Technical Note: High Nash Sutcliffe Efficiencies conceal poor simulations of interannual variance in seasonal regimes

Sacha W. Ruzzante[1], Wouter J. M. Knoben[2], Thorsten Wagener[3], Tom Gleeson[1], Markus Schnorbus[4]

[1]Department of Civil Engineering, University of Victoria, Canada
[2]Department of Civil Engineering, University of Calgary, Canada
[3]Institute for Environmental Science and Geography, University of Potsdam, Germany
[4]Pacific Climate Impacts Consortium, University of Victoria, Canada

Correspondence to: Sacha W. Ruzzante (sruzzante@uvic.ca)

Abstract. In highly seasonal regimes hydrologic models generally achieve high scores on common performance metrics such as the Nash-Sutcliffe Efficiency (NSE) and the Kling-Gupta Efficiency (KGE). However, variance in streamflow time series is composed of seasonal, interannual, and irregular variance, and the NSE and KGE do not differentiate between these components. Differences in performance on these three components have not been evaluated across a broad spectrum of hydrologic models and regions. We analyse open-access simulations from 18 regional and global hydrologic models. We find that these models consistently achieve the highest NSE and KGE in highly seasonal catchments where they are worse at simulating interannual variability, compared to less seasonal catchments. Simulated year-to-year changes in ecologically relevant hydrologic signatures are less accurate in highly seasonal catchments, and the NSE of the interannual variance component is usually lower. This suggests that these hydrologic models may struggle to predict long-term responses to climate change, especially in highly seasonal tropical, alpine, and polar regions, which are some of the most vulnerable to climate change. We encourage hydrologic modellers to explicitly evaluate skill at simulating interannual variability, rather than relying only on aggregate measures such as the NSE and KGE.

Short Summary:

Common metrics used to evaluate hydrologic models make it relatively easy to achieve high performance scores in highly seasonal catchments. However, we analysed 18 hydrologic models and found that almost all were worse at simulating interannual variability and change in seasonal streamflow regimes. This suggests that climate change impacts on streamflow may not be accurately predicted in highly seasonal tropical, alpine, and polar regions, which are highly vulnerable to climate change.

1 Introduction

Interannual variability and change in streamflow threatens social and ecological systems, as aquatic species are adapted to the natural streamflow regime (McMillan, 2021; Poff et al., 1997; Poff and Zimmerman, 2010), and unpredictable water availability hinders effective water resource management (Hall et al., 2014). The earth's changing climate is exposing non-stationarities in streamflow regimes (Milly et al., 2008), which have been detected in hydrologic signatures such as mean flows and the frequency, duration, and magnitude of hydrologic extremes (Gudmundsson et al., 2021; Ruzzante and Gleeson, 2025; Slater et al., 2021; Taye and Dyer, 2024; Xiong and Yang, 2024), and changes in the timing of streamflow (Berghuijs et al., 2025; Stewart et al., 2005; Wasko et al., 2020). However, there is a disconnect between how hydrologic change is assessed (using hydrologic signatures) and how hydrologic models are typically calibrated/trained and evaluated (using aggregate performance metrics), which may lead to inaccurate predictions of hydrologic change.

Streamflow time series can be conceptualized as the sum of seasonal, interannual, and irregular variance components, with very different driving mechanisms for each component. Stochasticity in individual weather events drives irregular variance (sometimes called remainder, noise, or high-frequency variance) while more regular seasonal cycles of temperature and precipitation drive seasonal fluctuations of the hydrograph (Dralle et al., 2017). Interannual variance, on the other hand, can be driven by climate oscillations, climate change, and other non-stationarities such as vegetation responses to climate change.

Hydrologic models, in contrast, are typically assessed on their aggregated ability to reproduce a historical timeseries of observations (Eker et al., 2018). In a climate change context, we desire hydrologic models that accurately predict historical interannual variability and change (Milly et al., 2008; Montanari et al., 2013; Wagener et al., 2010) and use performance on historical data as a proxy for their performance in future cases. However, hydrologists rarely consider interannual variability separately and the most popular performance metrics (the Nash-Sutcliffe and Kling-Gupta Efficiencies, NSE and KGE) evaluate all three variance components in an aggregated manner (i.e., as a single number; Gupta et al., 2009; Kling et al., 2012; Nash and Sutcliffe, 1970).

However, good performance on one variability component does not guarantee good performance on the others because the driving mechanisms of interannual, seasonal, and irregular variability are different. We hypothesize that in catchments where one component historically dominates variability, model calibration/training will emphasize simulation accuracy of the dominant component, whereas the other components will be more poorly simulated. Specifically, we aim to test whether interannual variance is poorly modelled in catchments with a strong seasonal cycle, because locations where strong seasonal cycles are common (*e.g.,* tropical basins, snow-dominated mountain ranges, and high-latitude locations) are also extremely vulnerable to climate change.

The concept of a 'strong' seasonal cycle has been discussed at some length and quantified in the context of climatological benchmark models, which are typically defined as the interannual mean flow for each calendar day. Garrick et al. (1978) were among the first to propose that a model should outperform the climatological benchmark, and subsequent authors found that the climatological benchmark NSE values, here denoted as NSE_{cb}, can be very high (sometimes greater than 0.8) in high-mountain catchments (Martinec and Rango, 1989; Schaefli and Gupta, 2007). Knoben (2024) similarly found that benchmark KGEs are high in snow-dominated regions.

We aim to answer three questions:

1) Where are the interannual, seasonal, and irregular components of streamflow variance dominant?
2) Where is the climatological benchmark NSE high?
3) What does this mean for our ability to simulate long-term change in a nonstationary world?

In Section 4.1 we address question 1. We use time series decomposition on global stream gauge data from 18 open-access datasets and calculate the variance fraction for each component. We then address question 2 in Section 4.2 and calculate the climatological benchmark NSE for 20,338 stream gauges. In Section 2.1 we also explain that the climatological benchmark NSE is equivalent to the seasonal variance fraction.

Regarding question 3, we expect that hydrologic models will be worse at representing interannual variability (and long-term change) in highly seasonal catchments because in these catchments ‘high’ NSE scores can be achieved without accurately representing the hydrologic processes that lead to interannual variability (for example, in the climatological benchmark model). In Section 4.3 we test this hypothesis with open-access simulations from 18 hydrologic models.

2 Methods

2.1 Time Series Decomposition

To address question 1, we applied time series decomposition techniques to streamflow data from 17,245 gauges. From our compilation of 28,406 daily discharge time series (see section 3.1), we selected the 17,245 with at least 10 years of data without missing days. We decomposed each time series into seasonal, interannual, and irregular components. First, we deseasonalized the data: that is, we calculated the seasonal component as the mean of each calendar day and subtracted this from the observed time series to extract the deseasonalized anomalies. We calculated the Fast Fourier Transform of the anomalies and separated the Fourier frequencies into interannual (frequencies smaller than 2 year^{-1}) and irregular components (frequencies greater than or equal to 2 year^{-1}). We chose a cutoff frequency of 2 year^{-1}, or a period of 6 months, in order to classify variations in seasonality (*e.g.,* a wetter than normal summer) as interannual variance. Additional details and a flowchart are provided in Section S3.

This decomposition is orthogonal so the sum of the variances of the components ($\sigma^2_{interannual} + \sigma^2_{seasonal} + \sigma^2_{irregular}$) is equal to the variance of the observed streamflow time series (σ^2_o). For each catchment we calculated the variance fraction associated with each component, (*e.g.,* $\sigma^2_{seasonal}/\sigma^2_o$). Because the decomposition is orthogonal the three variance fractions sum to 1. Another result of this orthogonality is that the variance fraction is identical to the NSE for each component. For example, the seasonal variance fraction is equivalent to the climatological benchmark NSE_{cb}:

$$NSE_{cb} = 1 - \frac{\sigma_\epsilon^2}{\sigma_o^2} = 1 - \frac{\sigma_o^2 - \sigma_{seasonal}^2}{\sigma_o^2} = \frac{\sigma_{seasonal}^2}{\sigma_o^2}, \tag{1}$$

Where σ_ϵ^2 is the error variance of the climatological benchmark model.

This particular decomposition method is a novel contribution but has similarities to previous work. De-seasonalizing data to obtain anomalies is a standard technique (Wilks, 2006), and spectral analysis (including using Fourier transforms) has been applied to streamflow to characterize hydrologic regimes (Brown et al., 2023; Smith et al., 1998).

We considered other time series decomposition methods, including classical decomposition (Kendall and Stuart, 1966) and Seasonality and Trend decomposition using Loess (STL, Cleveland et al., 1990). Details are provided in Section S3, along with examples of streamflow time series decomposed with each method. However, classical decomposition does not allow the interannual component to vary seasonally, which means that the interannual component only represents changes in mean annual flow. In addition, neither classical nor STL decomposition result in orthogonal components, so the variance fractions do not necessarily sum to 1. Other decomposition methods are possible, including using wavelet transforms instead of Fourier analysis, or allowing the seasonal component to vary with time. However, such methods are left to further work, as the main points we wish to make are readily supported by our current approach.

2.2 Climatological Benchmark Performance

To answer our second question, we calculated the NSE_{cb} (the NSE for a climatological benchmark model defined as the interannual mean flow for each calendar day) for 20,338 catchments. We used all catchments with at least 10 years of observed daily discharge data; for this analysis we permitted gaps in the data, as long as each calendar day was observed at least 10 times. We used a leave-one-out cross validation scheme: the discharge for each year was predicted using observed discharge for all other years. We then concatenated the predictions and calculated the NSE_{cb} on the full time series. This cross-validation reduces NSE_{cb} such that it is no longer identical to the seasonal variance fraction, and $NSE_{cb} \leq \sigma_{seasonal}^2/\sigma_o^2$.

We also calculated the climatological benchmark KGE'_{cb} (the apostrophe denotes the modification by Kling et al., 2012) and include these results in Sect. S1. We focus on the NSE_{cb} here for brevity, because the KGE'_{cb} and NSE_{cb} exhibit similar global patterns, and because the NSE_{cb} is so closely related to the seasonal variance fraction. Lastly, we tested the robustness of NSE_{cb} to a differential split sample methodology and include these methods and results in Sect S9.

2.3 Representation of interannual and seasonal variability in hydrologic models

Our third question asks how the strength of the seasonal cycle affects the ability of hydrologic models to simulate interannual change. We analysed the simulated streamflow from 18 regional and global hydrologic models to investigate differences in interannual performance between highly seasonal and less seasonal catchments. The models are described in section 3.2.

2.3.1 Variance component NSE values

For each model, we calculated the overall NSE for each simulated catchment. We then decomposed both the simulated and observed time series using the strategy in Sect. 2.1 and calculated the NSE for each variance component. For example:

$$NSE_{interannual} = 1 - \frac{\sum_{t=1}^{N}(I_o(t)-I_s(t))^2}{\sum_{t=1}^{N}(I_o(t)-\bar{I_o})^2} \tag{2}$$

Where I_o and I_s are the observed and simulated interannual components as derived from time series decomposition (Section 2.1). The seasonal and irregular NSEs are calculated similarly. Section S4 shows that the overall NSE is equal to the weighted sum of the three component NSEs, where the weights are the variance fractions discussed in Sect. 2.1:

$$NSE = \frac{\sigma_{interannual}^2}{\sigma_o^2} NSE_{interannual} + \frac{\sigma_{seasonal}^2}{\sigma_o^2} NSE_{seasonal} + \frac{\sigma_{irregular}^2}{\sigma_o^2} NSE_{irregular} \tag{3}$$

We wanted to test whether the models perform better or worse in highly seasonal catchments, so for each model we classified the catchments into highly seasonal ($\sigma_{seasonal}^2/\sigma_o^2 > 0.5$) and less seasonal ($\sigma_{seasonal}^2/\sigma_o^2 \leq 0.5$) subsets. We then compared the NSE values between the highly seasonal and less seasonal subsets using the non-parametric Mann-Whitney U test (Mann and Whitney, 1947).

