Contents

S1: Benchmark KGEs

For the climatological benchmark model, the NSE and KGE are closely related. Figure S1 shows the benchmark KGEs for the 20,338 catchments.

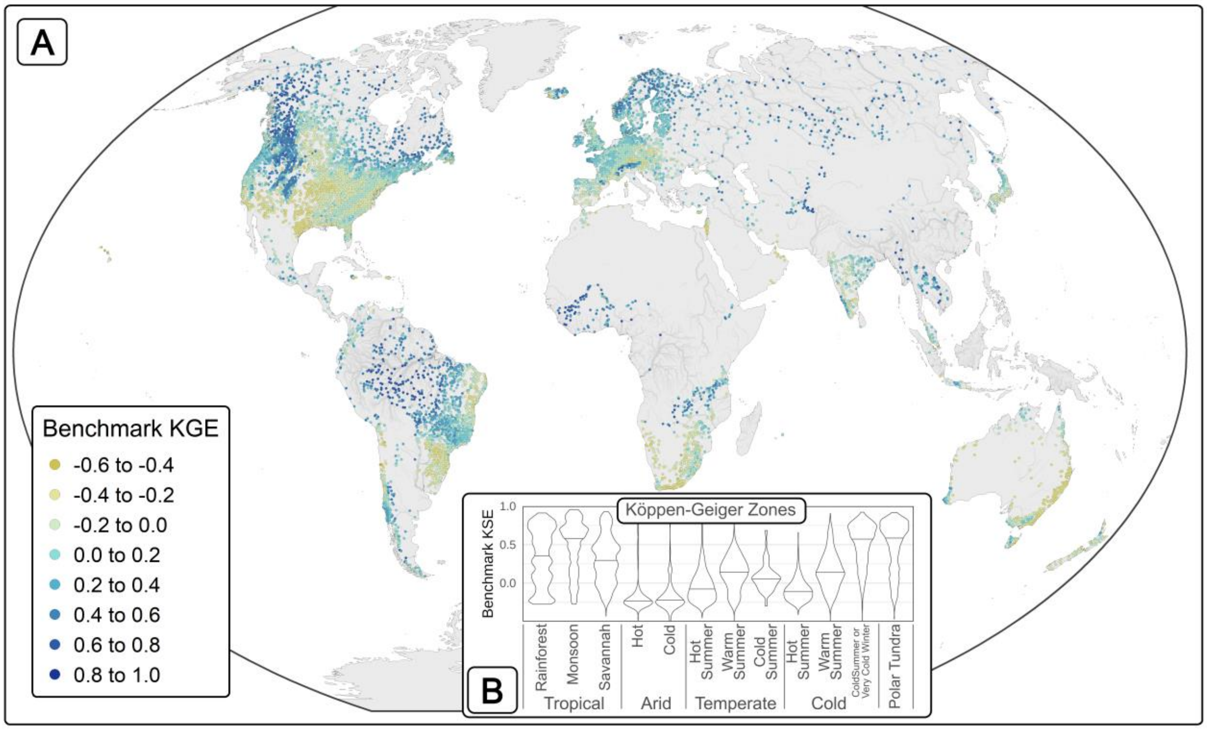

Figure S1: A: Climatological Benchmark KGEs for 20,338 catchments. B: Benchmark KGEs by Koppen-Geiger climate classification.

For a climatological benchmark model, where the model is calculated and tested on the same data, the KGE is uniquely and monotonically defined by the NSE. The NSE can be written as:

$$NSE = 2r\alpha - \alpha^2 - \beta^2 \qquad [S1]$$

Where *r* is the Pearson correlation coefficient, α is the ratio of standard deviations σ in the simulated (s) and observed (o) time series, and β is the bias.

$$\alpha = \frac{\sigma_s}{\sigma_o} \qquad [S2]$$

$$\beta_n = \frac{\mu_s - \mu_o}{\sigma_o} \qquad [S3]$$

Since the climatological model is generated by averaging the observed time series, $\beta_n = 0$ for the benchmark NSE$_{cb}$.

$$NSE_{cb} = 2r\alpha - \alpha^2 \qquad [S4]$$

The KGE is defined as:

