and increase in precipitation leads to an increase in streamflow, and (ii) and increase in temperature leads to a decrease in streamflow. This behavior is consistent with the expected response of a catchment to climate change, although there is a caveat: the decrease in temperature leads to snowpack accumulation, which leads to a decrease in streamflow. This happens in the snow-dominated catchment, where the snowpack plays an important role in the streamflow response. In the rain-dominated catchment, the same decrease in temperature leads to a

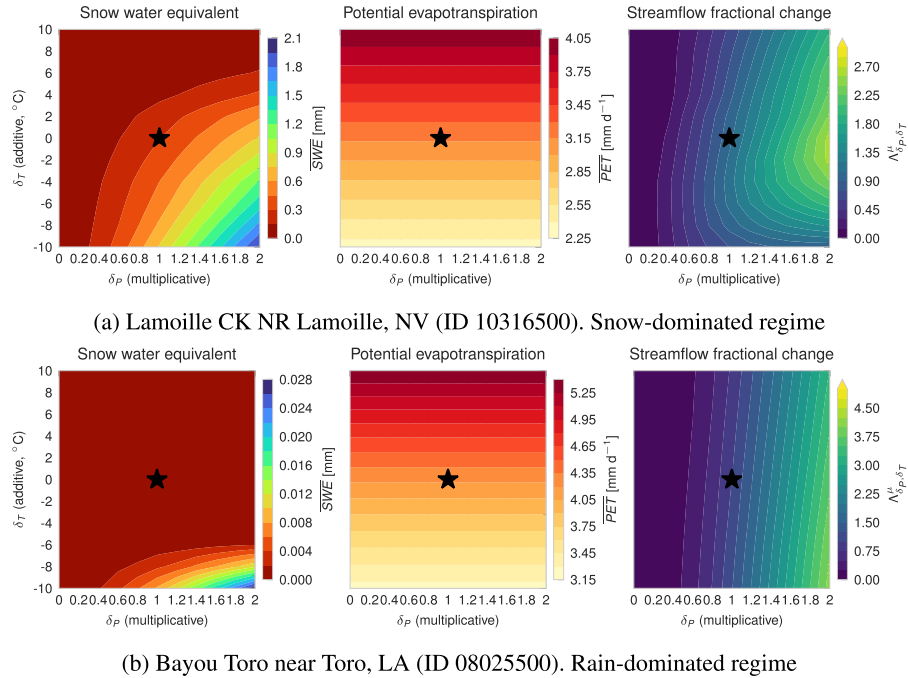


Figure 5. Mean snow water equivalent ($\overline{\text{SWE}}$), mean potential evapotranspiration ($\overline{\text{PET}}$), and mean fractional change ($\Lambda_{\delta_P, \delta_T}^\mu$) for the perturbation grid of precipitation and temperature. The star indicates the position of $\Lambda_{\delta_P, \delta_T}^\mu = 1$ ($\delta_P = 1$ and $\delta_T = 0$).

marginal accumulation of snow (~ 1 % of the snow-dominated catchment), which is not enough to affect significantly the streamflow response.

The seasonal analysis shows a completely different behavior (Fig. 6) for the catchment 10316500. As a remainder, each perturbation is applied to the entire 360-day period for each streamflow record, then the effect of the temperature and precipitation is affecting the previous modeled streamflow equally. During the winter, the streamflow response is minimal due to the temperature, then any positive change in streamflow affects the modeled streamflow increasing this response several times. During the summer, an increase in temperature leads to an anticipation of the freshet, reducing the streamflow response, while a decrease in temperature delays the freshet of spring. Both behaviors, although are opposite, are consistent with the expected response of a catchment to climate change, as some authors have reported (Sha et al., 2024).

For the catchment 8025500, the seasonal analysis shows a similar behavior than the yearly mean analysis, with a slight increase in streamflow during the summer, but following the same trend. Here, we demonstrate that the expected change in the streamflow depends largely of the regime of the catchment due to the snow processes represented by the model SNOW-17, a process that is not represented in the other models.

Figure 7 shows the model comparison of the transects of the perturbation of each factor with the other one fixed at normal conditions ($\delta_P = 1$ and $\delta_T = 0$) for catchment 10316500 and 8025500, respectively. There is some sort of agreement between

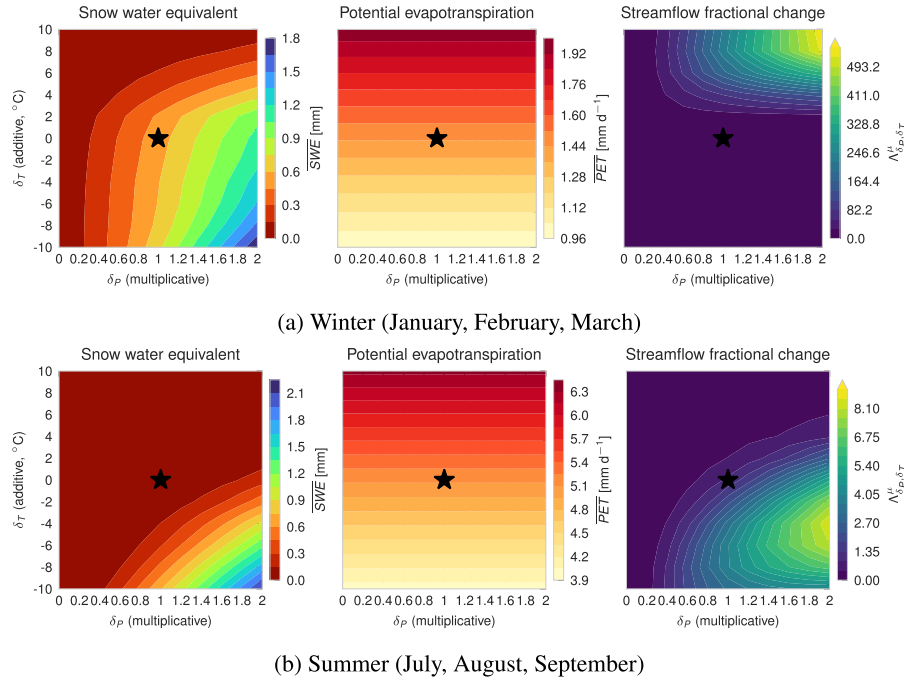


Figure 6. Mean snow water equivalent ($\overline{\text{SWE}}$), mean potential evapotranspiration ($\overline{\text{PET}}$), and mean fractional change ($\Lambda^{\mu}_{\delta_P, \delta_T}$) for the perturbation grid of precipitation and temperature of catchment 10316500, split by season. The star indicates the position of $\Lambda^{\mu}_{\delta_P, \delta_T} = 1$ ($\delta_P = 1$ and $\delta_T = 0$).

the models, exposing the same general trend of the fractional change, although the magnitude of the response is different. There are several points to highlight:

- For the snow-dominated catchment (10316500), when $\delta_P = 0$ the DL models are exhibiting a response up to $\sim 40\%$ of the response compared with $\delta_P = 1$, while SAC-SMA is exhibiting a response very close to 0% . The DL models assumes the snowpack is still present even after 360 days without precipitation, while SAC-SMA does not.
- Although the response turns close to a linear form after a threshold for the precipitation perturbation, the slope is different. DL models show a lower the response with increased precipitation.
- There is a large divergence of the response for the temperature perturbation, with the DL models exhibiting a response that does not follow SAC-SMA behavior. In general, the curves trend are similar, but the magnitude of the response is different, with the LSTM being the least sensitive and minGRU being the most sensitive.
- The precipitation perturbation is well represented by the DL models for the rain-dominated catchment (8025500), all the models exhibit a response close to the SAC-SMA model with a $\delta_P = 0$. There is a divergence with $\delta_P > 1$, but the magnitude of the response is similar to the SAC-SMA model.

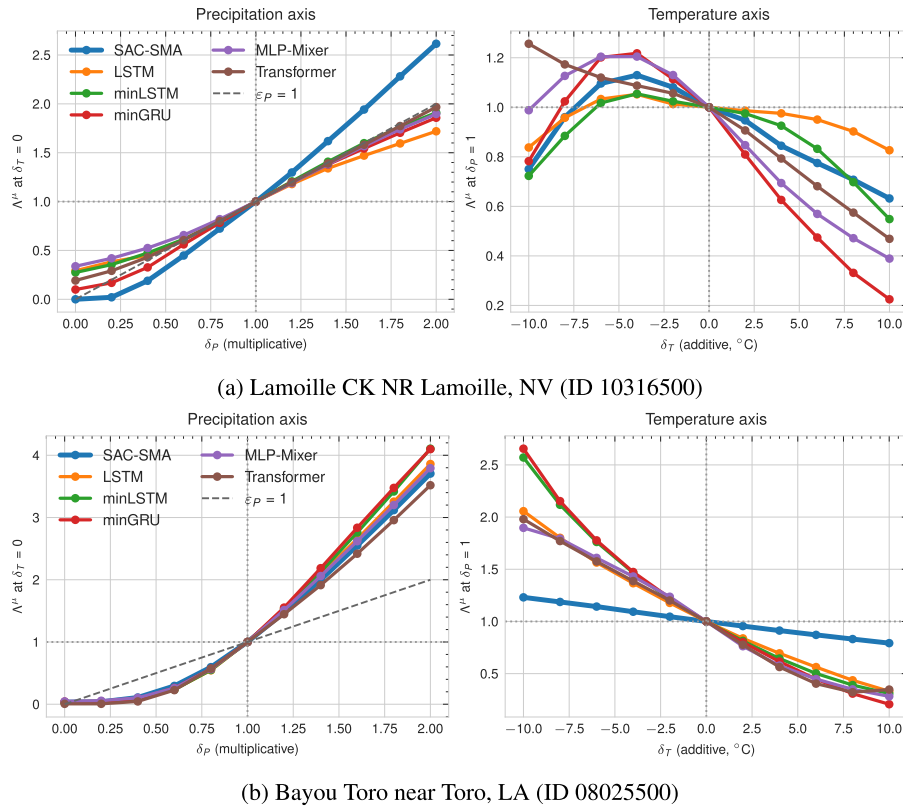


Figure 7. Transects of the perturbation of each factor with the other one fixed at normal conditions ($\delta_P = 1$ and $\delta_T = 0$) for catchments 10316500 and 8025500. The origin on the dotted lines axes indicates the position of $\Lambda^\mu_{\delta_P, \delta_T} = 1$ ($\delta_P = 1$ and $\delta_T = 0$). The dashed line represents the precipitation perturbation in the same plot as a way to facilitate the comparison with the streamflow change

- For the temperature perturbation, the behavior is quite different for this catchment. Since this one does not build up snowpack, the response of the temperature perturbation is closely linked to the change in the potential evapotranspiration. The DL models exhibiting a response larger to the same temperature change than the SAC-SMA model.