2.3.2 Simulated changes in hydrologic signatures

The NSEs for the interannual, seasonal, and irregular components provide a concise and holistic summary of performance for each type of variance. However, studies of hydrologic change are often concerned with predicting changes to hydrologic signatures relevant to ecology or water management, so it is useful to evaluate how well models simulate changes in these signatures over the historical period. To this end, we compared simulated and observed values of 41 hydrologic signatures calculated on an annual basis (Table 1). The ecologically-relevant signatures that we consider are the 32 indicators of hydrologic alteration proposed by Richter et al. (1996) in addition to the total number of days below the 25th percentile (Number of low flow days), the total number of days above the 75th percentile (Number of high flow days), the rising and falling limb densities, the streamflow concentration index, the half flow day, the mean annual flow, the slope of the midsegment of the flow duration curve, and the baseflow index (see Table 1 for references). These additional metrics have been widely used by hydrologists to characterize hydrologic regimes and to detect trends.

We are interested in whether the hydrologic models accurately reproduce interannual variability in these 41 signatures, so we calculated non-parametric correlation coefficients between the simulated and observed annual series of hydrologic signatures, using Spearman's ρ for most metrics. Two metrics (No. high pulses and No. low pulses) frequently have tied ranks, so for these we used Kendall's τ.

We also calculated five other popular performance metrics that evaluate interannual, seasonal, and irregular variance jointly: the KGE', KGE'(1/Q), and the three components of the KGE': Pearson r, the mean bias β and the ratio of coefficients of variation γ (Kling et al., 2012). To be consistent with the other metrics (for which values near 1 are better), we transformed β

and γ to the range $(-\infty, 1]$ using the transforms $\left(1-\sqrt{(\beta-1)^2}\right)$ and $\left(1-\sqrt{(\gamma-1)^2}\right)$. These transforms are analogous to the use of these terms in the KGE.

We applied the same tests here as for the variance component NSE values (Section 2.3.1): we compared the values of each metric between the highly seasonal and less seasonal subsets using Mann-Whitney U tests. We also performed the same analysis after splitting the catchments based on thresholds of 0.4 and 0.6 for the seasonal variance fractions, and splitting on the streamflow concentration index (Han et al., 2024), the coefficient of variation of the average hydrograph, the aridity seasonality, and the fraction of precipitation as snow (Knoben et al., 2018).

Table 1: Hydrologic signatures used to evaluate models' ability to reproduce interannual variability

Category	**Signature**	**Definition**	**Source**
Magnitude of Seasonal water conditions	$\bar{Q}_{Jan} \dots \bar{Q}_{Dec}$	Mean monthly flow for each calendar month	(Richter et al., 1996)
Magnitude and timing of annual extreme water conditions	Max n-day Q	Annual maximum flow for 1, 3, 7, 30, and 90 day rolling averaging	(Richter et al., 1996)
	Min n-day Q	Annual minimum flow for 1, 3, 7, 30, and 90 day rolling averaging	(Richter et al., 1996)
	$day\ of\ annual\ maximum$	Calendar day of annual 1-day maximum flow[1]	(Richter et al., 1996)
	$day\ of\ annual\ maximum$	Calendar day of annual 1-day minimum flow[1]	(Richter et al., 1996)
Frequency and duration of high and low pulses	$No.\ high/low\ pulses$	Number of periods with flow above 75th percentile (high pulse) or below 25th percentile (low pulse)	(Richter et al., 1996)
	$High/low\ pulse\ duration$	Average duration of high/low pulse in days	(Richter et al., 1996)
	$No.\ of\ high/low\ flow\ days$	Annual total days above 75th percentile (high flow) and below 25th percentile (low flow)	Less sensitive indicator of pulse duration than the two above; similar to Simeone et al

---

[1] The water year is rotated to begin 183 days before the maximum (minimum) flow day, to prevent artificially large disagreements between simulated and observed time series arising from maximum or minimum flow dates occurring just before or just after the beginning of the water year.

			(2024) Yin et al (2024)
Rate and frequency of water condition changes	$Mean\ daily\ rise/fall$	Average 1-day difference of all days with positive (rise) and negative (fall) differences	(Richter et al., 1996)
	$No.rises\ (falls)$	Annual number of rising/falling limbs	(Richter et al., 1996)
	$rising\ (falling)\ limb\ density$	Number of rising (falling) limbs divided by total number of days with increasing (decreasing) flow	(Shamir et al., 2005)
Other hydrologic signatures	QCI	Streamflow concentration index	(Han et al., 2024)
	$half\ flow\ day$	Day of water year at which half of the total annual streamflow has passed the gauge	(Court, 1962)
	$\bar{Q}_{annual}$	Mean annual streamflow	(Döll and Schmied, 2012)
	$slope\ FDC$	Slope of the flow duration curve between the 33rd and 66th percentiles	(Yadav et al., 2007)
	BFI	Annual baseflow index, using 3-pass Lyne-Hollick filter	(Ladson et al., 2013; Lyne and Hollick, 1979)

3 Data

3.1 Streamflow Data

We compiled streamflow data from 16 CAMELS-type datasets from Australia (Fowler et al., 2024), Brazil (Chagas et al., 2020), Central Europe (Klingler et al., 2021), Chile (Alvarez-Garreton et al., 2018), Denmark (Liu et al., 2024), France (Delaigue et al., 2024), Germany (Loritz et al., 2024), Great Britain (Coxon et al., 2020), Iceland (Helgason and Nijssen, 2024), India (Mangukiya et al., 2025), Israel/West Bank/Golan Heights (Efrat, 2025), Luxembourg (Nijzink et al., 2024), the United States (Addor et al., 2017; Newman et al., 2015), North America (Arsenault et al., 2020), Spain (Casado Rodríguez, 2023), and Switzerland (Höge et al., 2023). For countries not represented in the above datasets we used streamflow data from the Global Runoff Data Centre (GRDC, https://grdc.bafg.de/). We also added data from 138 Russian stations (Lammers and Shiklomanov, 2000) and ensured they were not duplicates of the GRDC stations. In total we compiled records from 28,406 stations worldwide.

3.2 Streamflow Simulations

We searched Google Scholar and used ChatGPT to identify freely available datasets of simulated streamflow at gauged locations. We required that some of the gauges be in the tropical, alpine, or polar regions where we expect the seasonal variance fraction to be high. Where publications reported the results of multiple versions of the same or very similar models, we selected the version identified by the authors as having the best performance.

In total we compiled simulations from 18 models this way: six Long Short-Term Memory Models (LSTMs), eleven process-based hydrologic models, and one hybrid model. These models are listed in Table 2, and Table S1 provides more extensive details about the calibration procedures for each model and the degree of human alteration of the catchments.

Where possible we included only near-natural catchments in the evaluation of each model, either as defined by the authors or by referencing other published lists of near-natural catchments (Falcone, 2011; Newman et al., 2015; Pellerin and Nzokou Tanekou, 2020). For the two Brazilian models, we used only catchments without regulation, with less than 5% impervious surfaces, and consumptive use less than 5% of annual streamflow. For the two global models published by Nearing et al (2024) we included all available catchments since we lacked a reliable way to identify near-natural catchments.

Where models used a split-sample approach to training and testing we used the testing subset. The type of split sample is listed in Table S1. Some models were tested in unseen catchments, some were tested in unseen time periods, and some did not employ a split-sample approach or did not provide sufficient detail to determine which subset of the simulation data was not seen during training/calibration. Six of the models (CE-COSERO, US-FUSE, US-HBV, US-mHM, US-SAC-SMA, and US-VIC) were basin-calibrated, while the rest were globally calibrated.

Table 2: 18 models for which we reanalysed simulations to test performance on interannual, seasonal, and irregular variance components.

Model	Type	Region	Model reference
GLOB-LSTM1	Lumped LSTM	Global	(Nearing et al., 2024)
GLOB-LSTM2	Lumped LSTM	Global	(Yang et al., 2025)
BR-LSTM	Lumped LSTM	Brazil	Section S7
CH-LSTM	Lumped LSTM	Switzerland	(Kraft et al., 2025)
ENA-LSTM	Lumped LSTM	Northeast North America	(Arsenault et al., 2023)
US-LSTM	Lumped LSTM	Conterminous United States	(Kratzert et al., 2024)
US-δHBV2.0UH	Hybrid: Semi-distributed differentiable process-based model	Conterminous United States	(Song et al., 2025)
GLOB-GloFAS	Distributed process-based model	Global	(Nearing et al., 2024)

BR-MGB-SA	Semi-distributed process-based model	Brazil	(Chagas et al., 2020; Siqueira et al., 2018)
CE-COSERO	Lumped process-based model	Central Europe	(Klingler et al., 2021)
CH-PREVAH	Distributed process-based model	Switzerland	(Kraft et al., 2025)
US-NHM	Distributed process-based model	Conterminous United States	(Regan et al., 2019)
US-FUSE	Lumped process-based model	Conterminous United States	(Kratzert, 2019)
US-HBV	Lumped process-based model	Conterminous United States	(Kratzert, 2019; Seibert et al., 2018)
US-mHM	Lumped process-based model	Conterminous United States	(Kratzert, 2019; Mizukami et al., 2019)
US-SAC-SMA	Lumped process-based model	Conterminous United States	(Kratzert, 2019; Newman et al., 2017)
US-VIC	Lumped process-based model	Conterminous United States	(Kratzert, 2019; Newman et al., 2017)
WNA-VIC-GL	Distributed process-based model	Western North America	(Schnorbus, 2018, 2020)

4 Results and Discussion

4.1 Global Distribution of seasonal, interannual and irregular variance

Figure 1 (a) shows the fraction of variance associated with seasonal, interannual, and irregular variance for 17,245 catchments. Globally, irregular variance dominates: more than half the variance is irregular in 70% of the catchments. Figure S19 shows histograms of each variance fraction.

Streams in arid regions (such as the ephemeral Oued Kert in Morocco, Figure 1 (d)) are especially irregular, because the streamflow time series are composed of infrequent flash floods driven by episodic heavy rainfall (D’Odorico and Bhattachan, 2012; Smith et al., 2015). However, flashy catchments in humid regions can also have high irregular variance fractions. Additional examples of highly irregular streams are included in Figs. S26-S31, from arid catchments in Telangana (India), New Mexico (USA), and Kunene (Namibia), and humid catchments in Newfoundland (Canada), Narvik (Norway) and Westland (New Zealand).