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \qquad [S5]$$

Where α is as defined above and:

$$\beta = \frac{\mu_s}{\mu_o} \quad \text{[S6]}$$

Again the bias term $\beta - 1 = 0$. We can therefore expand and rearrange the KGE$_{cb}$:

$$KGE_{cb} = 1 - \sqrt{r^2 + \alpha^2 - 2r - 2\alpha + 2} \quad \text{[S7]}$$

Substituting for α^2 we get KGE as a function of NSE:

$$KGE_{cb} = 1 - \sqrt{r^2 - 2r + 2r\alpha - NSE - 2\alpha + 2} \quad \text{[S8]}$$

Now *r* is defined as:

$$r = \frac{cov(o,s)}{\sigma_s \sigma_o} \quad \text{[S9]}$$

Where cov(o,s) is the covariance of observed and simulated (climatological) time series. Substituting α:

$$r = \frac{cov(o,s)}{\sigma_s^2}\alpha \quad \text{[S10]}$$

For a climatological model, the covariance of the observed and simulated time series is equal to the variance in simulated time series. This can be shown as follows[1] (equations S11 to S20):

$$cov(o,s) = \frac{1}{n}\sum_{t=1}^{n}(s_t - \bar{s})(o_t - \bar{o}) \quad \text{[S11]}$$

Now consider that the observed time series is the climatological model s_t plus some zero-mean noise ϵ_t:

$$o_t = s_t + \epsilon_t \quad \text{[S12]}$$

And:

$$\bar{s} = \bar{o} \quad \text{[S13]}$$

$$cov(o,s) = \frac{1}{n}\sum_{t=1}^{n}(s_t - \bar{s})(s_t + \epsilon_t - \bar{s}) \quad \text{[S14]}$$

Now expanding and simplifying:

$$cov(o,s) = \frac{1}{n}\sum_{t=1}^{n}\left((s_t - \bar{s})^2 + \epsilon_t(s_t - \bar{s})\right) \quad \text{[S15]}$$

$$cov(o,s) = \sigma_s^2 + \frac{1}{n}\sum_{t=1}^{n}\epsilon_t(s_t - \bar{s}) \quad \text{[S16]}$$

Assuming a 365-day year and no missing data for any year, this summation can be written over d=365 days and y=Y years.

[1] We used ChatGPT to assist with this derivation: https://chatgpt.com/share/6839f23b-62fc-800c-b18a-48aa91df2b80

$$cov(o,s) = \sigma_s^2 + \frac{1}{365 * Y} \sum_{d=1}^{365} \sum_{y=1}^{Y} \epsilon_{d,y}(s_d - \bar{s}) \qquad [S17]$$

Since $(s_d - \bar{s})$ depends only on d we can take it out of the right-most summation:

$$cov(o,s) = \sigma_s^2 + \frac{1}{365 * Y} \sum_{d=1}^{365} \left((s_d - \bar{s}) \sum_{y=1}^{Y} \epsilon_{d,y} \right) \qquad [S18]$$

By construction the noise is zero-mean, $\sum_{y=1}^{Y} \epsilon_{d,y} = 0$, so:

$$cov(o,s) = \sigma_s^2 \qquad [S19]$$

Therefore equation S10 simplifies to:

$$r = \alpha \qquad [S20]$$

And equation S4 simplifies to:

$$NSE_{cb} = r^2 = \alpha^2 \qquad [S21]$$

Equation S8 can then be simplified:

$$KGE_{cb} = 1 - \sqrt{2} + \sqrt{2 \times NSE_{cb}} \qquad [S22]$$

Using leave-one-out cross-validation

In our analysis we used leave-one-out cross-validation to construct the climatological time series, which means that in equation S17 above we must replace $(s_d - \bar{s})$ with $\left(s_{d,y} - \bar{s_y}\right)$ since the climatology changes with each analysed year. The second term in equation S17 is then always less than or equal to zero, since the noise correlates negatively with the climatology.

In Figure S2 we plot the KGE$_{cb}$ against the NSE$_{cb}$ for the 20,338 catchments analyzed. For long time series The KGE approaches the ideal line (equation S22).