435 All this evidence indicates that the DL models are not robust enough to mimic the behavior of the SAC-SMA model, due to the non-linearity of the response of the snow-dominated catchment and possibly overfitting on the temperature perturbation for the rain-dominated catchment. In both cases, DL models exhibit deficiencies in the correct representation of the response of the catchment to the perturbation of the factors.

3.2.2 CAMELS testing period

440 The available Hydrological Years (HYs) in the testing period were classified into wet, normal and dry years based on the annual precipitation, and into warm, normal and cold years based on the annual mean temperature. The classification was done by computing the median of each category relative to the mean of the normal category, both in absolute and relative terms. The wet

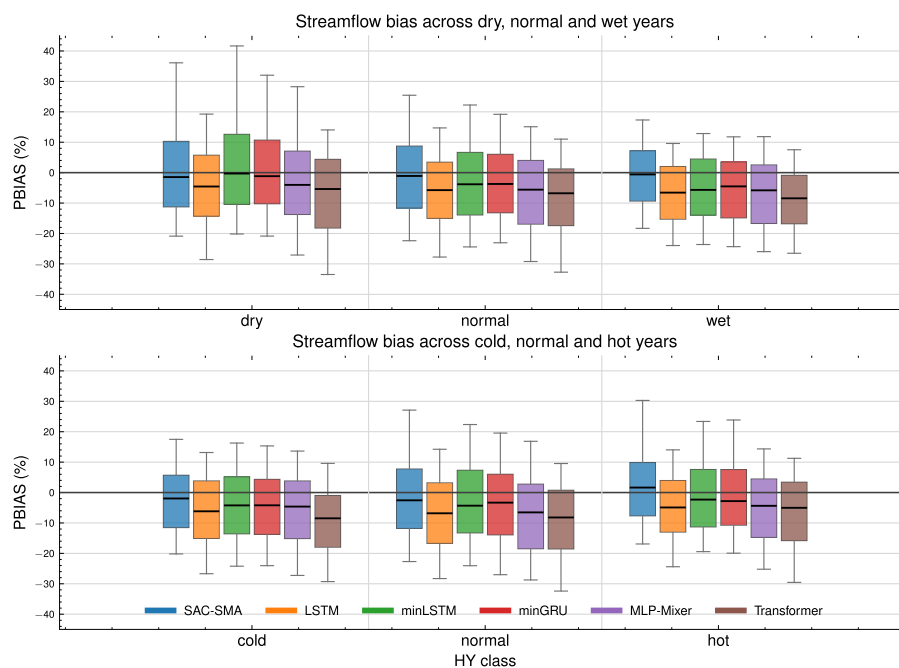


Figure 8. PBIAS of the evaluated models on the CAMELS dataset, by hydrological year classification

years had a median precipitation that was $\sim 15\%$ higher than the normal years, while the dry years had a median precipitation that was $\sim 15\%$ lower than the normal years. The warm years had a median temperature that was $\sim 0.7^\circ\text{C}$ higher than the normal years, while the cold years had a median temperature that was $\sim 0.7^\circ\text{C}$ lower than the normal years.

Figure 8 shows the percent bias (PBIAS) of each model computed against the observed streamflow, across the wet, normal and dry (and warm, normal and cold) hydrological-year classes. Because the models are here evaluated on year classes that differ from the average conditions of the training period, a model that keeps its PBIAS close to zero across all classes can be considered robust, since its bias does not depend on the climatic condition of the year. Overall, SAC-SMA stayed closest to zero, while all the data-driven models shared a moderate negative bias of comparable magnitude (median PBIAS between -3.39% and -7.13% , being Transformer the model with the largest bias). Most of the data-driven models were more sensitive to the precipitation classes than to the temperature classes: their negative bias get larger from dry to wet years and was smallest in warm years, whereas SAC-SMA showed the opposite temperature response, turning slightly positive in hot years (median $+1.6\%$). The largest PBIAS spread was achieved in dry and hot years, which is consistent with the fact that in these years the available streamflow is lower, and every change in the streamflow has a larger impact on the PBIAS. SAC-SMA and the MLP-Mixer were the most stable across precipitation classes (median-PBIAS range of 0.9 and 1.8% , respectively).



3.2.3 Runoff coefficient on Historical period

The models were evaluated on the historical period of the NEX-GDDP-CMIP6 dataset, after the bias-correction. The historical period, common to all GCMs, is used as a baseline to compare the future projections against, and it is also used to evaluate if the models exhibit any systematic bias using a different dataset.

Therefore, as an independent water-balance plausibility check, the mean runoff coefficient (Eq. (8)) was computed using the modeled results and the CMIP6 GCM precipitation forcing bias-corrected with Daymet, and compared with the reported static RC from CAMELS dataset (Newman et al., 2015; Addor et al., 2017). Fig. 9 shows the mean runoff coefficient separated by model for every catchment, colored by the aridity index. The runoff coefficient followed the expected gradient with aridity (Sankarasubramanian et al., 2001), with wetter catchments converting a larger fraction of precipitation into streamflow. SAC-SMA exhibits a good correlation between the Runoff Coefficient informed by the CAMELS dataset and the Runoff Coefficient computed from the historical period of the NEX-GDDP-CMIP6 dataset, with a Spearman correlation of 0.964. This model obtained an under-estimation of the RC coefficient for a set of catchments with low aridity index.

In the case of DL models, the Spearman correlation ranges between 0.875 to 0.949, for MLP-Mixer and LSTM, respectively. For energy-limited catchments, with an aridity index < 1 , the RC computed matches close to the reference RC. In contrast, water-limited catchments over-estimate the RC, exhibiting a larger fraction of precipitation converted to streamflow. It is unclear if this over-estimation is driven by precipitation or temperature, although the perturbation experiments exposed higher sensitivity in temperature for the rain-dominated catchment as exhibited for catchment id 8025500. More examples can be found in the Supplementary Material.

3.3 Fractional change in mean streamflow

Each model was evaluated on the historical period and on three future periods (Period 1: 2015–2045, Period 2: 2045–2070, and Period 3: 2071–2100) for each of the six GCMs and four SSP scenarios. The fractional change in mean streamflow (Λ , Eq. (9)) was computed for each of these factors, and the results are shown in Fig. 10. The median change was small across all models and scenarios, with a median Λ of 0.974, indicating a small decrease in mean streamflow. The response varied by scenario, deepening as the emission level increased: 0.992 (SSP1-2.6), 0.979 (SSP2-4.5), 0.965 (SSP3-7.0), and 0.956 (SSP5-8.5). Across periods the differences were minor ($\bar{\Lambda}$ of 0.977, 0.972 and 0.972 for Periods 1–3), with the largest departure from the historical baseline occurring in Period 2.

The hydrological models generally agreed on the sign of the change within each scenario, but differed in its magnitude, with the Transformer projecting the largest change (Fig. 10). This spread widened under the higher-emission scenarios, indicating that model choice becomes more influential as the climate signal grows.