Highly seasonal catchments (those with a seasonal variance fraction above 0.5) are found primarily in cold (polar and alpine), and tropical climates, where seasonality is driven either by snow accumulation and melt or by strong monsoons. Figure 1 (c) shows the Candeias River, a highly seasonal tropical catchment with some variation from year to year. The seasonal variance fraction is very high (greater than 0.8) in only 1% of catchments, but these extremely seasonal catchments are found on all continents except Oceania. Extremely seasonal catchments are found in the Arctic regions of Nunavut, Nunavik, Iceland, Sápmi, and Siberia, the alpine ranges of western North America, Europe, southern Patagonia, and central Asia, and the tropical Orinoco, Amazon, Niger, Congo, Irrawaddy, and Mekong basins. Additional examples of decomposed time series from extremely seasonal cold and tropical catchments are provided in Figures S32-S38.

High interannual variability is rare and occurs mainly in catchments with large surface or groundwater storage and/or strong connections to climate oscillations. Figure 1 (b) shows the decomposition for the Sturgeon Weir River in northern Canada, where strong connections to the Arctic Oscillation drive decadal-scale variability (St. Jacques et al., 2014) and seasonal as well as irregular variation is dampened by a large lake. Other regions with high interannual variance are: i) semi-arid north-central Chile (*e.g.,* Fig. S19), where warm phases of El Niño Southern Oscillation (ENSO) are associated with heavy rainfall, including major floods in 1997 (Araya et al., 2022), ii) the Paraguay River Basin (*e.g.,* Fig S22), where interannual persistence in dry and wet conditions is linked to the extensive Pantanal (wetland) hydrology as well as ENSO, the Pacific Decadal Oscillation, and the Atlantic Multidecadal Oscillation (Santos and Slater, 2025), and iii) southeastern England and northwestern France (*e.g.,* Fig S20), where variability is driven by the North Atlantic Oscillation (Rodwell et al., 1999; West et al., 2022), and the occurrence of record flooding from 2000-2001 (Marsh and Dale, 2002). Anthropogenic impacts also have the potential to cause interannual variability, such as in the Syr Darya (Kazakhstan) where water abstraction increased beginning with the expansion of irrigation canals in 1973 (Zou et al., 2019) (Fig S25). The preceding examples are selected to illustrate the diverse drivers of interannual variability around the world. In-depth analysis of their hydrologic conditions is considered out of scope for this work.

Hydrologic models should be capable of simulating all three variance components, but accurate simulation of interannual variance is arguably the most important when the objective is to predict long-term changes in statistical properties of streamflow, such as for climate change impact research. Accurately simulating interannual variance is probably an easier task in catchments that have historically been very interannually variable (such as the Sturgeon Weir River) than it is in catchments that have been interannually stationary, because there is more variance with which to calibrate hydrologic models. Nevertheless, historically stationary regimes are not guaranteed to remain stationary, so we believe this difficult task is worthwhile (Gudmundsson et al., 2012; Milly et al., 2008; Safeeq et al., 2014).

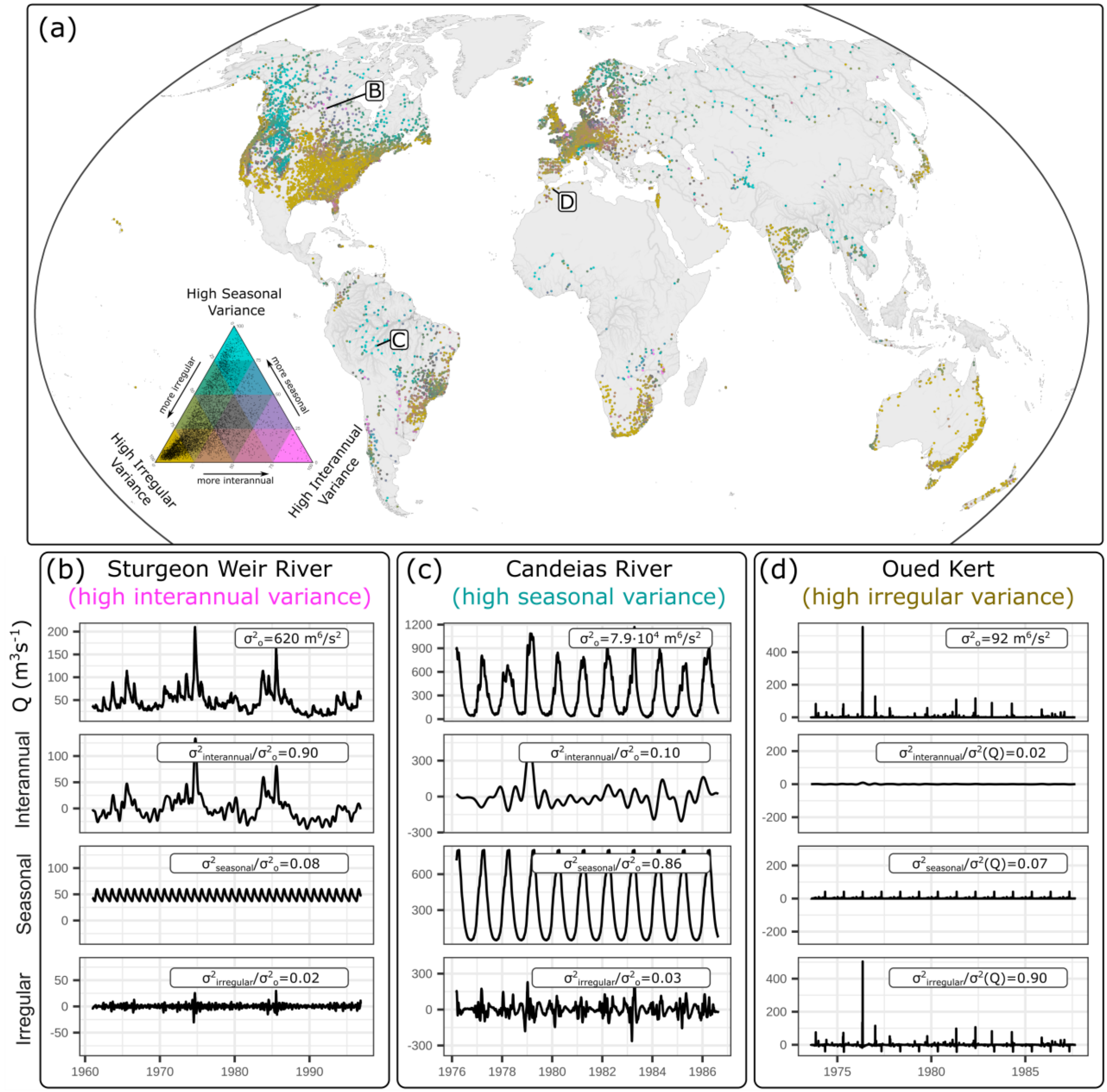


Figure 1: (a) The fraction of variance that is seasonal, interannual, and irregular. To reduce overplotting, gauges have been aggregated to 1 per 2500 km^2 using the mean of each variance fraction. The three panels (b), (c), and (d) show the decomposed time series for three example rivers that exhibit high variance fractions for each of the three components. (b) The Sturgeon Weir River at the outlet of Amisk Lake (Water Survey of Canada ID 05KG002, catchment area 14,600 km^2), an interannually variable stream (90% interannual variance). (c) Santa Isabel (Candeias River at Candeias do Jamari, Agência Nacional de Águas e Saneamento Básico ID 15550000, catchment area 12,700 km^2), a highly seasonal stream (86% seasonal variance). (d) Oued Kert at Driouch, an ephemeral stream in Morocco (Global Runoff Data Centre ID 1304800, catchment area 1,353 km^2), where 90% of the variance is irregular. The mapped river network is derived from HydroRIVERS v1.0 (Lehner and Grill, 2013).

4.2 Out-of-sample climatological benchmark NSE (NSE_{cb})

Figure 2 (a) shows the NSE_{cb} for 20,338 catchments, based on leave-one-out cross-validation. Overall, the median is small with a value of 0.11, which implies that streamflow in most catchments is largely unpredictable based only on climatology from other years. On the other hand, the NSE_{cb} is high (greater than 0.5) for 10% of the gauges. We argue that special care is warranted when modelling these catchments to ensure they add information beyond what is contained in the climatology.

Figure 2 (b) shows that high NSE_{cb} values tend to occur in tropical monsoon and cold/polar Köppen-Geiger climate zones. Figure 2 insets (c)-(i) show hydrographs from several catchments with $NSE_{cb}>0.8$, from arctic, alpine, and tropical locations. Very high NSE_{cb} values do rarely occur elsewhere, although we note that some of these catchments are large and cover multiple climate zones. For example, the Mekong and Irrawaddy are classified as 'temperate' based on catchment average climate data, but they are more accurately described as a mix of polar climate zones (at their headwaters in the Tibetan Plateau), temperate zones through their midsections, and tropical zones nearer to the gauging stations.

Figure S1 in the supplementary material shows the KGE'_{cb} for the 20,338 catchments. The patterns are very similar to those seen with the NSE_{cb}, and Sect. S1 shows that KGE'_{cb} and NSE_{cb} are uniquely and monotonically related if no cross-validation scheme is used. After implementing the leave-one-out cross-validation scheme, this relationship is modified by the number of years of data. Thus, in this article we focus on the NSE_{cb} but we point out that the KGE'_{cb} behaves similarly.

The cross-validation scheme ensures the climatological model is always tested on unseen data, but a more difficult test is the differential split sample, where models are tested on data outside of their calibration conditions. This has been widely applied and recommended to test models used for climate change impact assessment (Klemeš, 1986; Krysanova et al., 2018; Refsgaard et al., 2014; Seibert, 2003). In Sect. S9 we show that the NSE_{cb} remains high in tropical, alpine, and polar catchments when evaluated using a differential split sample methodology. The climatological benchmark model is, by definition, unable to simulate interannual variance or change, so the fact that it can achieve high NSE values when tested on data outside of its calibration conditions further reinforces that high NSE values do not guarantee a model is useful for making hydrologic predictions under climate change.

This is, to the best of our knowledge, the largest and most geographically extensive compilation of benchmark performance values for streamflow gauging stations to date. Figure 2 serves as a reminder the NSE is not an absolute measure of performance, and that comparing NSE values across catchments is challenging, because baseline performance varies substantially (Knoben, 2024; Martinec and Rango, 1989; Schaefli and Gupta, 2007; Seibert, 2001). Our analysis builds on previous work by showing that NSE_{cb} can be high even when evaluated on unseen data. In this work, however, we are primarily interested in analysing if the ease of achieving high NSE scores in some catchments jeopardizes the modelling of interannual variance. This is the subject of the following section.