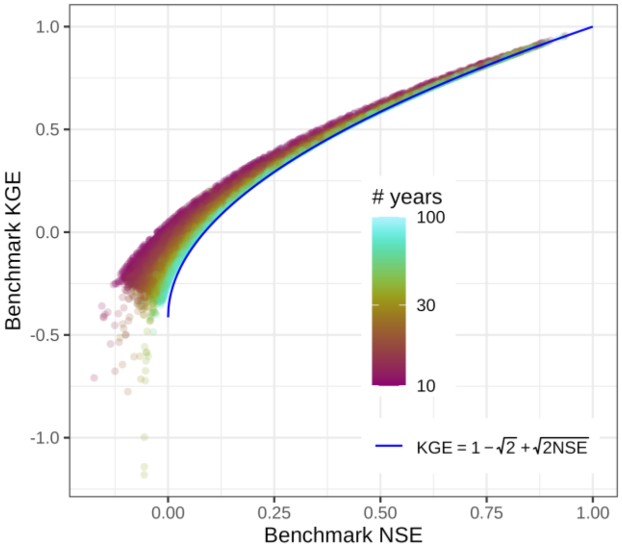

Figure S2: Scatterplot of benchmark KGE and NSE values. The idealized relationship is shown in blue, which is derived for a climatological model that is calculated and tested on the same data without cross-validation. For long time series the points plot near to the idealized line.

S2: Relationship between benchmark NSE and climate indices

We calculated three climate indices for each catchment following Knoben et al. (2018). We used WorldClim 2.1 data (Fick & Hijmans, 2017) for temperature (T) and precipitation (P) and the Global Aridity and PET database (Zomer et al., 2022) for potential evaporation (PET). The resolution of these data are 30 seconds (approximately 1 km at the equator). All calculations are performed on the raster data and then the indices are averaged over each catchment.

First, Thornthwaite's moisture index $\mathrm{MI(t)}$ was calculated for each month.

$$\mathrm{MI(t)} = \begin{cases} 1 - \frac{\mathrm{PET(t)}}{\mathrm{P(t)}}, & \mathrm{P(t)} \geq \mathrm{PET(t)} \\ \frac{\mathrm{P(t)}}{\mathrm{PET(t)}} - 1, & \mathrm{PET(t)} < \mathrm{P(t)} \end{cases} \quad \text{[S23]}$$

Then the aridity $\mathbf{I_m}$, the seasonality $\mathbf{I_{m,r}}$, and the fraction of precipitation as snow $\mathbf{f_s}$ are calculated:

$$\mathrm{I_m} = \frac{1}{12}\sum_{t=1}^{12}\mathrm{MI(t)} \quad \text{[S24]}$$

$$\mathrm{I_{m,r}} = \max\big(\mathrm{MI}(1,2,\ldots,12)\big) - \min\big(\mathrm{MI}(1,2,\ldots,12)\big) \quad \text{[S25]}$$

$$\mathrm{f_s} = \frac{\sum \mathrm{P(T(t) \leq 0°C)}}{\sum_{\mathrm{t}=1}^{12}\mathrm{P(t)}} \quad \text{[S26]}$$

Figure S3 shows scatterplots of the Benchmark NSE against these three climate indices. Figure S4 shows the benchmark NSE as a function of seasonality and snow fraction. We binned the catchments by $\mathrm{I_{m,r}}$ and $\mathrm{f_s}$ and took the median benchmark NSE for each 2D bin. Figure S5 shows the benchmark NSE as a function of seasonality and aridity, for snow-free catchments ($\mathrm{f_s} = 0$).