3.4 Robustness evaluation under non-stationary conditions

After computing the fractional change in mean streamflow for each model, catchment, GCM, scenario, and period, the divergence of the deep learning models from the SAC-SMA benchmark was evaluated using PBIAS (Eq. (10)). The results are

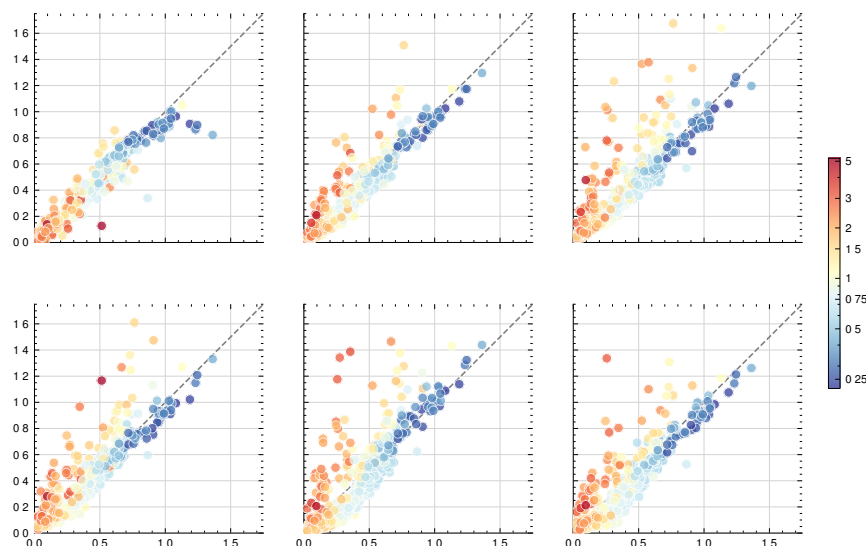


Figure 9. Mean runoff coefficient for the historical period across the CMIP6 GCM dataset, by model and catchment. The color represents the aridity index extracted from the CAMELS dataset. The observed runoff coefficient was computed using Daymet data and USGS discharge observations during the hydrological years 1990 to 2009 (Addor et al., 2017), the runoff coefficients computed here use the same period

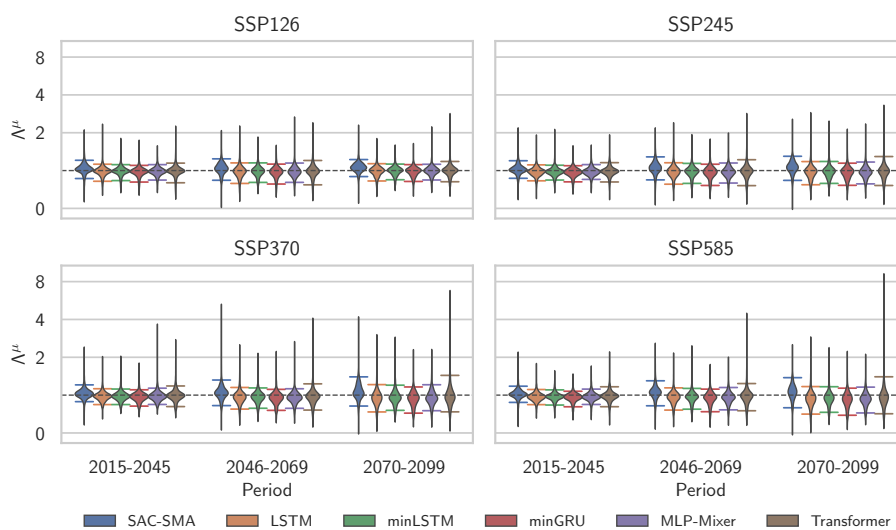


Figure 10. Fractional change in mean streamflow for the future period relative to the historical period, by model, catchment, and scenario. Horizontal colored lines denote the 5 %-95 % percentile.

summarized in Fig. 11, showing the robustness rating of the deep learning models, aggregated by period and scenario. The

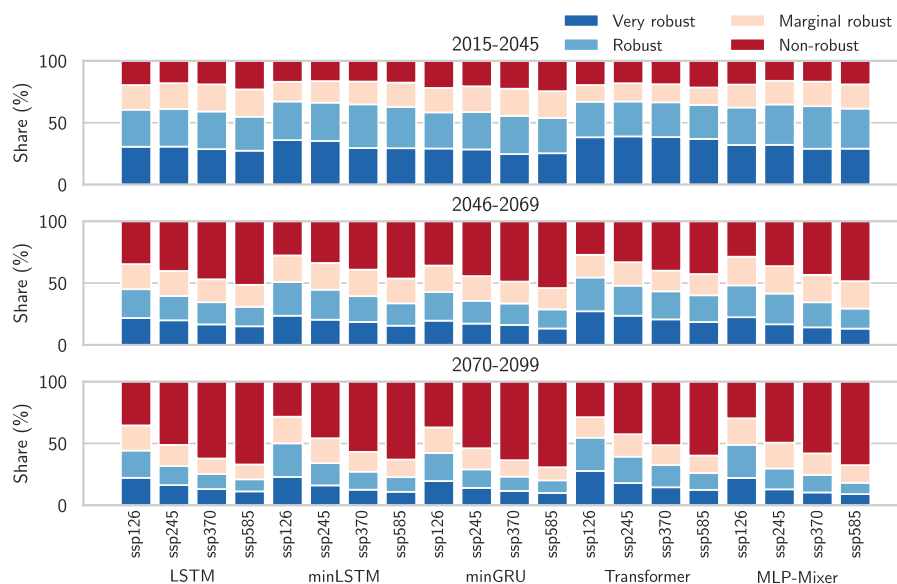


Figure 11. Deep learning model robustness evaluation, by scenario and period

490 rating is derived from the divergence of each model’s fractional change from the SAC-SMA benchmark, classified into the four robustness bands defined in Sect. 2.5.5, thus a higher rating indicates that a model departs less from the physics-based benchmark as the climate signal moves away from the historical conditions.

Interestingly, the ranking is broadly the reverse of the predictive-skill ranking of Sect. 3.1: The most robust model using the mean divergence analysis is the Transformer, with 50.3 % of the catchment–scenario–period cases rated robust or very robust. It is followed by minLSTM with 47.0 % of catchments under the same criteria. Transformer also showed the minimum divergence from the weak benchmark, with a $\overline{\text{PBIAS}}$ of -8.40 %. The LSTM and the minGRU—the best performers on the CAMELS test period—were the least robust (42.3 % and 40.1 % of cases robust or very robust, and the largest mean divergence, -11.3 % and -13.1 %, respectively). All models became markedly less robust as the projection period advanced from Period 1 to Period 3 and as the scenario intensified; under SSP5-8.5, for example, the fraction of robust or very robust cases fell from 62.7 % to 22.7 % for the minLSTM and from 54.9 % to 21.0 % for the LSTM. This degradation is consistent with the reported difficulty of data-driven models in extrapolating beyond the conditions seen during training (O et al., 2020; Wi and Steinschneider, 2022; Baste et al., 2025).

3.5 Robustness in the day-of-year climatology

The results above collapse each future window—the combination of scenario and period—to a single scalar per catchment, hence they capture shifts in the total volume of water but are blind to *when* that water arrives. This is a good way to summarize the robustness of the models in terms of the annual water balance, but it does not capture the seasonal redistribution of stream-

flow that is projected under climate change, which is a key feature of the hydrological response to warming (Barnett et al., 2005; Christensen and Lettenmaier, 2007). The day-of-year (DOY) climatology signature resolves this intra-annual timing and seasonality, consequently we also evaluate robustness on this more demanding signature.

510 To resolve the above, the same fractional-change framework (Eq. (9)) is applied to the day-of-year (DOY) climatology signature: for every catchment the mean streamflow on each calendar day (1–366) is computed for the historical and each future window, obtaining a fractional-change as Λ^{doy} over the year instead of a single value. The per-day divergence of each deep learning model from SAC-SMA (Eq. (10)) is then a function of the day of year. Fig. 12 illustrates the DOY climatology for the most extreme case (SSP5-8.5, Period 3), averaged across catchments and GCMs. The physics-based benchmark projects
 515 a pronounced increase in late-winter and early-spring streamflow ($\Lambda^{\text{doy}} \approx 2.3$ around day 22), the signature of an earlier snowmelt under warming, followed by a decrease through the summer. For the wrong reasons, Transformer is also the model with the smallest departure from SAC-SMA. The $\overline{\text{PBIAS}}^{\text{doy}}$ is -6.5 %, with a $|\overline{\text{PBIAS}}|^{\text{doy}}$ of 40.9 %, not reproducing the seasonal timing and showing a larger streamflow during the fall season. This increase compensates with the decrease during the winter, resulting in the one with the smallest mean divergence. All the DL models show lower seasonal change during the
 520 winter, with relatively a similar behavior. The differences among the DL models are exhibited during summer and fall seasons. For this extreme case, the model with the smallest $|\overline{\text{PBIAS}}|^{\text{doy}}$ is LSTM, with a magnitude of 34.2 %, followed by minLSTM with 35.9 %. Examples for individual catchments (the 18 selected) are included in the Supplementary Material, for the first period and lowest emissions and the third period and highest emissions scenarios. In most of the cases, the response of the DL models are similar to the averaged behavior presented in Fig. 12.

525 In terms of robustness, the DOY signature is more demanding than the mean signature because it resolves the intra-annual timing of streamflow. On this more demanding signature the robust-or-better share collapses for every architecture: fewer than 4 % of the catchment–scenario–period cases are rated robust or very robust (MLP-Mixer 2.9 %, LSTM 3.4 %, minLSTM 3.2 %, minGRU 3.9 % and Transformer 3.9 %), with median $|\overline{\text{PBIAS}}^{\text{doy}}|$ between 23 % and 27 %. An agreement on annual volumes does not imply agreement on the seasonal timing of streamflow.

530 3.6 Statistical assessment

The assessment used the 199 440 PBIAS values obtained from the combination of 554 catchments, 5 architectures, 6 climate models, 4 scenarios and 3 periods. A per-catchment Kruskal–Wallis test was applied to determine which factor—architecture, climate model, scenario or period—had the largest rank-based effect size η_H^2 on the PBIAS, in the spirit of uncertainty partitioning for climate-impact ensembles (Chegwidden et al., 2019).