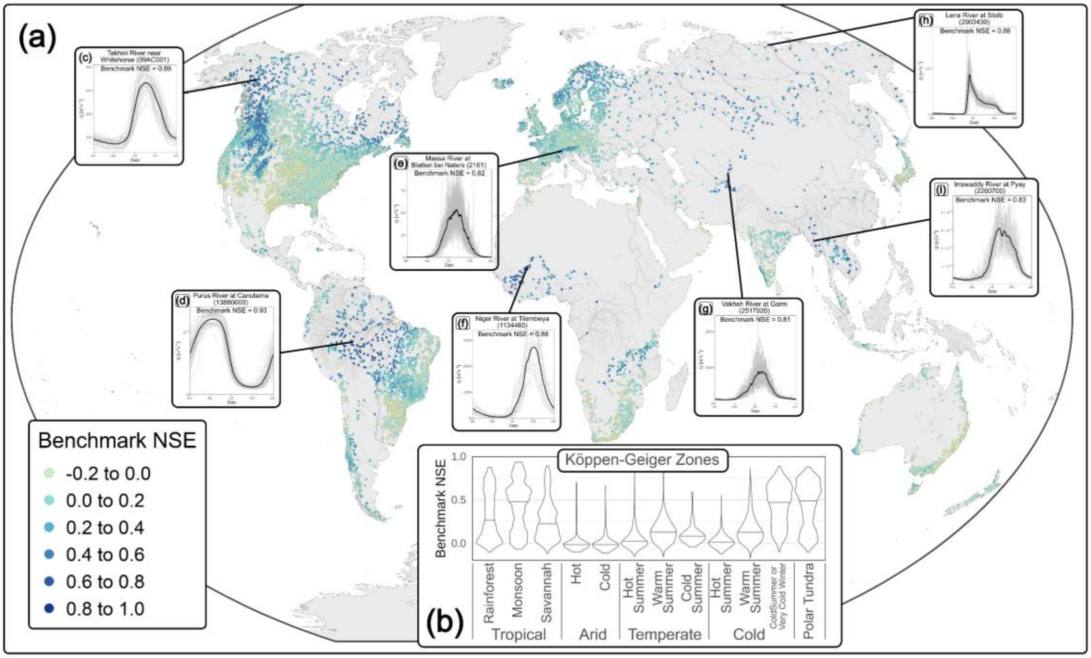


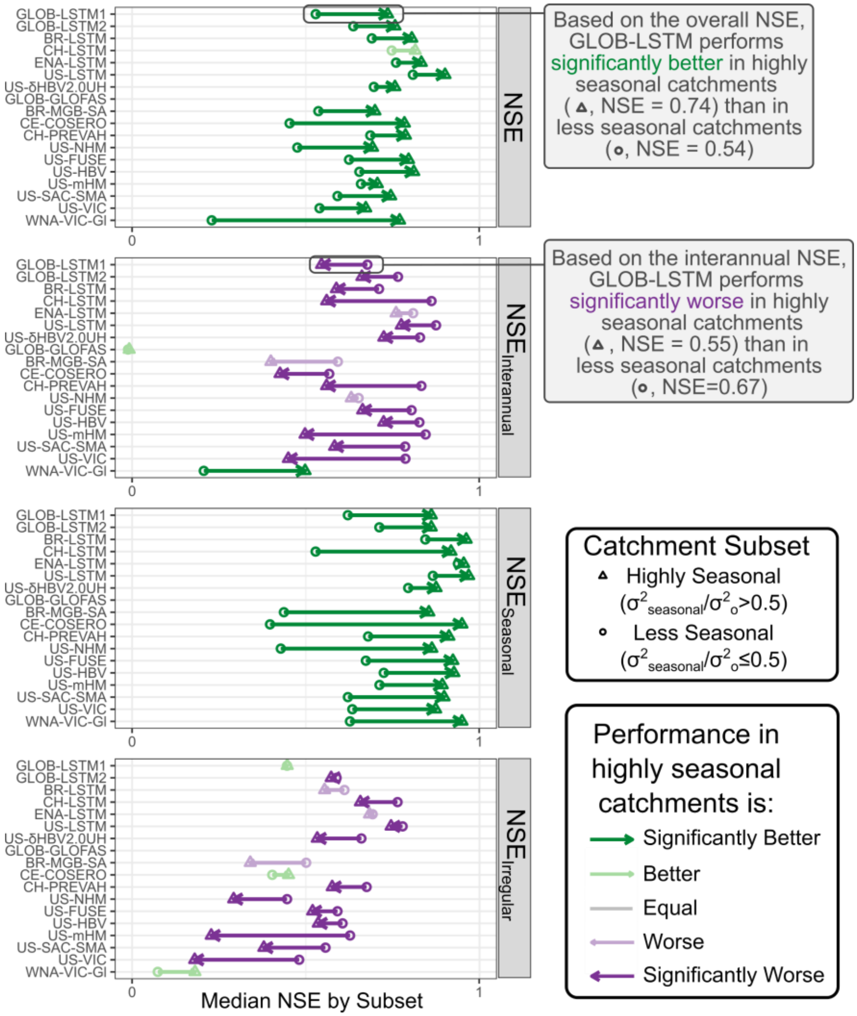
Figure 2 (a): Climatological benchmark Nash Sutcliffe Efficiencies for 20,338 catchments. To reduce overplotting, for panel A gauges have been aggregated to 1 per 2500 km^2 using the median NSE. (b): Distribution of the climatological benchmark NSE by Köppen-Geiger climate zone. Cold (high alpine and polar) and tropical climates have high benchmark NSE values, often upwards of 0.5 and occasionally upwards of 0.8. (c)-(i): Annual hydrographs from catchments with very high benchmarks (BNSE>0.8). The grey lines are individual years and the solid black line is the mean flow for each calendar day. The mapped river network is derived from HydroRIVERS v1.0 (Lehner and Grill, 2013).

4.3 High NSEs can hide poor representations of interannual variance

4.3.1: Variance component NSE values

Figure 3 shows that high NSE values often hide inferior simulations of interannual and irregular variability. The overall NSE is consistently higher in highly seasonal catchments, but this is mainly driven by an improvement in modelling the seasonal component ($NSE_{seasonal}$). $NSE_{interannual}$ values are lower in highly seasonal catchments for all models except GLOB-GLOFAS (where the median $NSE_{interannual}$ value is near zero for both highly seasonal and less seasonal catchments), and WNA-VIC-GL. $NSE_{irregular}$ values are also usually lower in highly seasonal catchments.

One might expect that higher NSEs can be achieved in highly seasonal catchments because seasonal variability is easy to model, and there is more of it. However, Fig. 3 suggests this is only part of the story: in highly seasonal catchments there is more seasonal variance, *and* a larger fraction of it is modelled accurately ($NSE_{seasonal}$ is larger).



Based on the overall NSE, GLOB-LSTM performs significantly better in highly seasonal catchments (△, NSE = 0.74) than in less seasonal catchments (○, NSE = 0.54)

Based on the interannual NSE, GLOB-LSTM performs significantly worse in highly seasonal catchments (△ , NSE = 0.55) than in less seasonal catchments (○, NSE=0.67)

Figure 3: Comparison of NSE scores between highly seasonal and less seasonal catchments, across 18 hydrologic models (vertical axis labels correspond to Table 2). For all models the NSEs of the overall time series and of the seasonal component are better in highly seasonal catchments, but the NSEs of the interannual and irregular components are usually significantly worse. The arrows point from the median value for less-seasonal catchments to the median value for highly seasonal catchments. Significance is determined at $p<0.05$ using the unpaired Mann-Whitney U-test. The median NSE was negative for GLOB-GLOFAS for all components and subsets.

4.3.2: Simulated changes in hydrologic signatures

Figure 4 shows the expanded analysis for 6 typical performance metrics and 41 interannual signature metrics (see Table 1). The typical metrics (top row) are higher in highly seasonal catchments. NSE, KGE', Pearson r, and variance ratio γ are better in highly seasonal catchments across all 13 models, and most of these differences are statistically significant. KGE' (1/Q) shows a similar pattern except for some of the process-based models for the United States, which struggle to simulate winter low flows in colder regions.

In the seasonal catchments where typical performance metrics are high, we again see that interannual variability is more poorly simulated: the correlations between observed and simulated annual values of 41 hydrologic signatures across 13 models are worse in highly seasonal catchments for 80% of the cases (63% at a significance level of $p=0.05$). All models except WNA-VIC-GL perform worse in highly seasonal catchments across most (>50%) of the interannual signature metrics, and all but one of these metrics are lower in highly seasonal catchments across most of the models.

Some of the differences in performance are quite large. For example, the Spearman rank correlations for monthly flows from November to April are close to 1 (perfect) for less seasonal catchments and below 0.5 for some models in highly seasonal catchments. The predictions of annual minima, high and low pulses, and rising and falling limbs are also substantially worse in highly seasonal catchments.

These non-parametric correlations test if the models correctly predict the direction of change, but not the absolute values of each signature. We also calculated the NSE of the simulated and observed annual series of hydrologic signatures, but all models struggle to simulate absolute values of at least some of these signatures: across 738 model-by-metric combinations, the NSE of the simulated hydrologic signatures is negative for both the highly-seasonal and less-seasonal subsets 48% of the time (Fig. S11). We view large, positive correlations with historical observations as a minimum requirement to consider a model useful for simulating hydrologic responses to a changing climate.

In Section S6 we present various robustness tests of these results. We tried changing the splitting threshold (seasonal variance fractions of 0.4 and 0.6, instead of 0.5). Our analysis and conclusions are insensitive to these changes. We also tried splitting the catchments using the snow fraction, and for snow-dominated regions the results were similar to Fig. 4. Splitting on several indices did not produce clear patterns.

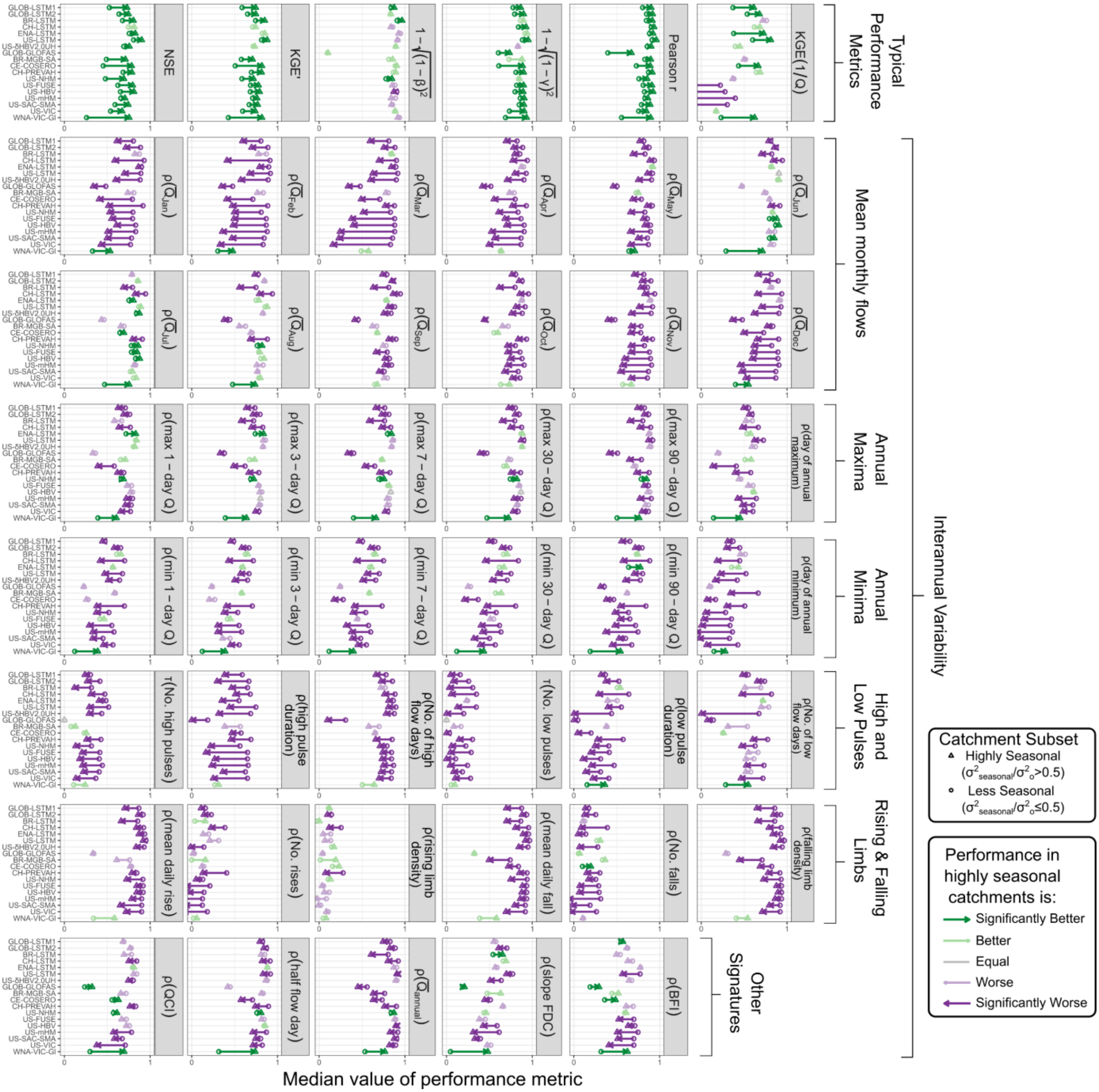


Figure 4: Catchments in high-benchmark NSE regions generally perform better than low-benchmark catchments on typical performance metrics which combine intra and interannual variance (top row) but are significantly worse when evaluated on other metrics that focus on interannual variability. The arrows point from the median value for low-benchmark catchments to the median value for high-benchmark catchments. Significance is determined at $p<0.05$ using the unpaired Mann-Whitney U-test. Note that the median NSE, KGE', and KGE' (1/Q) are negative for GLOB-GLOFAS in both highly seasonal and less seasonal catchments.