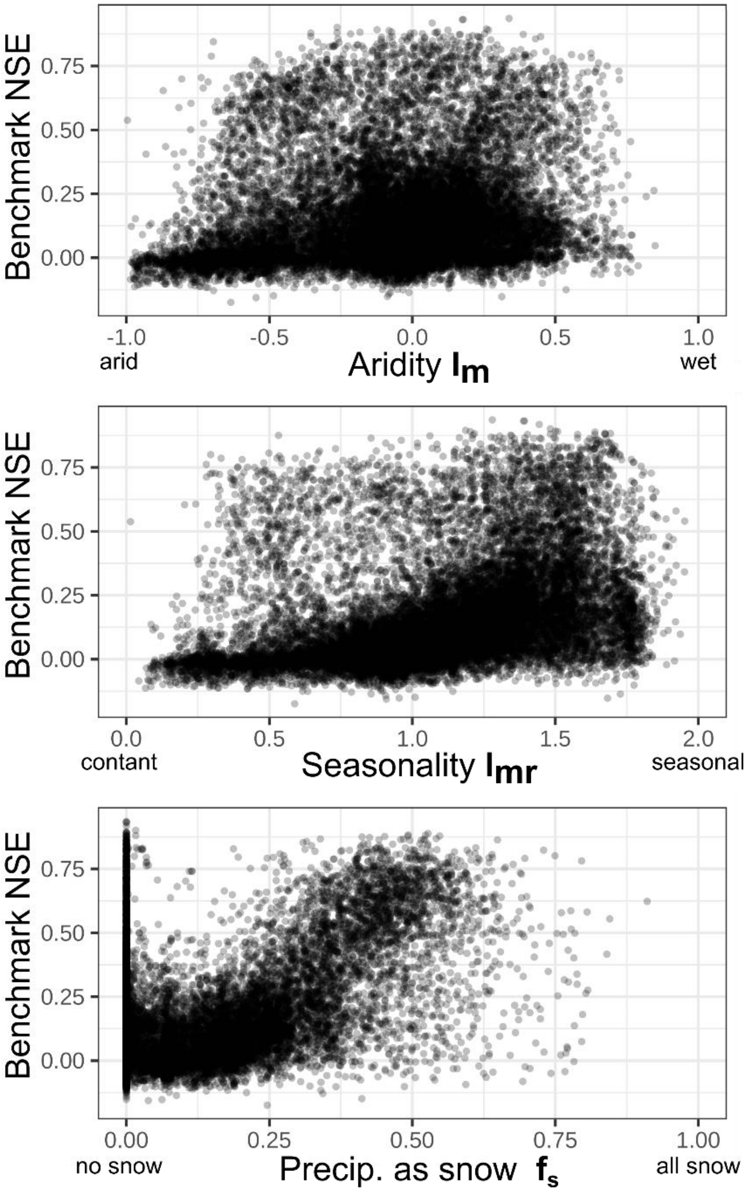

Figure S3: Scatterplots of the Benchmark NSE against aridity, aridity seasonality, and snow fraction. There is no clear relationship to aridity. Higher seasonality is associated with higher benchmark NSEs, but the relationship is noisy and many highly seasonal catchments have near-zero benchmark NSEs . On the other hand, increasing snow fraction (above about 0.25) is strongly associated with higher benchmark NSEs.

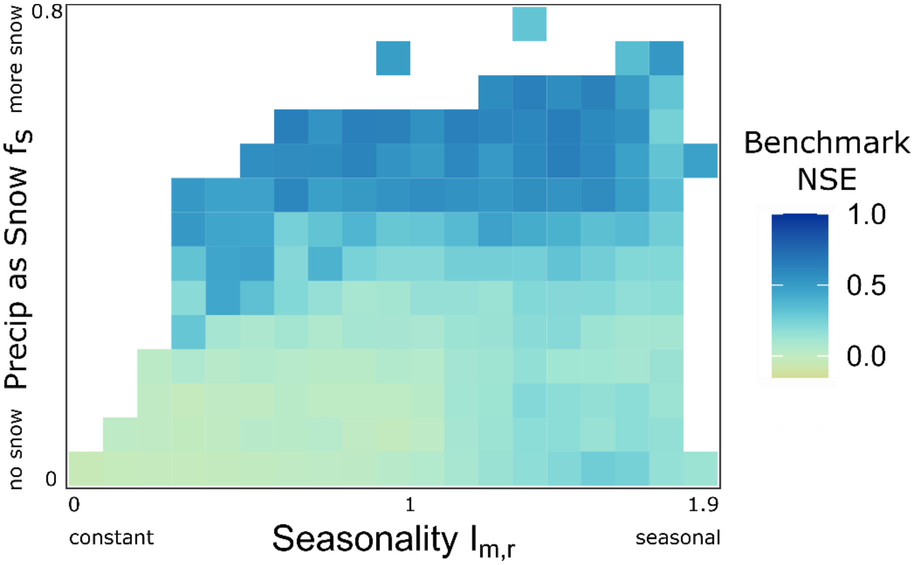

Figure S4: The median benchmark NSE for each cell in the climate space defined by seasonality $I_{m,r}$ and fraction of precipitation as snow f_s. Catchments with higher snow fractions have higher benchmark NSEs. There is a slight gradient in the seasonality, with more seasonal catchments exhibiting slightly higher benchmark NSEs overall.