535 Taking, for each catchment, the effect sizes of the four factors from the Kruskal–Wallis test and the rank-based effect size η_H^2 , the dominant driver of the divergence from SAC-SMA is shown in Fig. 13, which maps the dominant factor across CONUS. The deep learning architecture was the dominant source of variability in most catchments 45 % (247 of the 554 catchments), followed by period with 33 % (182 catchments), and the GCM 23 % (125 catchments); the SSP scenario was never the dominant factor. This indicates that, at catchment level, the deep learning architecture is the main driver of divergence from
 540 the physics-based benchmark, rather than the climate forcing in close to half of the cases. The effect of the architecture is

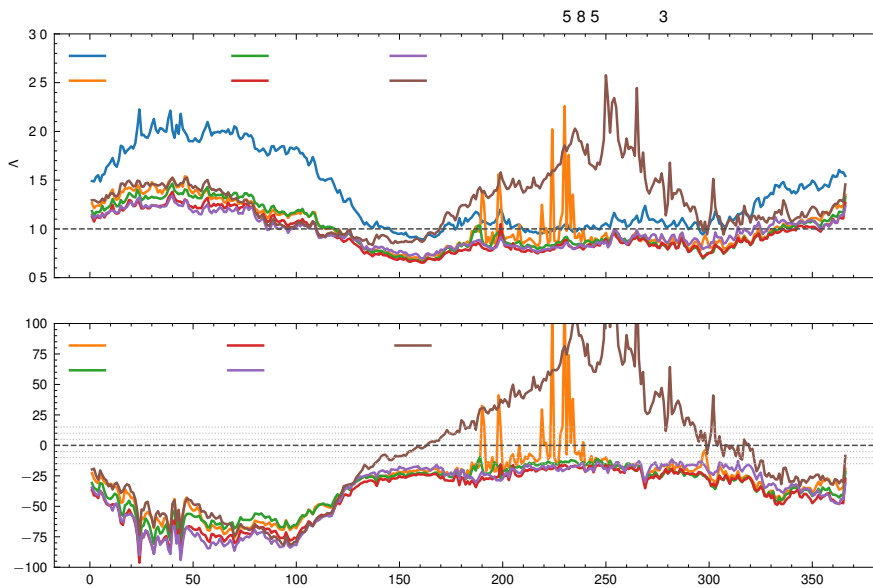


Figure 12. Day-of-year climatology fractional change Δ^{day} (top) and per-day divergence from SAC-SMA ($\text{PBIAS}^{\text{day}}$, bottom) under SSP5-8.5, Period 3, averaged across catchments and GCMs.

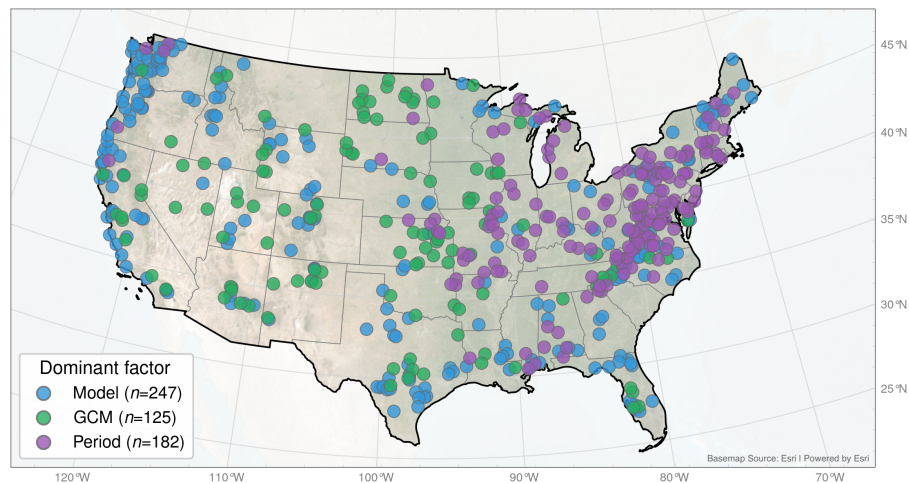


Figure 13. Per-catchment Kruskal–Wallis factor with maximum η_H^2 . Sources: US Census Bureau, USGS, ESRI World Imagery.

particularly pronounced in the west coast catchments. In the arid catchments of the southwest, the GCM is the dominant factor, which is consistent with the fact that these catchments are more sensitive to precipitation than to temperature. Period was a common dominant factor in the Appalachian region and areas close to it.

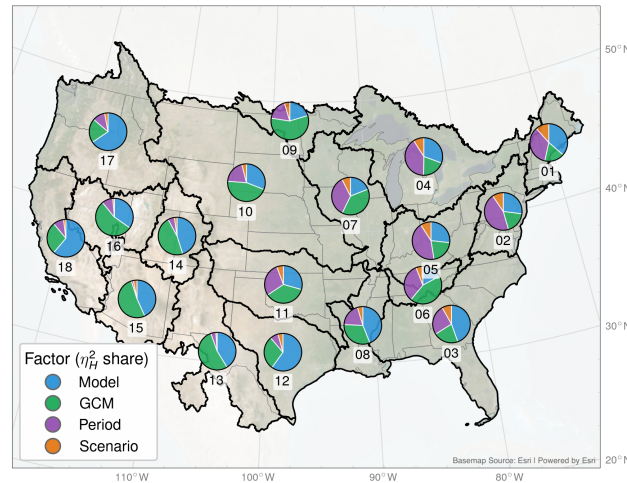


Figure 14. Kruskal–Wallis shared η_H^2 aggregated at HUC level. Sources: US Census Bureau, USGS, ESRI World Imagery.

To summarize the effect in a more interpretable way, the shared η_H^2 was aggregated at Hydrologic Unit Code to show the individual effect of each factor. The results in Fig. 14 shows the mean shared η_H^2 by HUC, to highlight the spatial distribution of the effect of each factor.

The HUC with the largest mean shared η_H^2 for the architecture is HUC 17 (Pacific Northwest Region), with a share of 65.6 % of the total effect, followed by HUC 18 (California Region) with 61.1 %. The GCM has the largest mean shared η_H^2 in HUC 9 (Souris-Red-Rainy Region), with a share of 57.0 %, followed by HUC 16 (Great Basin Region) with 54.4 %. The period has the largest mean shared η_H^2 in HUC 2 (Mid Atlantic Region), with a share of 45.0 %, followed by HUC 5 (Ohio Region) with 43.3 %, both are located in Appalachian region. The scenario has the smallest mean shared η_H^2 across all HUCs, with a maximum of only 11.5 % in HUC 1 (New England Region).

Each HUC has a different dominant factor, but also a different forcing signal. Since the forcing signal is not uniform across the country, it is expected that the effect of each factor will also vary spatially. Since the scope of the current work is to evaluate the robustness of the deep learning models, we focus on the effect of the architecture, but to give a context of the forcing signal, here a summary of the two main forcing factors is presented (summarized by HUC code):

- Precipitation: all HUCs presented a higher precipitation in the scenario SSP1-2.6, with a mean increase of 6.6 % across the country, ranging from +1.6 % (HUC 18, California Region) to +11.5 % (HUC 5, Ohio Region). There are divergences from the scenarios SSP2-4.5 to SSP5-8.5. In the scenario with highest emissions, the HUCs 12, 13 and 18 exhibited -2.5 %, -1.2 % and -0.7 %, respectively. On the other hand, the HUCs 1, 6 and 16 presented a precipitation of $\sim +14$ % (+14.3 %, +13.6 % and +13.5 %), exposing the differences across scenarios and HUCs, although the scenario factor was the one with the least effect. In the same scenario, the change between periods is increasing in most of the cases—with the exception of the HUCs mentioned before—going from a country-wide mean of +3.4 % in Period 1 to +7.7 % in Period 3; by the last period the mean increase ranges from +6.6 % in the scenario SSP1-2.6 up to +7.7 % in the scenario SSP5-8.5.



565 – Temperature: The temperature dynamic is more straightforward to track, with a clear increase in temperature across all HUCs and scenarios. The mean increase is +2.1°C in the scenario SSP1-2.6, +2.9°C in the scenario SSP2-4.5, +3.8°C in the scenario SSP3-7.0, and +4.7°C in the scenario SSP5-8.5, considering the increase from the historical period to the last period. The HUCs 9, 4 and 1 are the ones with the largest increase in temperature, with +5.7°C, +5.5°C and +5.3°C, respectively, in the scenario SSP5-8.5.

570 4 Discussion

4.1 Deep learning against process-based benchmarks in large-sample hydrology

A standard protocol in rainfall-runoff modeling is to benchmark the data-driven models against a calibrated conceptual model over a large sample of catchments. The physically based reference for the CAMELS sample was established by Newman et al. (2017), who documented the skill of SAC-SMA and other conceptual models across the contiguous United States, which later
575 on was used by Kratzert et al. (2018) to show that a LSTM can match or exceed a per-basin calibrated SAC-SMA+Snow-17. Multiple benchmarks have been developed with this approach (Klotz et al., 2022; Liu et al., 2024, 2025), including metrics like NSE, KGE and upper/lower performance benchmarks (Gupta et al., 2009; Seibert et al., 2018), and datasets (Addor et al., 2017; Newman et al., 2015; Clerc-Schwarzenbach et al., 2024). All these community efforts are based on the assumption that the data-driven models are able to capture the physical processes that are not explicitly represented in the conceptual models,
580 and that the data-driven models are able to learn the relationships between the input and output variables.

The approach of training Deep Learning models for a single catchment, similar to the approach followed by physics-based models, was adopted initially by Kratzert et al. (2018), which later transitioned to regional training, extending data-driven models to ungauged basins (Kratzert et al., 2024). LSTMs trained jointly on hundreds of basins learn transferable, regionally coherent behavior (Kratzert et al., 2019c) and outperform locally calibrated conceptual and process-based models even in
585 ungauged conditions (Kratzert et al., 2019b; Arsenault et al., 2023), requiring large and diverse training samples (Gauch et al., 2021; Kratzert et al., 2024; Clerc-Schwarzenbach et al., 2024).