4.3.3: Causes of poor interannual performance

Our analysis does not directly identify the causes of poor interannual performance in highly seasonal catchments, but we hypothesize that model optimization algorithms, model structures, and data quality and quantity all play a role.

First, model optimization algorithms may bias training away from highly seasonal catchments. Seasonal variance is predictable and can be reproduced quite accurately with models of low complexity (Knoben, 2024). Where models are optimized simultaneously across many catchments, optimization algorithms may therefore greedily simulate the seasonal variance in seasonal catchments and neglect interannual and irregular variance. When seasonal variance is well-simulated, algorithms will prioritize improvements to modelling the irregular and interannual components in less-seasonal catchments where these improvements result in the largest increase to the average NSE. Learned human biases for what a ‘good’ simulation looks like could also bias modellers to neglect catchments where errors appear small compared to the seasonal pattern, since expert opinions and current quantitative metrics have been shown to be mostly consistent (Gauch et al., 2023). However, six of the models that we analysed were calibrated individually to each catchment, so this explanation is not sufficient on its own.

Second, model structures often do not include important hydrologic components for highly seasonal catchments, such as glacier change, avalanching, and vegetation-moisture feedbacks (D’Odorico et al., 2007; Köplin et al., 2013; Staal et al., 2020). These components can be drivers of interannual variability and change, and we hypothesize that their inclusion could improve simulations of interannual variability.

Third, data quality and quantity are often lower in remote polar, alpine, and tropical regions. Long-term weather stations tend to be scarce in these regions, which means that the gridded meteorological data used to run models may not accurately capture interannual climate variability (Burton et al., 2018), and therefore interannual streamflow variability will also be poorly modelled. There are also few gauged streams in highly seasonal regimes globally (Krabbenhoft et al., 2022), and optimization algorithms that aim to maximize the average performance across many catchments will not prioritize improving simulations for a small number of highly seasonal catchments.

Indeed, the global distribution of stream gauges is very biased (Krabbenhoft et al., 2022), and highly seasonal catchments are underrepresented in the datasets we analysed. Cold climates with cold or very cold winters (as defined in the Köppen-Geiger climate classification by Beck et al., 2018) represent 15% of global non-frozen lands (excluding perennially ice-covered regions) but only 10% of the catchments in our selected datasets, and polar tundra occupies 6.7% of global land and just 2.4% of catchments. The tropical zone is even more underrepresented: tropical rainforest and tropical monsoon climates represent 6.2% and 4% of global lands, and collectively generate 39% of global runoff (based on Beck et al., 2018 and Ghiggi et al., 2019), but only 0.9% and 0.8% of catchments in our datasets are located in these zones, respectively.

Other compilations of streamflow data are even more biased, particularly in underrepresenting tropical regions. Caravan, a popular compilation of 6830 catchments (Kratzert et al., 2023), included only 8 tropical rainforest and 5 tropical monsoon catchments in its initial version. Several ‘community extensions’ (*e.g.,* Färber et al., 2025) have improved the global coverage of the Caravan project, but the distribution remains highly biased. Another initiative, the Reference Observatory of Basins for

INternational hydrological climate change detection (ROBIN), includes streamflow data and catchment polygons for 2265 near-natural streams worldwide, but only 19 tropical rainforest and 6 tropical monsoon catchments (Turner et al., 2025). Our understanding of hydrologic processes and change in highly seasonal regimes is thus impeded not just by poor modelling of interannual variability, but also by the underrepresentation of these regimes in the datasets that are currently available and widely used in the hydrological modelling community. Efforts to increase the representation of these regimes could include making existing data publicly available (*e.g.,* Lin et al., 2023), digitizing paper records (*e.g.,* Bathelemy et al., 2024; Henck et al., 2011), or installing new stream gauges. However, we acknowledge that at least some of the causes of global gauge biases are not easily overcome (Krabbenhoft et al., 2022), and that we must continue to rely on approaches to predictions in ungauged basins (Hrachowitz et al., 2013).

5 Conclusions & Recommendations

Streamflow time series are made up of interannual, seasonal, and irregular components, and models can perform very differently with respect to these three components. These differences are obscured when aggregated performance metrics such as NSE and KGE are used. Our findings have several relevant consequences.

First, we provide further evidence that using the same performance thresholds to judge hydrological models across catchments is ill advised. Especially in tropical, alpine, and polar climates it is generally easier to achieve high NSE values: the climatological benchmark NSE_{cb} is often higher than 0.5 and occasionally even higher than 0.8. We observe, in Figures 3 and 4, that hydrologic models do achieve higher NSE and KGE values in these highly seasonal, high-benchmark catchments. It is therefore important to contextualize model performance with climatological and other benchmark models (Knoben, 2024).

Second, it is critical to choose evaluation metrics that are suitable for both the study location and model purpose. We show that high NSE values often hide inferior simulations of interannual variance, including changes in ecologically relevant hydrologic signatures. This is most evident in tropical, alpine, and polar regions, where most of the variance in streamflow is seasonal. Poor interannual performance in these regions (and in some cases almost complete failure to simulate year-to-year variability) raises concerns about the ability of these models to accurately simulate nonstationary hydrologic processes and responses to climate change. This is especially worrying because these regions may be some of the most vulnerable to climate change (Flores et al., 2024; Pepin et al., 2022; Rantanen et al., 2022) and are historically less-well studied regarding hydrologic extremes (Stein et al., 2024). We recommend that authors evaluate how well models simulate interannual variance to understand how well a model may extrapolate to different locations or climate conditions.

We encourage the community to pay more attention to interannual variance and to the highly seasonal regimes where it is most poorly modelled. This could include developing new calibration targets and objective functions to train models that improve the representation of interannual variance. Some suitable calibration targets include the interannual NSE introduced in section 2.3, and the correlations of the hydrologic signatures in Table 1. Lastly, we stress the need to collect and publish observations from more tropical, alpine, and polar regions, which are underrepresented in global datasets.

Code availability: Codes necessary to reproduce the analyses in this study are available at https://doi.org/10.5281/zenodo.16761320

Data availability: All data used in this study are available from their original sources. Streamflow data are available from CAMELS-AUS v2 (https://doi.org/10.5281/zenodo.14289037) (Fowler et al., 2024);, CAMELS-BR (https://doi.org/10.5281/zenodo.3709337) (Chagas et al., 2020), LamaH-CE (https://doi.org/10.5281/zenodo.4525244) (Klingler et al., 2021), CAMELS-CL (https://doi.pangaea.de/10.1594/PANGAEA.894885) (Alvarez-Garreton et al., 2018), CAMELS-DK (https://doi.org/10.22008/FK2/AZXSYP) (Liu et al., 2024), CAMELS-DE (https://doi.org/10.5281/zenodo.12733967) (Loritz et al., 2024), CAMELS-FR (https://doi.org/10.57745/WH7FJR) (Delaigue et al., 2024), CAMELS-GB (https://doi.org/10.5285/8344e4f3-d2ea-44f5-8afa-86d2987543a9) (Coxon et al., 2020), LamaH-ICE (https://doi.org/10.4211/hs.86117a5f36cc4b7c90a5d54e18161c91) (Helgason & Nijssen, 2024), CAMELS-IND (https://doi.org/10.5281/zenodo.14005378) (Mangukiya et al., 2025), CAMELS-LUX (https://doi.org/10.5281/zenodo.13846619) (Nijzink et al., 2024), CAMELS (https://doi.org/10.5065/D6G73C3Q) (Newman et al., 2015), HYSETS (https://doi.org/10.17605/OSF.IO/RPC3W) (Arsenault et al., 2020), CAMELS-CH (https://doi.org/10.5281/zenodo.7784632) (Höge et al., 2023), R-ArcticNET (https://www.r-arcticnet.sr.unh.edu/v4.0/index.html) (Lammers and Shiklomanov, 2000), the Global Runoff Data Centre (https://grdc.bafg.de/), and three Caravan community extensions not associated with peer-reviewed publications (https://doi.org/10.5281/zenodo.15181680, https://doi.org/10.5281/zenodo.13320514, and https://doi.org/10.5281/zenodo.15040948).

Model simulations are available from GLOB-LSTM1 and GloFAS (https://doi.org/10.5281/zenodo.8139379) (Nearing et al., 2024), GLOB-LSTM2 (https://doi.org/10.5281/zenodo.15272903) (Yang et al., 2025), BR-LSTM (https://github.com/sruzzante/NSE-and-Variance-Components), CH-LSTM and CH-PREVAH (Basil Kraft, personal communication) (Kraft et al., 2025), ENA-LSTM (https://doi.org/10.17605/OSF.IO/3S2PQ) (Arsenault et al., 2023), US-LSTM (https://doi.org/10.5281/zenodo.10139248) (Kratzert et al., 2024), US-δHBV2.0UH (https://doi.org/10.5281/zenodo.13774373) (Song et al., 2025), BR-MGB-SA (https://doi.org/10.5281/zenodo.15025488) (Chagas et al., 2020; Siqueira et al., 2018), CE-COSERO (https://doi.org/10.5281/zenodo.4525244) (Klingler et al., 2021), the US-NHM (https://doi.org/10.5066/P9PGZE0S) (Regan et al., 2019), WNA-VIC-Gl (https://data.pacificclimate.org/portal/hydro_stn_cmip5/map/) (Schnorbus, 2018), and US-FUSE, US-HBV, US-mHM,US-SAC-SMA, and US-VIC (https://doi.org/10.4211/hs.474ecc37e7db45baa425cdb4fc1b61e1) (Kratzert, 2019; Mizukami et al., 2019; Newman et al., 2017; Seibert et al., 2018).

Author Contributions: SWR designed the study, performed the analysis, and wrote most of the manuscript. WJMK ensured the results in Fig. 4 are reproducible. WJMK, TW, TG, and MS helped with the interpretation of the results and contributed to writing of the manuscript.