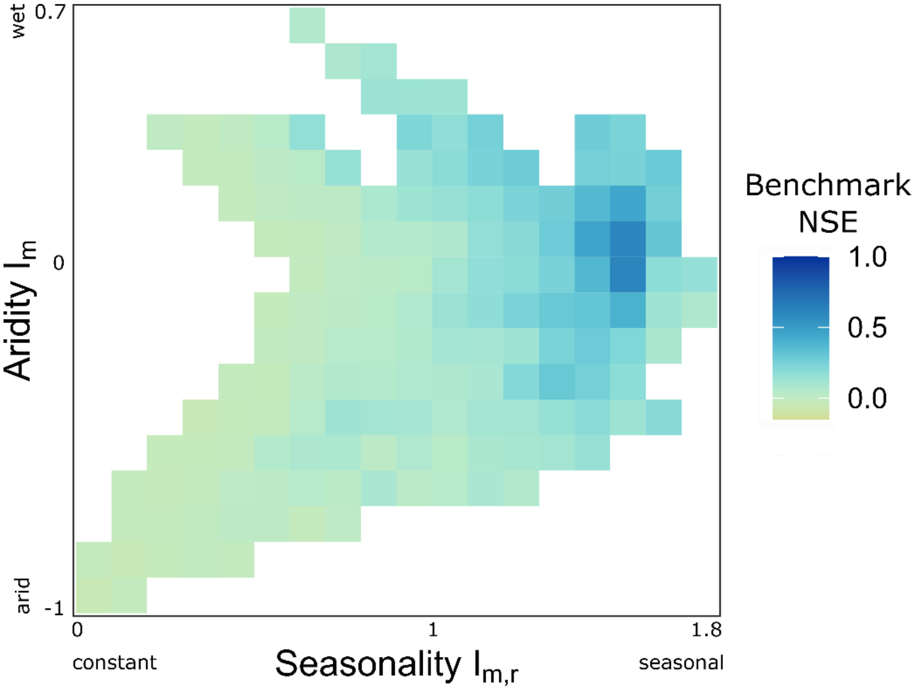

Figure S5: The median benchmark NSE for each cell in the climate space defined by seasonality and aridity, for catchments where the fraction of precipitation as snow is 0. In general more seasonal catchments have higher benchmark NSEs, although some very seasonal catchments have low benchmark NSEs.

S3: The overall NSE is the weighted mean of the component NSEs

We begin with the following definition for the NSE:

$$NSE = 1 - \frac{\sigma_\epsilon^2}{\sigma_o^2} \tag{S27}$$

Where σ_ϵ^2 is the error variance of the σ_o^2 is the variance of the observations. Then replace σ_ϵ^2 by its definition:

$$NSE = 1 - \frac{1}{\sigma_o^2} \times \frac{1}{N-1} \sum_{t=1}^{N} (q_o - q_s)^2 \tag{S28}$$

Where $q_o$and q_s are the observed and simulated discharge. $q_o$and q_s can also be written as the sum of their components: interannual i_o and i_s, seasonal s_o and s_s, and irregular r_o and r_s.

$$NSE = 1 - \frac{1}{\sigma_o^2} \times \frac{1}{N-1} \sum_{t=1}^{N} (i_o + s_o + r_o - i_s - s_s - s_r)^2 \tag{S29}$$

We can expand the brackets and use the orthogonality of the decomposition to cancel terms. In particular, all cross-component terms (eg $i_o \times s_o$ or $i_o \times s_s$) can be cancelled because the decomposition (both the Fast Fourier Transform and the calculation of the seasonal component) is based on orthogonal basis functions[2].