4.2 Robustness evaluation

For testing the reliability of data-driven models, three different approaches have been taken in this work. The first applies controlled forcing perturbations (uniform temperature offset and multiplicative precipitation scaling) inspecting the resulting
590 change in streamflow (Wi and Steinschneider, 2022, 2024; Martel et al., 2025). The second approach follows the differential split-sample test (DSST) of conceptual hydrology, training on one hydroclimatic period and validating on a contrasting one (Natel de Moura et al., 2022; Behfar et al., 2024; Bohl et al., 2026), but with a limited set of hydrologic years in this case, as an initial exploration of model robustness. The third one utilizes the trained models with downscaled Global Climate Model (GCM) projections under emission scenarios and evaluates the divergence of the projected streamflow (Singh et al., 2023; Ullah et al., 2023; Kartal, 2024; Wei et al., 2025). Each approach provides a different insight: the dependence of the changes



in temperature and precipitation of the perturbation experiments, the contrasted-climate behavior of the second one, and the application of a long and projected non-stationary condition of the third one.

As a complementary plausibility check, the runoff coefficient was computed using the bias-corrected GCM datasets to identify any systematic bias using the models trained with CAMELS dataset. In general, the agreement between the runoff
600 coefficient observed and computed from this study is good. Only catchments with high aridity index exhibited a positive systematic bias, mostly related with the high sensitivity to precipitation and the possible over-fit identified from the perturbation experiments.

4.2.1 Robustness under controlled perturbations

The perturbation experiments, like the one presented here, are the most direct probe of physical realism, and they are able to
605 expose the physical limitations of the data-driven models. The results show that the data-driven models reproduce the main hydrological responses reasonably well, even those not explicitly represented in the conceptual model. However, their lack of an explicit process characterization remains a fundamental limitation. This is particularly evident in the snow-dominated catchments, where the data-driven models are not able to reproduce the first-order response of SAC-SMA + Snow-17, creating a ghost snowpack and a wrong precipitation response. In the case of the rain-dominated catchments, the data-driven models
610 exhibited a more reasonable response for the precipitation, but the temperature perturbation exhibited a larger streamflow reduction than the Priestley-Taylor evapotranspiration pathway inflicts on the SAC-SMA + Snow-17 response. Although the two most important factors that drive streamflow were tested, and after finding the over-sensitivity of the data-driven models to temperature, is still unclear if the other forcings are generating over-fitting behaviors that are outside the expected behavior of the data-driven models, but that is outside the scope of the current work.

615 The results obtained from the perturbation experiments are consistent with other studies that have explored the physical realism of the data-driven models. Wi and Steinschneider (2022) showed that the LSTM trained on 15 Californian watersheds was able to reproduce the expected shift toward more winter runoff, but in several basins it projected an unrealistic increase in the runoff coefficient that HYMOD, SAC-SMA, and VIC did not. Wi and Steinschneider (2024) found that the LSTM constrained to respect the PET-evaporation relationship (MC-LSTM-PET) corrected the excessive water loss under warming
620 with an energy-budget PET. Cen et al. (2026) reported that purely data-driven LSTMs give physically implausible projections in warming cases in the Pearl River Basin, wrongly interpreting a temperature rise as a runoff-generating signal. These studies show that in-training accuracy does not guarantee a reasonable extrapolation.

4.2.2 Robustness under CAMELS testing period

The split-sample test results are consistent with other studies that have explored the robustness of the data-driven models under
625 contrasted climate conditions. Bohl et al. (2026) trained LSTM, HBV, and differentiable hybrid models on a cold 1960-1990 period and tested them on a warmer 2000-2023 period; the LSTM was the most accurate but had the largest performance drop and an over-sensitivity of mean and high flows to temperature that the other models did not presented. Natel de Moura et al. (2022) found the same in six snow-influenced Swiss catchments where the performance of the LSTM was similar or worse



than the conceptual model benchmark, while the later one requires less data to achieve that skill. In the same line, Behfar et al. (2024) found conceptual models to outperform a data-driven model under the DSST scheme in the Meuse basin. Our contrasted-climate perturbation over the CAMELS testing period reproduces this ordering: SAC-SMA stayed closest to zero bias across the wet, dry, warm, and cold hydrological-year classes, while all five data-driven architectures shared a moderate negative bias that grew from dry to wet years and was largest in dry and warm years. The largest penalty falls in the snow-dominated regime regions, which our statistical assessment identifies the architecture, not the climate forcing, drives the divergence from the physics-based benchmark.

4.2.3 Robustness under climate-change projections (NEX-GDDP-CMIP6)

This study is one of the few that evaluates the robustness of the data-driven models under climate-change projections. The results indicated that a good skill in the calibration period does not guarantee a good skill in the projection period, which has been reported in other studies (Martel et al., 2025; Natel de Moura et al., 2022). The results also showed that the data-driven models are able to reproduce the expected shift toward more winter runoff, but the magnitude differs from the physics-based benchmark.

On the mean-flow signature, the data-driven models diverged systematically toward lower streamflow than SAC-SMA as the climate signal intensified, showing a similar result of the excessive water loss of Wi and Steinschneider (2024). Because our metric is a within-model fractional change, this reflects each architecture's response to the forcing, not its static historical bias. Overall, the divergence widened with the strength of the climate signal, in line with the changes projected for these regions, reducing the robustness of the data-driven models through the time (Vano et al., 2012; Martel et al., 2025).

Almost all of the robustness studies above evaluate a single architecture (LSTM), therefore they cannot separate robustness that is intrinsic to the data from robustness that is a property of the architecture. Interestingly, the results show that the predictive-skill order on the CAMELS test period is broadly reversed in the mean-signature robustness order, with the LSTM and minGRU the most accurate yet least robust, while Transformer was the model with the highest overall robustness. This is a direct argument for evaluating multiple architectures whenever robustness, not just accuracy, is the goal. The multi-architecture design also introduces the newer models in context: the Transformer shows the smallest mean-signature divergence from SAC-SMA (-8.40 %), had until now been assessed in hydrology mainly for accuracy (Liu et al., 2024, 2025) rather than for climate robustness. Hence the architecture, not the GCM or the scenario, is the dominant driver of divergence in 45 % of catchments, which is a direct argument for evaluating multiple architectures whenever robustness is the goal.

The use of data-driven models for climate-change projections is a growing trend, but the lack of a physics-based benchmark is a concern. Some studies have reported future streamflow for specific basins using data-driven models without a strong intercomparison to a physics-based benchmark (Singh et al., 2023; Chanda et al., 2024; Ullah et al., 2023; Kartal, 2024; Nguyen et al., 2023; Sha et al., 2024; Wang et al., 2025). Our results suggest that a calibration-period skill score is not enough to trust such projections, and that a within-model comparison against a physics-based benchmark, as used here, is a low-cost safeguard these workflows could adopt.



Even when some architectures exhibited robustness, the results showed that the seasonal timing of streamflow is not well reproduced by the data-driven models. The physical benchmark shows a clear shift toward earlier snowmelt and a compensating summer decline of a warming climate, which is not well reproduced by the data-driven models. The results are even more surprising considering that the Transformer, the strongest model on skill and mean-flow robustness, is the weakest on seasonal timing, departing from the physical benchmark during different seasons in both directions. These divergence of predictions is summarized by a absolute bias of 40.9 %, while the mean bias is -6.5 %. Under the robustness metric proposed in this study, this model is still the most robust, by the wrong reasons. Recurrent architectures exhibited a lower absolute bias (LSTM with 34.2 %), exhibiting a permanent negative bias, which gets more negative during the winter. This split between volume robustness and timing robustness is invisible to annual signatures, and it directly affects the flood-forecasting, reservoir-operation, and irrigation applications for which the timing of streamflow, not just its annual total, is the decision-relevant quantity.

The statistical assessment results are consistent with other studies that have explored the robustness of the data-driven models under contrasted climate conditions. Martel et al. (2025) showed that an LSTM trained on 148 catchments has a less credible climate sensitivity than a continental LSTM trained on 1000 catchments spanning the study domain's future climate zones, and concluded that the continental model is better suited to climate-impact studies. Natel de Moura et al. (2022) found that a longer calibration series improves LSTM transferability. Increasing the sample size and the hydroclimatic diversity of the training sample could help to improve the robustness of the data-driven models, and that would require either extending the training period, or increasing the number of catchments in the training sample. The first option, extending the training period, is the most feasible option since the catchments present in the CAMELS database represents areas with minimal human impact and low complexity, therefore the addition of more catchments with more complexity and human impact would either require a larger training sample.

4.2.4 Strengths and limitations

Where the training sample cannot be extended, the use of physical constraints offers a way to improve the robustness of the data-driven models. Differentiable process-based models keep the accuracy of deep networks while guaranteeing mass conservation and a full set of diagnostic variables. Feng et al. (2023) showed this approach surpassing LSTM in ungauged basins and giving more credible decadal trends. In the same way, the MC-LSTM-PET of Wi and Steinschneider (2024), the physics-encoded LSTM of Cen et al. (2026), and the differentiable hybrid of Bohl et al. (2026) all recover physically admissible warming responses that their purely data-driven versions fail to produce. Our study uses only purely data-driven architectures, thus it cannot test these approaches directly.