Acknowledgements: We thank Basil Kraft for sharing unpublished gauge-based model outputs from CH-RUN, and the authors of all open-access datasets used in this study.

Financial Support: SWR and TG acknowledge support from the Natural Sciences and Engineering Research Council of Canada. TW acknowledges support from the Alexander von Humboldt Foundation in the framework of the Alexander von Humboldt Professorship endowed by the German Federal Ministry of Education and Research (BMBF). WJMK was supported by the Cooperative Institute for Research to Operations in Hydrology (CIROH) with funding under award NA22NWS4320003 from the NOAA Cooperative Institute Program. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the opinions of NOAA.

Competing interests: The authors declare that they have no conflict of interest.

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.

Alvarez-Garreton, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., Lara, A., Puelma, C., Cortes, G., Garreaud, R., McPhee, J., and Ayala, A.: The CAMELS-CL dataset: catchment attributes and meteorology for large sample studies – Chile dataset, Hydrology and Earth System Sciences, 22, 5817–5846, https://doi.org/10.5194/hess-22-5817-2018, 2018.

Araya, K., Muñoz, P., Dezileau, L., Maldonado, A., Campos-Caba, R., Rebolledo, L., Cardenas, P., and Salamanca, M.: Extreme Sea Surges, Tsunamis and Pluvial Flooding Events during the Last ~1000 Years in the Semi-Arid Wetland, Coquimbo Chile, Geosciences, 12, 135, https://doi.org/10.3390/geosciences12030135, 2022.

Arsenault, R., Brissette, F., Martel, J.-L., Troin, M., Lévesque, G., Davidson-Chaput, J., Gonzalez, M. C., Ameli, A., and Poulin, A.: A comprehensive, multisource database for hydrometeorological modeling of 14,425 North American watersheds, Sci Data, 7, 243, https://doi.org/10.1038/s41597-020-00583-2, 2020.

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.

Bathelemy, R., Brigode, P., Andréassian, V., Perrin, C., Moron, V., Gaucherel, C., Tric, E., and Boisson, D.: Simbi: historical hydro-meteorological time series and signatures for 24 catchments in Haiti, Earth System Science Data, 16, 2073–2098, https://doi.org/10.5194/essd-16-2073-2024, 2024.

Beck, H. E., Zimmermann, N. E., McVicar, T. R., Vergopolan, N., Berg, A., and Wood, E. F.: Present and future Köppen-Geiger climate classification maps at 1-km resolution, Sci Data, 5, 180214, https://doi.org/10.1038/sdata.2018.214, 2018.

Berghuijs, W. R., Hale, K., and Beria, H.: Technical note: Streamflow seasonality using directional statistics, Hydrology and Earth System Sciences, 29, 2851–2862, https://doi.org/10.5194/hess-29-2851-2025, 2025.

Brown, B. C., Fullerton, A. H., Kopp, D., Tromboni, F., Shogren, A. J., Webb, J. A., Ruffing, C., Heaton, M., Kuglerová, L., Allen, D. C., McGill, L., Zarnetske, J. P., Whiles, M. R., Jones Jr., J. B., and Abbott, B. W.: The Music of Rivers: The Mathematics of Waves Reveals Global Structure and Drivers of Streamflow Regime, Water Resources Research, 59, e2023WR034484, https://doi.org/10.1029/2023WR034484, 2023.

Burton, C., Rifai, S., and Malhi, Y.: Inter-comparison and assessment of gridded climate products over tropical forests during the 2015/2016 El Niño, Philosophical Transactions of the Royal Society B: Biological Sciences, 373, 20170406, https://doi.org/10.1098/rstb.2017.0406, 2018.

Casado Rodríguez, J.: CAMELS-ES: Catchment Attributes and Meteorology for Large-Sample Studies – Spain (1.0.2), https://doi.org/10.5281/zenodo.8428374, 2023.

Chagas, V. B. P., Chaffe, P. L. B., Addor, N., Fan, F. M., Fleischmann, A. S., Paiva, R. C. D., and Siqueira, V. A.: CAMELS-BR: hydrometeorological time series and landscape attributes for 897 catchments in Brazil, Earth System Science Data, 12, 2075–2096, https://doi.org/10.5194/essd-12-2075-2020, 2020.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I.: STL: A Seasonal-Trend Decomposition Procedure Based on Loess, Journal of Official Statistics, 6, 3–73, 1990.

Court, A.: Measures of streamflow timing, Journal of Geophysical Research (1896-1977), 67, 4335–4339, https://doi.org/10.1029/JZ067i011p04335, 1962.

Coxon, G., Addor, N., Bloomfield, J. P., Freer, J., Fry, M., Hannaford, J., Howden, N. J. K., Lane, R., Lewis, M., Robinson, E. L., Wagener, T., and Woods, R.: CAMELS-GB: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain, Earth System Science Data, 12, 2459–2483, https://doi.org/10.5194/essd-12-2459-2020, 2020.

Delaigue, O., Guimarães, G. M., Brigode, P., Génot, B., Perrin, C., Soubeyroux, J.-M., Janet, B., Addor, N., and Andréassian, V.: CAMELS-FR dataset: A large-sample hydroclimatic dataset for France to explore hydrological diversity and support model benchmarking, Earth System Science Data Discussions, 1–27, https://doi.org/10.5194/essd-2024-415, 2024.

D'Odorico, P. and Bhattachan, A.: Hydrologic variability in dryland regions: impacts on ecosystem dynamics and food security, Philosophical Transactions of the Royal Society B: Biological Sciences, 367, 3145–3157, https://doi.org/10.1098/rstb.2012.0016, 2012.

D'Odorico, P., Caylor, K., Okin, G. S., and Scanlon, T. M.: On soil moisture–vegetation feedbacks and their possible effects on the dynamics of dryland ecosystems, Journal of Geophysical Research: Biogeosciences, 112, https://doi.org/10.1029/2006JG000379, 2007.

Döll, P. and Schmied, H. M.: How is the impact of climate change on river flow regimes related to the impact on mean annual runoff? A global-scale analysis, Environ. Res. Lett., 7, 014037, https://doi.org/10.1088/1748-9326/7/1/014037, 2012.

Dralle, D., Karst, N., Müller, M., Vico, G., and Thompson, S. E.: Stochastic modeling of interannual variation of hydrologic variables, Geophysical Research Letters, 44, 7285–7294, https://doi.org/10.1002/2017GL074139, 2017.

Efrat, M.: Caravan extension Israel - Israel dataset for large-sample hydrology, https://doi.org/10.5281/zenodo.15003600, 2025.

Eker, S., Rovenskaya, E., Obersteiner, M., and Langan, S.: Practice and perspectives in the validation of resource management models, Nat Commun, 9, 5359, https://doi.org/10.1038/s41467-018-07811-9, 2018.

Falcone, J. A.: GAGES-II: Geospatial Attributes of Gages for Evaluating Streamflow, U.S. Geological Survey, https://doi.org/10.3133/70046617, 2011.

Färber, C., Plessow, H., Mischel, S., Kratzert, F., Addor, N., Shalev, G., and Looser, U.: GRDC-Caravan: extending the original dataset with data from the Global Runoff Data Centre (0.5), https://doi.org/10.5281/zenodo.15124865, 2025.

Flores, B. M., Montoya, E., Sakschewski, B., Nascimento, N., Staal, A., Betts, R. A., Levis, C., Lapola, D. M., Esquível-Muelbert, A., Jakovac, C., Nobre, C. A., Oliveira, R. S., Borma, L. S., Nian, D., Boers, N., Hecht, S. B., ter Steege, H., Arieira, J., Lucas, I. L., Berenguer, E., Marengo, J. A., Gatti, L. V., Mattos, C. R. C., and Hirota, M.: Critical transitions in the Amazon forest system, Nature, 626, 555–564, https://doi.org/10.1038/s41586-023-06970-0, 2024.

Fowler, K. J. A., Zhang, Z., and Hou, X.: CAMELS-AUS v2: updated hydrometeorological timeseries and landscape attributes for an enlarged set of catchments in Australia, Earth System Science Data Discussions, 1–21, https://doi.org/10.5194/essd-2024-263, 2024.

Garrick, M., Cunnane, C., and Nash, J. E.: A criterion of efficiency for rainfall-runoff models, Journal of Hydrology, 36, 375–381, https://doi.org/10.1016/0022-1694(78)90155-5, 1978.

Gauch, M., Kratzert, F., Gilon, O., Gupta, H., Mai, J., Nearing, G., Tolson, B., Hochreiter, S., and Klotz, D.: In Defense of Metrics: Metrics Sufficiently Encode Typical Human Preferences Regarding Hydrological Model Performance, Water Resources Research, 59, e2022WR033918, https://doi.org/10.1029/2022WR033918, 2023.

Ghiggi, G., Humphrey, V., Seneviratne, S. I., and Gudmundsson, L.: GRUN: an observation-based global gridded runoff dataset from 1902 to 2014, Earth System Science Data, 11, 1655–1674, https://doi.org/10.5194/essd-11-1655-2019, 2019.

Gudmundsson, L., Tallaksen, L. M., Stahl, K., Clark, D. B., Dumont, E., Hagemann, S., Bertrand, N., Gerten, D., Heinke, J., Hanasaki, N., Voss, F., and Koirala, S.: Comparing Large-Scale Hydrological Model Simulations to Observed Runoff Percentiles in Europe, https://doi.org/10.1175/JHM-D-11-083.1, 2012.

Gudmundsson, L., Boulange, J., Do, H. X., Gosling, S. N., Grillakis, M. G., Koutroulis, A. G., Leonard, M., Liu, J., Müller Schmied, H., Papadimitriou, L., Pokhrel, Y., Seneviratne, S. I., Satoh, Y., Thiery, W., Westra, S., Zhang, X., and Zhao, F.: Globally observed trends in mean and extreme river flow attributed to climate change, Science, 371, 1159–1162, https://doi.org/10.1126/science.aba3996, 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.

Hall, J. W., Grey, D., Garrick, D., Fung, F., Brown, C., Dadson, S. J., and Sadoff, C. W.: Coping with the curse of freshwater variability, Science, https://doi.org/10.1126/science.1257890, 2014.

Han, J., Liu, Z., Woods, R., McVicar, T. R., Yang, D., Wang, T., Hou, Y., Guo, Y., Li, C., and Yang, Y.: Streamflow seasonality in a snow-dwindling world, Nature, 629, 1075–1081, https://doi.org/10.1038/s41586-024-07299-y, 2024.

Helgason, H. B. and Nijssen, B.: LamaH-Ice: LArge-SaMple DAta for Hydrology and Environmental Sciences for Iceland, Earth System Science Data, 16, 2741–2771, https://doi.org/10.5194/essd-16-2741-2024, 2024.

Henck, A. C., Huntington, K. W., Stone, J. O., Montgomery, D. R., and Hallet, B.: Spatial controls on erosion in the Three Rivers Region, southeastern Tibet and southwestern China, Earth and Planetary Science Letters, 303, 71–83, https://doi.org/10.1016/j.epsl.2010.12.038, 2011.

Höge, M., Kauzlaric, M., Siber, R., Schönenberger, U., Horton, P., Schwanbeck, J., Floriancic, M. G., Viviroli, D., Wilhelm, S., Sikorska-Senoner, A. E., Addor, N., Brunner, M., Pool, S., Zappa, M., and Fenicia, F.: CAMELS-CH: hydro-meteorological time series and landscape attributes for 331 catchments in hydrologic Switzerland, Earth System Science Data, 15, 5755–5784, https://doi.org/10.5194/essd-15-5755-2023, 2023.