Against this body of work, the present study offers four methodological strengths. First, it evaluates five architectures from three model families under one training, calibration, and evaluation protocol, hence that differences in robustness can be attributed to the architecture rather than to the setup. Second, it drives the models with a spatially coherent, bias-corrected six-GCM, four-SSP ensemble rather than idealized offsets, consequently the forcing is internally consistent while spanning a wide range of climate signals. Third, the within-model fractional-change design removes each model's static historical bias, thus the metric isolates the response to the climate signal and is not confounded by calibration skill (unlike direct-projection



studies that compare absolute flows). Fourth, the evaluation is resolved both spatially, over 554 CONUS catchments with an explicit partitioning of the sources of divergence, and temporally, on both the annual-mean and day-of-year signatures, which together expose the volume-timing split that single-signature assessments miss.

These strengths come with limitations. The central one is shared with the whole field: future streamflow cannot be observed, hence robustness can only be judged against a benchmark that is itself imperfect under non-stationary forcing, since a conceptual model calibrated on the historical period carries its own structural and parameter drift into the projection (Coron et al., 2012; Thirel et al., 2015). SAC-SMA + Snow-17 is a defensible but not infallible reference, and it receives temperature only through PET and snow partitioning while the data-driven models get raw temperature directly, an asymmetry that may inflate the apparent divergence in snow-dominated catchments. The robustness bands are an interpretable but arbitrary convention, the CAMELS testing window used for the contrasted-climate perturbation is short, and the analysis is limited to purely data-driven, lumped models, therefore it neither tests the hybrid and physically constrained architectures proposed as remedies nor addresses distributed or ungauged settings. Finally, the ensemble samples six GCMs rather than the full CMIP6 archive, therefore the GCM contribution to the divergence, though never dominant here, is likely a lower bound, and the statistical bias correction does not fully remove systematic error: a residual bias can persist (Cannon et al., 2015; Sofokleous et al., 2025) and propagate into the projected signatures.

5 Conclusions

In this study, we presented a robustness assessment following multiple approaches and deep learning architectures for streamflow simulation under climate-change projections for 554 catchments across CONUS. Using a within-model fractional change metric, we isolated each architecture's response to the forcing with its systematic bias across six GCMs, four SSP scenarios, and three future periods. SAC-SMA coupled with Snow-17 was used as physical benchmark, and the within-model metric was compared with the process-based model to establish the robustness property of each deep learning architecture.

We observed that a high skill score over the calibration period does not guarantee a high robustness in non-stationary conditions. For example, LSTM, the model with the highest KGE (median of 0.785) is the least robust for the period-mean streamflow, while Transformer showed the highest robustness for this signature. The data-driven models are able to reproduce the shift in seasonality for the day-of-year climatology, although the timing and magnitude depart drastically from the process-based benchmark. This departure is amplified for the most extreme case (scenario SSP5-8.5 and period 2071–2100), showing that the average robustness of all data-driven models is marginal under this signature evaluation. This signature also exposed the reason of the Transformer's robustness, showing that a compensating of the winter's lower predictions is compensated by a higher predictions of the summer streamflow, indicating that a multi-approach evaluation is recommended to fully understand the the divergence between models. To complement the previous statement, the perturbation analysis helps us to understand the DL models' behavior if the main forcing that explain streamflow process generation are altered (precipitation and temperature). The observed response in the selected catchments exhibited unrealistic responses for the DL models—such as ghost snowpack



when the precipitation is forced to 0 for a snow-dominated catchment or and over-fit to temperature for the rain-dominated catchment—and when the response was similar then the process-based model, the elasticity of those models differed.

730 The results indicates that the selection of a data-driven rainfall-runoff hydrological model based on accuracy alone, to be used to evaluate climate projections for a single architecture without a physical reference, is not necessarily reliable. We recommend that: (i) evaluate multiple architecture rather than using the default LSTM, (ii) use a within-model comparison for comparing the projected signals with a process-based benchmark, and (iii) report timing-based signature alongside with volume-based ones, as the period-mean complemented with day-of-year presented here. Finally, robustness could be improved extending the training period and including more catchments to improve the process representation of the data-driven models, as has been demonstrated with previous studies.

735

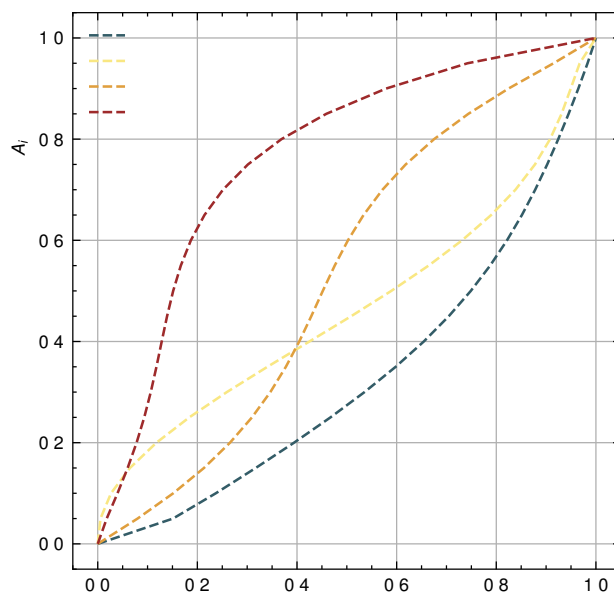


Figure A1. Areal Depletion Curve (ADC) of 4 different shapes described in SNOW-17 documentation, recreated with the ADC equation (Eq. (A1))

Appendix A: Appendix A

To calibrate the Snow Areal Depletion Curve (ADC), we used a two-Beta CDF ensemble with a weighting parameter as follows:

$$740 \quad \text{ADC} = \omega \cdot B_1(\alpha_1, \beta_1) + (1 - \omega) \cdot B_2(\alpha_2, \beta_2) \quad (\text{A1})$$

where $B_1(\alpha_1, \beta_1)$ and $B_2(\alpha_2, \beta_2)$ are two Beta CDF distributions with shape parameters α_1, β_1 and α_2, β_2 , respectively, and ω is a weight parameter that controls the contribution of each distribution to the overall ADC. This equation is able to represent 4 different shapes of the ADC, as described in the documentation of SNOW-17 (Anderson, 1973), and is illustrated in Fig. A1.

The calibration of these super-parameters ($\alpha_1, \beta_1, \alpha_2, \beta_2$, and ω) allows a more flexible representation of the snow areal depletion curve, which can improve the model's ability to capture the snowmelt dynamics in different catchments, and also, reduces the number of parameters to fit (Snow-17 uses 11 parameters to define the shape of the ADC curve).

The search of optimal fit for each super-parameter was conducted by the ranges presented in Table A1.



Table A1. Lower and upper limits for the ADC optimal parameter search.

Parameter	Lower limit	Upper limit
α_1	0.3	5.0
β_1	0.3	5.0
α_2	0.3	5.0
β_3	0.3	5.0
w	0.0	1.0



Code and data availability. The data to reproduce the results of this study are provided in <https://doi.org/10.4211/hs.9ff1c91ad7234bb8a8edbdde4e30bff3> (last access: 25 September 2026), including: (i) The model performance skill for individual models; (ii) the SAC-SMA +
750 Snow-17 re-calibrated parameters; (iii) the DL model files, (iv) Λ and PBIAS for the signatures evaluated, (v) mean forcing and streamflow for each model by hydrological year. The code to train and execute the DL models is available in https://github.com/aldotapia/Rainfall_Runnoff_DL_Robustness (last access: 25 September 2026)

Author contributions. All authors contributed to the development of the study ideas. AB initialized the DL models development, edited and reviewed the manuscript. AT re-calibrated the physically-based benchmark, carried out the experiments, and wrote the manuscript.

755 *Competing interests.* The contact author has declared that none of the authors has any competing interests.

Acknowledgements. The authors thank the University of Arizona College of Science for providing startup funds to support this work



References

- Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS Data Set: Catchment Attributes and Meteorology for Large-Sample Studies, *Hydrology and Earth System Sciences*, 21, 5293–5313, <https://doi.org/10.5194/hess-21-5293-2017>, 2017.
- 760 Allen, R., Pereira, L., Raes, D., and Smith, M.: Crop Evapotranspiration Guidelines for Computing Crop Requirements. FAO Irrig. Drain. Report Modeling and Application, *J. Hydrol.*, 285, 19–40, 1998.
- Anderson, B. J., Brunner, M. I., Slater, L. J., and Dadson, S. J.: Elasticity Curves Describe Streamflow Sensitivity to Precipitation across the Entire Flow Distribution, *Hydrology and Earth System Sciences*, 28, 1567–1583, <https://doi.org/10.5194/hess-28-1567-2024>, 2024.
- Anderson, E. A.: National Weather Service River Forecast System: Snow Accumulation and Ablation Model, Tech. Rep. 17, National
 765 Weather Service, Washington, D.C., 1973.
- Arsenault, R., Martel, J.-L., Brunet, F., Brissette, F., and Mai, J.: Continuous Streamflow Prediction in Ungauged Basins: Long Short-Term Memory Neural Networks Clearly Outperform Traditional Hydrological Models, *Hydrology and Earth System Sciences*, 27, 139–157, <https://doi.org/10.5194/hess-27-139-2023>, 2023.
- Ballarin, A. S., Sone, J. S., Gesualdo, G. C., Schwaback, D., Reis, A., Almagro, A., and Wendland, E. C.: CLIMBra - Climate Change
 770 Dataset for Brazil, *Scientific Data*, 10, 47, <https://doi.org/10.1038/s41597-023-01956-z>, 2023.
- Barnett, T. P., Adam, J. C., and Lettenmaier, D. P.: Potential Impacts of a Warming Climate on Water Availability in Snow-Dominated Regions, *Nature*, 438, 303–309, <https://doi.org/10.1038/nature04141>, 2005.
- Baste, S., Klotz, D., Acuña Espinoza, E., Bardossy, A., and Loritz, R.: Unveiling the Limits of Deep Learning Models in Hydrological Extrapolation Tasks, *Hydrology and Earth System Sciences*, 29, 5871–5891, <https://doi.org/10.5194/hess-29-5871-2025>, 2025.
- 775 Behfar, N., Booij, M. J., and Nourani, V.: Assessing Rainfall-Runoff Models for Climate Change: Simple and Differential Split-Sample Tests for Conceptual and Artificial Intelligence Models, *Hydrological Sciences Journal*, 69, 861–877, <https://doi.org/10.1080/02626667.2024.2345151>, 2024.
- Benjamini, Y. and Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society: Series B (Methodological)*, 57, 289–300, <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>, 1995.
- 780 Beven, K.: Deep Learning, Hydrological Processes and the Uniqueness of Place, *Hydrological Processes*, 34, 3608–3613, <https://doi.org/10.1002/hyp.13805>, 2020.
- Bohl, J. P., Wood, R. R., Frank, C., Astagneau, P. C., Peters, J., and Brunner, M. I.: Hybrid Models Generalize Better to Warmer Climate Conditions than Process-Based and Purely Data-Driven Models, *Hydrology and Earth System Sciences*, 30, 4667–4698, <https://doi.org/10.5194/hess-30-4667-2026>, 2026.
- 785 Brown, M. B. and Forsythe, A. B.: Robust Tests for the Equality of Variances, *Journal of the American Statistical Association*, 69, 364–367, <https://doi.org/10.1080/01621459.1974.10482955>, 1974.
- Burnash, R. J. C., Ferral, R. L., and McGuire, R. A.: A Generalized Streamflow Simulation System: Conceptual Modeling for Digital Computers, U. S. Department of Commerce, National Weather Service, and State of California, Department of Water Resources, 1973.
- Cannon, A. J., Sobie, S. R., and Murdock, T. Q.: Bias Correction of GCM Precipitation by Quantile Mapping: How Well Do Methods
 790 Preserve Changes in Quantiles and Extremes?, *Journal of Climate*, 28, 6938–6959, <https://doi.org/10.1175/JCLI-D-14-00754.1>, 2015.
- Cen, Y., Yu, B., Xiong, R., Lu, M., Tian, Y., Han, F., Xu, P., and Zheng, Y.: Bridging Training–Projection Gaps in Purely Data-Driven Deep Learning for Runoff under Climate Change, *Journal of Hydrology*, 674, 135 508, <https://doi.org/10.1016/j.jhydrol.2026.135508>, 2026.



- Chanda, N., Chintalacheruvu, M. R., and Choudhary, A. K.: Exploring Climate-Change Impacts on Streamflow and Hydropower Potential: Insights from CMIP6 Multi-GCM Analysis, *Journal of Water and Climate Change*, 15, 4476–4498, <https://doi.org/10.2166/wcc.2024.150>, 2024.
- 795
- Chegwidden, O. S., Nijssen, B., Rupp, D. E., Arnold, J. R., Clark, M. P., Hamman, J. J., Kao, S.-C., Mao, Y., Mizukami, N., Mote, P. W., Pan, M., Pytlak, E., and Xiao, M.: How Do Modeling Decisions Affect the Spread Among Hydrologic Climate Change Projections? Exploring a Large Ensemble of Simulations Across a Diversity of Hydroclimates, *Earth's Future*, 7, 623–637, <https://doi.org/10.1029/2018EF001047>, 2019.
- 800
- Christensen, N. S. and Lettenmaier, D. P.: A Multimodel Ensemble Approach to Assessment of Climate Change Impacts on the Hydrology and Water Resources of the Colorado River Basin, *Hydrology and Earth System Sciences*, 11, 1417–1434, <https://doi.org/10.5194/hess-11-1417-2007>, 2007.
- Clerc-Schwarzenbach, F., Selleri, G., Neri, M., Toth, E., van Meerveld, I., and Seibert, J.: Large-Sample Hydrology – a Few Camels or a Whole Caravan?, *Hydrology and Earth System Sciences*, 28, 4219–4237, <https://doi.org/10.5194/hess-28-4219-2024>, 2024.
- 805
- Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., and Hendrickx, F.: Crash Testing Hydrological Models in Contrasted Climate Conditions: An Experiment on 216 Australian Catchments, *Water Resources Research*, 48, <https://doi.org/10.1029/2011WR011721>, 2012.
- De la Fuente, L. A., Ehsani, M. R., Gupta, H. V., and Condon, L. E.: Toward Interpretable LSTM-Based Modeling of Hydrological Systems, *Hydrology and Earth System Sciences*, 28, 945–971, <https://doi.org/10.5194/hess-28-945-2024>, 2024.
- 810
- Dinno, A.: Nonparametric Pairwise Multiple Comparisons in Independent Groups Using Dunn's Test, *The Stata Journal*, 15, 292–300, <https://doi.org/10.1177/1536867X1501500117>, 2015.
- Duan, Q., Sorooshian, S., and Gupta, V.: Effective and Efficient Global Optimization for Conceptual Rainfall-Runoff Models, *Water Resources Research*, 28, 1015–1031, <https://doi.org/10.1029/91WR02985>, 1992.
- Dunn, O. J.: Multiple Comparisons Using Rank Sums, *Technometrics*, 6, 241–252, <https://doi.org/10.1080/00401706.1964.10490181>, 1964.
- 815
- Ekambaram, V., Jati, A., Nguyen, N., Sinthong, P., and Kalagnanam, J.: TSMixer: Lightweight MLP-Mixer Model for Multivariate Time Series Forecasting, in: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '23*, pp. 459–469, Association for Computing Machinery, New York, NY, USA, ISBN 979-8-4007-0103-0, <https://doi.org/10.1145/3580305.3599533>, 2023.
- Feng, D., Beck, H., Lawson, K., and Shen, C.: The Suitability of Differentiable, Physics-Informed Machine Learning Hydrologic Models for Ungauged Regions and Climate Change Impact Assessment, *Hydrology and Earth System Sciences*, 27, 2357–2373, <https://doi.org/10.5194/hess-27-2357-2023>, 2023.
- 820
- Feng, L., Tung, F., Ahmed, M. O., Bengio, Y., and Hajimirsadeghi, H.: Were RNNs All We Needed?, <https://doi.org/10.48550/arXiv.2410.01201>, 2024.
- Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., Qualls, L. M., Gupta, H. V., and Nearing, G. S.: 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>, 2022.
- 825
- Gauch, M., Mai, J., and Lin, J.: 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>, 2021.



- Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the Mean Squared Error and NSE Performance Criteria: Implications for Improving Hydrological Modelling, *Journal of Hydrology*, 377, 80–91, <https://doi.org/10.1016/j.jhydrol.2009.08.003>, 2009.
- Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, *Neural Computation*, 9, 1735–1780, <https://doi.org/10.1162/neco.1997.9.8.1735>, 1997.
- Hong, Y.-C., Xiao, B., and Chen, Y.: TSKANMixer: Kolmogorov-Arnold Networks with MLP-Mixer Model for Time Series Forecasting, <https://doi.org/10.48550/arXiv.2502.18410>, 2025.
- Houska, T., Kraft, P., Chamorro-Chavez, A., and Breuer, L.: SPOTting Model Parameters Using a Ready-Made Python Package, *PLOS ONE*, 10, e0145180, <https://doi.org/10.1371/journal.pone.0145180>, 2015.
- Kartal, V.: Machine Learning-Based Streamflow Forecasting Using CMIP6 Scenarios: Assessing Performance and Improving Hydrological Projections and Climate Change, *Hydrological Processes*, 38, e15204, <https://doi.org/10.1002/hyp.15204>, 2024.
- Klemeš, V.: Operational Testing of Hydrological Simulation Models, *Hydrological Sciences Journal*, 31, 13–24, <https://doi.org/10.1080/02626668609491024>, 1986.
- Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., and Nearing, G.: Uncertainty Estimation with Deep Learning for Rainfall–Runoff Modeling, *Hydrology and Earth System Sciences*, 26, 1673–1693, <https://doi.org/10.5194/hess-26-1673-2022>, 2022.
- Knoben, W. J. M., Freer, J. E., and Woods, R. A.: Technical Note: Inherent Benchmark or Not? Comparing Nash–Sutcliffe and Kling–Gupta Efficiency Scores, *Hydrology and Earth System Sciences*, 23, 4323–4331, <https://doi.org/10.5194/hess-23-4323-2019>, 2019.
- Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall–Runoff Modelling Using Long Short-Term Memory (LSTM) Networks, *Hydrology and Earth System Sciences*, 22, 6005–6022, <https://doi.org/10.5194/hess-22-6005-2018>, 2018.
- Kratzert, F., Herrnegger, M., Klotz, D., Hochreiter, S., and Klambauer, G.: NeuralHydrology – Interpreting LSTMs in Hydrology, in: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, edited by Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R., pp. 347–362, Springer International Publishing, Cham, ISBN 978-3-030-28954-6, https://doi.org/10.1007/978-3-030-28954-6_19, 2019a.
- Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning, *Water Resources Research*, 55, 11344–11354, <https://doi.org/10.1029/2019WR026065>, 2019b.
- Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: 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>, 2019c.
- Kratzert, F., Gauch, M., Klotz, D., and Nearing, G.: HESS Opinions: Never Train an LSTM on a Single Basin, *Hydrology and Earth System Sciences Discussions*, pp. 1–19, <https://doi.org/10.5194/hess-2023-275>, 2024.
- Kruskal, W. H. and Wallis, W. A.: Use of Ranks in One-Criterion Variance Analysis, *Journal of the American Statistical Association*, <https://doi.org/10.1080/01621459.1952.10483441>, 1952.
- Liu, J., Bian, Y., Lawson, K., and Shen, C.: Probing the Limit of Hydrologic Predictability with the Transformer Network, *Journal of Hydrology*, 637, 131389, <https://doi.org/10.1016/j.jhydrol.2024.131389>, 2024.