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., Arheimer, B., Blume, T., Clark, M. P., Ehret, U., Fenicia, F., Freer, J. E., Gelfan, A., Gupta, H. V., Hughes, D. A., Hut, R. W., Montanari, A., Pande, S., Tetzlaff, D., Troch, P. A., Uhlenbrook, S., Wagener, T., Winsemius, H. C., Woods, R. A., Zehe, E., and Cudennec, C.: A decade of Predictions in Ungauged Basins (PUB)—a review, Hydrological Sciences Journal, 58, 1198–1255, https://doi.org/10.1080/02626667.2013.803183, 2013.

Kendall, M. and Stuart, A.: Time Series: Trend and Seasonality, in: The Advanced Theory of Statistics, vol. 3, Griffin, London, 366–402, 1966.

Klemeš, V.: Operational testing of hydrological simulation models, Hydrological Sciences Journal, 31, 13–24, https://doi.org/10.1080/02626668609491024, 1986.

Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, Journal of Hydrology, 424–425, 264–277, https://doi.org/10.1016/j.jhydrol.2012.01.011, 2012.

Klingler, C., Schulz, K., and Herrnegger, M.: LamaH-CE: LArge-SaMple DAta for Hydrology and Environmental Sciences for Central Europe, Earth System Science Data, 13, 4529–4565, https://doi.org/10.5194/essd-13-4529-2021, 2021.

Knoben, W. J. M.: Setting expectations for hydrologic model performance with an ensemble of simple benchmarks, Hydrological Processes, 38, e15288, https://doi.org/10.1002/hyp.15288, 2024.

Knoben, W. J. M., Woods, R. A., and Freer, J. E.: A Quantitative Hydrological Climate Classification Evaluated With Independent Streamflow Data, Water Resources Research, 54, 5088–5109, https://doi.org/10.1029/2018WR022913, 2018.

Köplin, N., Schädler, B., Viviroli, D., and Weingartner, R.: The importance of glacier and forest change in hydrological climate-impact studies, Hydrology and Earth System Sciences, 17, 619–635, https://doi.org/10.5194/hess-17-619-2013, 2013.

Krabbenhoft, C. A., Allen, G. H., Lin, P., Godsey, S. E., Allen, D. C., Burrows, R. M., DelVecchia, A. G., Fritz, K. M., Shanafield, M., Burgin, A. J., Zimmer, M. A., Datry, T., Dodds, W. K., Jones, C. N., Mims, M. C., Franklin, C., Hammond, J. C., Zipper, S., Ward, A. S., Costigan, K. H., Beck, H. E., and Olden, J. D.: Assessing placement bias of the global river gauge network, Nat Sustain, 5, 586–592, https://doi.org/10.1038/s41893-022-00873-0, 2022.

Kraft, B., Schirmer, M., Aeberhard, W. H., Zappa, M., Seneviratne, S. I., and Gudmundsson, L.: CH-RUN: a deep-learning-based spatially contiguous runoff reconstruction for Switzerland, Hydrology and Earth System Sciences, 29, 1061–1082, https://doi.org/10.5194/hess-29-1061-2025, 2025.

Kratzert, F.: CAMELS benchmark models, https://doi.org/10.4211/hs.474ecc37e7db45baa425cdb4fc1b61e1, 2019.

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., and Matias, Y.: Caravan - A global community dataset for large-sample hydrology, Sci Data, 10, 61, https://doi.org/10.1038/s41597-023-01975-w, 2023.

Kratzert, F., Gauch, M., Klotz, D., and Nearing, G.: HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin, Hydrology and Earth System Sciences, 28, 4187–4201, https://doi.org/10.5194/hess-28-4187-2024, 2024.

Krysanova, V., Donnelly, C., Gelfan, A., Gerten, D., Arheimer, B., Hattermann, F., and Kundzewicz, Z. W.: How the performance of hydrological models relates to credibility of projections under climate change, Hydrological Sciences Journal, 2018.

Ladson, T., Brown, R., Neal, B., and Nathan, R.: A standard approach to baseflow separation using the Lyne and Hollick filter, Australian Journal of Water Resources, 17, https://doi.org/10.7158/W12-028.2013.17.1, 2013.

Lammers, R. B. and Shiklomanov, A. I.: R-ArcticNet, A Regional Hydrographic Data Network for the Pan-Arctic Region., 2000.

Lehner, B. and Grill, G.: Global river hydrography and network routing: baseline data and new approaches to study the world’s large river systems, Hydrological Processes, 27, 2171–2186, https://doi.org/10.1002/hyp.9740, 2013.

Lin, J., Bryan, B. A., Zhou, X., Lin, P., Do, H. X., Gao, L., Gu, X., Liu, Z., Wan, L., Tong, S., Huang, J., Wang, Q., Zhang, Y., Gao, H., Yin, J., Chen, Z., Duan, W., Xie, Z., Cui, T., Liu, J., Li, M., Li, X., Xu, Z., Guo, F., Shu, L., Li, B., Zhang, J., Zhang, P., Fan, B., Wang, Y., Zhang, Y., Huang, J., Li, X., Cai, Y., and Yang, Z.: Making China’s water data accessible, usable and shareable, Nat Water, 1, 328–335, https://doi.org/10.1038/s44221-023-00039-y, 2023.

Liu, J., Koch, J., Stisen, S., Troldborg, L., Højberg, A. L., Thodsen, H., Hansen, M. F. T., and Schneider, R. J. M.: CAMELS-DK: Hydrometeorological Time Series and Landscape Attributes for 3330 Catchments in Denmark, Earth System Science Data Discussions, 1–30, https://doi.org/10.5194/essd-2024-292, 2024.

Loritz, R., Dolich, A., Acuña Espinoza, E., Ebeling, P., Guse, B., Götte, J., Hassler, S. K., Hauffe, C., Heidbüchel, I., Kiesel, J., Mälicke, M., Müller-Thomy, H., Stölzle, M., and Tarasova, L.: CAMELS-DE: hydro-meteorological time series and attributes for 1582 catchments in Germany, Earth System Science Data, 16, 5625–5642, https://doi.org/10.5194/essd-16-5625-2024, 2024.

Lyne, V. and Hollick, M.: Stochastic time-variable rainfall-runoff modelling, Institute of engineers Australia national conference, 1979.

Mangukiya, N. K., Kumar, K. B., Dey, P., Sharma, S., Bejagam, V., Mujumdar, P. P., and Sharma, A.: CAMELS-IND: hydrometeorological time series and catchment attributes for 228 catchments in Peninsular India, Earth System Science Data, 17, 461–491, https://doi.org/10.5194/essd-17-461-2025, 2025.

Mann, H. B. and Whitney, D. R.: On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other, The Annals of Mathematical Statistics, 18, 50–60, https://doi.org/10.1214/aoms/1177730491, 1947.

Marsh, T. J. and Dale, M.: The UK Floods of 2000–2001: A Hydrometeorological Appraisal, Water and Environment Journal, 16, 180–188, https://doi.org/10.1111/j.1747-6593.2002.tb00392.x, 2002.

Martinec, J. and Rango, A.: Merits of Statistical Criteria for the Performance of Hydrological Models, JAWRA Journal of the American Water Resources Association, 25, 421–432, https://doi.org/10.1111/j.1752-1688.1989.tb03079.x, 1989.

McMillan, H. K.: A review of hydrologic signatures and their applications, WIREs Water, 8, e1499, https://doi.org/10.1002/wat2.1499, 2021.

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

Mizukami, N., Rakovec, O., Newman, A. J., Clark, M. P., Wood, A. W., Gupta, H. V., and Kumar, R.: On the choice of calibration metrics for “high-flow” estimation using hydrologic models, Hydrology and Earth System Sciences, 23, 2601–2614, https://doi.org/10.5194/hess-23-2601-2019, 2019.

Montanari, A., Young ,G., Savenije ,H.H.G., Hughes ,D., Wagener ,T., Ren ,L.L., Koutsoyiannis ,D., Cudennec ,C., Toth ,E., Grimaldi ,S., Blöschl ,G., Sivapalan ,M., Beven ,K., Gupta ,H., Hipsey ,M., Schaefli ,B., Arheimer ,B., Boegh ,E., Schymanski ,S.J., Di Baldassarre ,G., Yu ,B., Hubert ,P., Huang ,Y., Schumann ,A., Post ,D.A., Srinivasan ,V., Harman ,C., Thompson ,S., Rogger ,M., Viglione ,A., McMillan ,H., Characklis ,G., Pang ,Z., and and Belyaev, V.: “Panta Rhei—Everything Flows”: Change in hydrology and society—The IAHS Scientific Decade 2013–2022, Hydrological Sciences Journal, 58, 1256–1275, https://doi.org/10.1080/02626667.2013.809088, 2013.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I — A discussion of principles, Journal of Hydrology, 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C., Shalev, G., Shenzis, S., Tekalign, T. Y., Weitzner, D., and Matias, Y.: Global prediction of extreme floods in ungauged watersheds, Nature, 627, 559–563, https://doi.org/10.1038/s41586-024-07145-1, 2024.

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 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, https://doi.org/10.1175/JHM-D-16-0284.1, 2017.

Nijzink, J., Loritz, R., Gourdol, L., Zoccatelli, D., Iffly, J. F., and Pfister, L.: CAMELS-LUX: Highly Resolved Hydro-Meteorological and Atmospheric Data for Physiographically Characterized Catchments around Luxembourg, https://doi.org/10.5281/zenodo.13846620, 2024.

Pellerin, J. and Nzokou Tanekou, F.: Reference Hydrometric Basin Network Update, Environment and Climate Change Canada, Gatineau, QC, 2020.

Pepin, N. C., Arnone, E., Gobiet, A., Haslinger, K., Kotlarski, S., Notarnicola, C., Palazzi, E., Seibert, P., Serafin, S., Schöner, W., Terzago, S., Thornton, J. M., Vuille, M., and Adler, C.: Climate Changes and Their Elevational Patterns in the Mountains of the World, Reviews of Geophysics, 60, e2020RG000730, https://doi.org/10.1029/2020RG000730, 2022.

Poff, N. L. and Zimmerman, J. K. H.: Ecological responses to altered flow regimes: a literature review to inform the science and management of environmental flows, Freshwater Biology, 55, 194–205, https://doi.org/10.1111/j.1365-2427.2009.02272.x, 2010.

Poff, N. L., Allan, J. D., Bain, M. B., Karr, J. R., Prestegaard, K. L., Richter, B. D., Sparks, R. E., and Stromberg, J. C.: The Natural Flow Regime, BioScience, 47, 769–784, https://doi.org/10.2307/1313099, 1997.

Rantanen, M., Karpechko, A. Y., Lipponen, A., Nordling, K., Hyvärinen, O., Ruosteenoja, K., Vihma, T., and Laaksonen, A.: The Arctic has warmed nearly four times faster than the globe since 1979, Commun Earth Environ, 3, 1–10, https://doi.org/10.1038/s43247-022-00498-3, 2022.

Refsgaard, J. C., Madsen, H., Andréassian, V., Arnbjerg-Nielsen, K., Davidson, T. A., Drews, M., Hamilton, D. P., Jeppesen, E., Kjellström, E., Olesen, J. E., Sonnenborg, T. O., Trolle, D., Willems, P., and Christensen, J. H.: A framework for testing the ability of models to project climate change and its impacts, Climatic Change, 122, 271–282, https://doi.org/10.1007/s10584-013-0990-2, 2014.

Regan, R. S., Juracek, K. E., Hay, L. E., Markstrom, S. L., Viger, R. J., Driscoll, J. M., LaFontaine, J. H., and Norton, P. A.: The U. S. Geological Survey National Hydrologic Model infrastructure: Rationale, description, and application of a watershed-scale model for the conterminous United States, Environmental Modelling & Software, 111, 192–203, https://doi.org/10.1016/j.envsoft.2018.09.023, 2019.

Richter, B. D., Baumgartner, J. V., Powell, J., and Braun, D. P.: A Method for Assessing Hydrologic Alteration within Ecosystems, Conservation Biology, 10, 1163–1174, https://doi.org/10.1046/j.1523-1739.1996.10041163.x, 1996.

Rodwell, M. J., Rowell, D. P., and Folland, C. K.: Oceanic forcing of the wintertime North Atlantic Oscillation and European climate, Nature, 398, 320–323, https://doi.org/10.1038/18648, 1999.

Ruzzante, S. W. and Gleeson, T.: Rising Temperatures Drive Lower Summer Minimum Flows Across Hydrologically Diverse Catchments in British Columbia, Water Resources Research, 61, e2024WR038057, https://doi.org/10.1029/2024WR038057, 2025.

Safeeq, M., Grant, G. E., Lewis, S. L., Kramer, M. G., and Staab, B.: A hydrogeologic framework for characterizing summer streamflow sensitivity to climate warming in the Pacific Northwest, USA, Hydrology and Earth System Sciences, 18, 3693–3710, https://doi.org/10.5194/hess-18-3693-2014, 2014.

Santos, M. S. and Slater, L. J.: Integrating Hidden Markov and Multinomial models for hydrological drought prediction under nonstationarity, Advances in Water Resources, 200, 104974, https://doi.org/10.1016/j.advwatres.2025.104974, 2025.

Schaefli, B. and Gupta, H. V.: Do Nash values have value?, Hydrological Processes, 21, 2075–2080, https://doi.org/10.1002/hyp.6825, 2007.

Schnorbus, M.: VIC Glacier (VIC-GL) - Description of VIC model changes and upgrades, VIC Generation 2 Deployment Report volume 1, Pacific Climate Impacts Consortium, University of Victoria, Victoria, BC, 2018.

Schnorbus, M.: VIC-Glacier (VIC-GL): Model set-up and deployment for the Peace, Fraser, and Columbia: VIC generation 2 deployment report, volume 6, Pacific Climate Impacts Consortium (PCIC), 2020.

Seibert, J.: On the need for benchmarks in hydrological modelling, Hydrological Processes, 15, 1063–1064, https://doi.org/10.1002/hyp.446, 2001.

Seibert, J.: Reliability of Model Predictions Outside Calibration Conditions: Paper presented at the Nordic Hydrological Conference (Røros, Norway 4-7 August 2002), Hydrology Research, 34, 477–492, https://doi.org/10.2166/nh.2003.0019, 2003.

Seibert, J., Vis, M. J. P., Lewis, E., and van Meerveld, H. j.: Upper and lower benchmarks in hydrological modelling, Hydrological Processes, 32, 1120–1125, https://doi.org/10.1002/hyp.11476, 2018.

Shamir, E., Imam, B., Morin, E., Gupta, H. V., and Sorooshian, S.: The role of hydrograph indices in parameter estimation of rainfall–runoff models, Hydrological Processes, 19, 2187–2207, https://doi.org/10.1002/hyp.5676, 2005.

Simeone, C., McCabe ,Greg, Hecht ,Jory, Hammond ,John, Hodgkins ,Glenn, Olson ,Carolyn, Wieczorek ,Mike, and and Wolock, D.: Low-flow period seasonality, trends, and climate linkages across the United States, Hydrological Sciences Journal, 69, 1387–1398, https://doi.org/10.1080/02626667.2024.2369639, 2024.

Siqueira, V. A., Paiva, R. C. D., Fleischmann, A. S., Fan, F. M., Ruhoff, A. L., Pontes, P. R. M., Paris, A., Calmant, S., and Collischonn, W.: Toward continental hydrologic–hydrodynamic modeling in South America, Hydrology and Earth System Sciences, 22, 4815–4842, https://doi.org/10.5194/hess-22-4815-2018, 2018.

Slater, L. J., Anderson, B., Buechel, M., Dadson, S., Han, S., Harrigan, S., Kelder, T., Kowal, K., Lees, T., Matthews, T., Murphy, C., and Wilby, R. L.: Nonstationary weather and water extremes: a review of methods for their detection, attribution, and management, Hydrology and Earth System Sciences, 25, 3897–3935, https://doi.org/10.5194/hess-25-3897-2021, 2021.

Smith, A., Sampson, C., and Bates, P.: Regional flood frequency analysis at the global scale, Water Resources Research, 51, 539–553, https://doi.org/10.1002/2014WR015814, 2015.

Smith, L. C., Turcotte, D. L., and Isacks, B. L.: Stream flow characterization and feature detection using a discrete wavelet transform, Hydrological Processes, 12, 233–249, https://doi.org/10.1002/(SICI)1099-1085(199802)12:2%3C233::AID-HYP573%3E3.0.CO;2-3, 1998.

Song, Y., Bindas, T., Shen, C., Ji, H., Knoben, W. J. M., Lonzarich, L., Clark, M. P., Liu, J., van Werkhoven, K., Lamont, S., Denno, M., Pan, M., Yang, Y., Rapp, J., Kumar, M., Rahmani, F., Thébault, C., Adkins, R., Halgren, J., Patel, T., Patel, A., Sawadekar, K. A., and Lawson, K.: High-Resolution National-Scale Water Modeling Is Enhanced by Multiscale Differentiable Physics-Informed Machine Learning, Water Resources Research, 61, e2024WR038928, https://doi.org/10.1029/2024WR038928, 2025.

St. Jacques, J.-M., Huang ,Yuhui Althea, Zhao ,Yang, Lapp ,Suzan L., and and Sauchyn, D. J.: Detection and attribution of variability and trends in streamflow records from the Canadian Prairie Provinces, Canadian Water Resources Journal / Revue canadienne des ressources hydriques, 39, 270–284, https://doi.org/10.1080/07011784.2014.942575, 2014.

Staal, A., Fetzer, I., Wang-Erlandsson, L., Bosmans, J. H. C., Dekker, S. C., van Nes, E. H., Rockström, J., and Tuinenburg, O. A.: Hysteresis of tropical forests in the 21st century, Nat Commun, 11, 4978, https://doi.org/10.1038/s41467-020-18728-7, 2020.

Stein, L., Mukkavilli, S. K., Pfitzmann, B. M., Staar, P. W. J., Ozturk, U., Berrospi, C., Brunschwiler, T., and Wagener, T.: Wealth Over Woe: Global Biases in Hydro-Hazard Research, Earth’s Future, 12, e2024EF004590, https://doi.org/10.1029/2024EF004590, 2024.

Stewart, I. T., Cayan, D. R., and Dettinger, M. D.: Changes toward Earlier Streamflow Timing across Western North America, https://doi.org/10.1175/JCLI3321.1, 2005.

Taye, M. T. and Dyer, E.: Hydrologic Extremes in a Changing Climate: a Review of Extremes in East Africa, Curr Clim Change Rep, 10, 1–11, https://doi.org/10.1007/s40641-024-00193-9, 2024.

Turner, S., Hannaford, J., Barker, L. J., Suman, G., Killeen, A., Armitage, R., Chan, W., Davies, H., Griffin, A., Kumar, A., Dixon, H., Albuquerque, M. T. D., Almeida Ribeiro, N., Alvarez-Garreton, C., Amoussou, E., Arheimer, B., Asano, Y., Berezowski, T., Bodian, A., Boutaghane, H., Capell, R., Dakhaoui, H., Daňhelka, J., Do, H. X., Ekkawatpanit, C., El Khalki, E. M., Fleig, A. K., Fonseca, R., Giraldo-Osorio, J. D., Goula, A. B. T., Hanel, M., Horton, S., Kan, C., Kingston, D. G., Laaha, G., Laugesen, R., Lopes, W., Mager, S., Rachdane, M., Markonis, Y., Medeiro, L., Midgley, G., Murphy, C., O'Connor, P., Pedersen, A. I., Pham, H. T., Piniewski, M., Renard, B., Saidi, M. E., Schmocker-Fackel, P., Stahl, K., Thyer, M., Toucher, M., Tramblay, Y., Uusikivi, J., Venegas-Cordero, N., Visessri, S., Watson, A., Westra, S., and Whitfield, P. H.: ROBIN: Reference observatory of basins for international hydrological climate change detection, Sci Data, 12, 654, https://doi.org/10.1038/s41597-025-04907-y, 2025.

Wagener, T., Sivapalan, M., Troch, P. A., McGlynn, B. L., Harman, C. J., Gupta, H. V., Kumar, P., Rao, P. S. C., Basu, N. B., and Wilson, J. S.: The future of hydrology: An evolving science for a changing world, Water Resources Research, 46, https://doi.org/10.1029/2009WR008906, 2010.

Wasko, C., Nathan, R., and Peel, M. C.: Trends in Global Flood and Streamflow Timing Based on Local Water Year, Water Resources Research, 56, e2020WR027233, https://doi.org/10.1029/2020WR027233, 2020.

West, H., Quinn ,Nevil, and and Horswell, M.: Spatio-temporal propagation of North Atlantic Oscillation (NAO) rainfall deviations to streamflow in British catchments, Hydrological Sciences Journal, 67, 676–688, https://doi.org/10.1080/02626667.2022.2038791, 2022.

Wilks, D. S.: Statistical methods in the atmospheric sciences, 2nd ed., Academic Press, Amsterdam ; Boston, 627 pp., 2006.

Xiong, J. and Yang, Y.: Climate Change and Hydrological Extremes, Curr Clim Change Rep, 11, 1, https://doi.org/10.1007/s40641-024-00198-4, 2024.

Yadav, M., Wagener, T., and Gupta, H.: Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, Advances in Water Resources, 30, 1756–1774, https://doi.org/10.1016/j.advwatres.2007.01.005, 2007.

Yang, Y., Feng, D., Beck, H. E., Hu, W., Abbas, A., Sengupta, A., Delle Monache, L., Hartman, R., Lin, P., Shen, C., and Pan, M.: Global Daily Discharge Estimation Based on Grid Long Short-Term Memory (LSTM) Model and River Routing, Water Resources Research, 61, e2024WR039764, https://doi.org/10.1029/2024WR039764, 2025.

Yin, Z., Lin, P., Riggs, R., Allen, G. H., Lei, X., Zheng, Z., and Cai, S.: A synthesis of Global Streamflow Characteristics, Hydrometeorology, and Catchment Attributes (GSHA) for large sample river-centric studies, Earth System Science Data, 16, 1559–1587, https://doi.org/10.5194/essd-16-1559-2024, 2024.

Zou, S., Jilili, A., Duan, W., Maeyer, P. D., and de Voorde, T. V.: Human and Natural Impacts on the Water Resources in the Syr Darya River Basin, Central Asia, Sustainability, 11, 3084, https://doi.org/10.3390/su11113084, 2019.