- 865 Liu, J., Shen, C., O'Donncha, F., Song, Y., Zhi, W., Beck, H. E., Bindas, T., Kraabel, N., and Lawson, K.: From RNNs to Transformers: Benchmarking Deep Learning Architectures for Hydrologic Prediction, *Hydrology and Earth System Sciences*, 29, 6811–6828, <https://doi.org/10.5194/hess-29-6811-2025>, 2025.
- Martel, J.-L., Brissette, F., Arsenault, R., Turcotte, R., Castañeda-Gonzalez, M., Armstrong, W., Mailhot, E., Pelletier-Dumont, J., Rondeau-Genesse, G., and Caron, L.-P.: Assessing the Adequacy of Traditional Hydrological Models for Climate Change Impact
870 Studies: A Case for Long Short-Term Memory (LSTM) Neural Networks, *Hydrology and Earth System Sciences*, 29, 2811–2836, <https://doi.org/10.5194/hess-29-2811-2025>, 2025.
- Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L.: Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations, *Transactions of the ASABE*, 50, 885–900, 2007.
- Moriasi, D. N., Gitau, M. W., Pai, G., and Daggupati, P.: Hydrologic and Water Quality Models: Performance Measures and Evaluation
875 Criteria, *Transactions of the ASABE*, 58, 1763–1785, 2015.
- Natel de Moura, C., Seibert, J., and Detzel, D. H. M.: Evaluating the Long Short-Term Memory (LSTM) Network for Discharge Prediction under Changing Climate Conditions, *Hydrology Research*, 53, 657–667, <https://doi.org/10.2166/nh.2022.044>, 2022.
- 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., and Duan, Q.: Development of a Large-Sample Watershed-Scale Hydrometeorological Data Set for the Contiguous USA: Data Set
880 Characteristics and Assessment of Regional Variability in Hydrologic Model Performance, *Hydrology and Earth System Sciences*, 19, 209–223, <https://doi.org/10.5194/hess-19-209-2015>, 2015.
- Newman, A. J., Mizukami, N., Clark, M. P., Wood, A. W., Nijssen, B., and Nearing, G.: Benchmarking of a Physically Based Hydrologic Model, *Journal of Hydrometeorology*, 18, 2215–2225, <https://doi.org/10.1175/JHM-D-16-0284.1>, 2017.
- Nguyen, Q., Shrestha, S., Ghimire, S., Sundaram, S. M., Xue, W., Virdis, S. G. P., and Maharjan, M.: Application of Machine Learning
885 Models in Assessing the Hydrological Changes under Climate Change in the Transboundary 3S River Basin, *Journal of Water and Climate Change*, 14, 2902–2918, <https://doi.org/10.2166/wcc.2023.313>, 2023.
- O, S., Dutra, E., and Orth, R.: Robustness of Process-Based versus Data-Driven Modeling in Changing Climatic Conditions, *Journal of Hydrometeorology*, 21, 1929–1944, <https://doi.org/10.1175/JHM-D-20-0072.1>, 2020.
- Priestley, C. H. B. and Taylor, R. J.: On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters, *Monthly
890 Weather Review*, 100, 81–92, [https://doi.org/10.1175/1520-0493\(1972\)100<0081:OTAOSH>2.3.CO;2](https://doi.org/10.1175/1520-0493(1972)100<0081:OTAOSH>2.3.CO;2), 1972.
- Rasiya Koya, S. and Roy, T.: Temporal Fusion Transformers for Streamflow Prediction: Value of Combining Attention with Recurrence, *Journal of Hydrology*, 637, 131 301, <https://doi.org/10.1016/j.jhydrol.2024.131301>, 2024.
- Sankarasubramanian, A., Vogel, R. M., and Limbrunner, J. F.: Climate Elasticity of Streamflow in the United States, *Water Resources Research*, 37, 1771–1781, <https://doi.org/10.1029/2000WR900330>, 2001.
- 895 Sarkar, A., Hakimi, A., Chen, X., Huang, H., Lu, C., Demir, I., and Jannesari, A.: HydroGAT: Distributed Heterogeneous Graph Attention Transformer for Spatiotemporal Flood Prediction, in: *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems*, pp. 1019–1030, <https://doi.org/10.1145/3748636.3764172>, 2025.
- Seibert, J., Vis, M. J. P., Lewis, E., and van Meerveld, H.: Upper and Lower Benchmarks in Hydrological Modelling, *Hydrological Processes*, 32, 1120–1125, <https://doi.org/10.1002/hyp.11476>, 2018.
- 900 Sha, J., Chang, Y., and Liu, Y.: Hydrological Response to Climate Change: McGAN for Multi-Site Scenario Weather Series Generation and LSTM for Streamflow Modeling, *Atmosphere*, 15, 1348, <https://doi.org/10.3390/atmos15111348>, 2024.



- Singh, D., Vardhan, M., Sahu, R., Chatterjee, D., Chauhan, P., and Liu, S.: Machine-Learning- and Deep-Learning-Based Streamflow Prediction in a Hilly Catchment for Future Scenarios Using CMIP6 GCM Data, *Hydrology and Earth System Sciences*, 27, 1047–1075, <https://doi.org/10.5194/hess-27-1047-2023>, 2023.
- 905 Sofokleous, I., Zittis, G., Dörflinger, G., and Bruggeman, A.: Evaluating the Performance of Objective Functions and Regional Climate Models for Hydrologic Climate Change Impact Studies: A Case Study in the Eastern Mediterranean, <https://doi.org/10.5194/egusphere-2025-2478>, 2025.
- Thirel, G., Andréassian, V., Perrin, C., Audouy, J.-N., Berthet, L., Edwards, P., Folton, N., Furusho, C., Kuentz, A., Lerat, J., Lindström, G., Martin, E., Mathevet, T., Merz, R., Parajka, J., Ruelland, D., and Vaze, J.: Hydrology under Change: An Evaluation
910 Protocol to Investigate How Hydrological Models Deal with Changing Catchments, *Hydrological Sciences Journal*, 60, 1184–1199, <https://doi.org/10.1080/02626667.2014.967248>, 2015.
- Tolstikhin, I. O., Hounsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lucic, M., and Dosovitskiy, A.: MLP-Mixer: An All-MLP Architecture for Vision, in: *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 261–24 272, Curran Associates, Inc., 2021.
- 915 Tomczak, M. and Tomczak-Łukaszewska, E.: The Need to Report Effect Size Estimates Revisited. An Overview of Some Recommended Measures of Effect Size, *TRENDS in Sport Sciences*, 21, 19–25, 2014.
- Ullah, B., Fawad, M., Khan, A. U., Mohamand, S. K., Khan, M., Iqbal, M. J., and Khan, J.: Futuristic Streamflow Prediction Based on CMIP6 Scenarios Using Machine Learning Models, *Water Resources Management*, 37, 6089–6106, <https://doi.org/10.1007/s11269-023-03645-3>, 2023.
- 920 Vano, J. A., Das, T., and Lettenmaier, D. P.: Hydrologic Sensitivities of Colorado River Runoff to Changes in Precipitation and Temperature, *Journal of Hydrometeorology*, 13, 932–949, <https://doi.org/10.1175/JHM-D-11-069.1>, 2012.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, A., and Polosukhin, I.: Attention Is All You Need, in: *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.
- Wang, X., Deng, C., Yin, X., Wei, J., and Zou, J.-c.: Assessing Climate Change Impacts on Streamflow in Upper Han River
925 Basin Using Deep Learning Models Ensembled with Bayesian Model Averaging, *Water Science and Engineering*, 18, 412–421, <https://doi.org/10.1016/j.wse.2025.08.004>, 2025.
- Wei, X., Wang, G., and Schmalz, B.: Future Projections of China Runoff Changes Based on CMIP6 and Deep Learning, *Journal of Hydrology: Regional Studies*, 62, 102 998, <https://doi.org/10.1016/j.ejrh.2025.102998>, 2025.
- Wi, S. and Steinschneider, S.: Assessing the Physical Realism of Deep Learning Hydrologic Model Projections Under Climate Change,
930 *Water Resources Research*, 58, e2022WR032 123, <https://doi.org/10.1029/2022WR032123>, 2022.
- Wi, S. and Steinschneider, S.: On the Need for Physical Constraints in Deep Learning Rainfall–Runoff Projections under Climate Change: A Sensitivity Analysis to Warming and Shifts in Potential Evapotranspiration, *Hydrology and Earth System Sciences*, 28, 479–503, <https://doi.org/10.5194/hess-28-479-2024>, 2024.
- Yin, H., Guo, Z., Zhang, X., Chen, J., and Zhang, Y.: RR-Former: Rainfall-Runoff Modeling Based on Transformer, *Journal of Hydrology*,
935 609, 127 781, <https://doi.org/10.1016/j.jhydrol.2022.127781>, 2022.
- Zhong, S., Song, S., Zhuo, W., Li, G., Liu, Y., and Chan, S.-H. G.: A Multi-Scale Decomposition MLP-Mixer for Time Series Analysis, *Proceedings of the VLDB Endowment*, 17, 1723–1736, <https://doi.org/10.14778/3654621.3654637>, 2024.