$$NSE = 1 - \frac{1}{\sigma_o^2} \times \frac{1}{N-1} \sum_{t=1}^{N} \left(\begin{array}{c} \boldsymbol{i_o^2} + \cancel{i_o s_o} + \cancel{i_o r_o} - \boldsymbol{i_o i_s} - \cancel{i_o s_s} - \cancel{i_o r_s} + \\ \cancel{s_o i_o} + \boldsymbol{s_o^2} + \cancel{s_o r_o} - \cancel{s_o i_s} - \boldsymbol{s_o s_s} - \cancel{s_o r_s} + \\ \cancel{r_o i_o} + \cancel{r_o s_o} + \boldsymbol{r_o^2} - \cancel{r_o i_s} - \cancel{r_o s_s} - \boldsymbol{r_o r_s} + \\ -\boldsymbol{i_s i_o} - \cancel{i_s s_o} - \cancel{i_s r_o} + \boldsymbol{i_s^2} + \cancel{i_s s_s} + \cancel{i_s r_s} + \\ -\cancel{s_s i_o} - \boldsymbol{s_s s_o} - \cancel{s_s r_o} + \cancel{s_s i_s} + \boldsymbol{s_s^2} + \cancel{s_s r_s} + \\ -\cancel{r_s i_o} - \cancel{r_s s_o} - \boldsymbol{r_s r_o} + \cancel{r_s i_s} + \cancel{r_s s_s} + \boldsymbol{r_s^2} \end{array} \right) \tag{S30}$$

$$NSE = 1 - \frac{1}{\sigma_o^2} \times \frac{1}{N-1} \sum_{t=1}^{N} (i_o^2 - 2 i_o i_s + i_s^2 + s_o^2 - 2 s_o s_s + s_s^2 + r_o^2 - 2 r_o r_s + r_s^2) \tag{S31}$$

$$NSE = 1 - \frac{1}{\sigma_o^2} \times \frac{1}{N-1} \sum_{t=1}^{N} \{(i_o - i_s)^2 + (s_o - s_s)^2 + (r_o - r_s)^2\} \tag{S32}$$

We can define the error variance of the interannual, seasonal, and irregular components as $\sigma_{\epsilon,i}^2$, $\sigma_{\epsilon,s}^2$, and $\sigma_{\epsilon,r}^2$, respectively:

$$\sigma_{\epsilon,i}^2 = \frac{1}{N-1} \sum_{t=1}^{N} (i_o - i_s)^2 \tag{S33}$$

$$\sigma_{\epsilon,s}^2 = \frac{1}{N-1} \sum_{t=1}^{N} (s_o - s_s)^2 \tag{S34}$$

$$\sigma_{\epsilon,r}^2 = \frac{1}{N-1} \sum_{t=1}^{N} (r_o - r_s)^2 \tag{S35}$$

Then rewrite equation S32:

[2] This ignores leap years and assumes an integer number of years of data. In practice when each year does not have exactly 365 days there can be small deviations from orthogonality. However, we found that these effects are negligible: across 16988 modeled catchments from the 18 models that we analysed, equation S39 was always accurate within an error of 3×10^{-8}.

$$NSE = 1 - \frac{\sigma_{\epsilon,i}^2+\sigma_{\epsilon,s}^2+\sigma_{\epsilon,r}^2}{\sigma_o^2} \qquad \text{[S36]}$$

Using the fact that the sum of the variance fractions is 1:

$$NSE = \frac{\sigma_{interannual}^2+\sigma_{seasonal}^2+\sigma_{irregular}^2}{\sigma_o^2} - \frac{\sigma_{\epsilon,i}^2+\sigma_{\epsilon,s}^2+\sigma_{\epsilon,r}^2}{\sigma_o^2} \qquad \text{[S37]}$$

$$NSE = \frac{\sigma_{interannual}^2}{\sigma_o^2}\left(1 - \frac{\sigma_{\epsilon,i}^2}{\sigma_{interannual}^2}\right) + \frac{\sigma_{seasonal}^2}{\sigma_o^2}\left(1 - \frac{\sigma_{\epsilon,s}^2}{\sigma_{seasonal}^2}\right) + \frac{\sigma_{irregular}^2}{\sigma_o^2}\left(1 - \frac{\sigma_{\epsilon,r}^2}{\sigma_{irregular}^2}\right) \qquad \text{[S38]}$$

$$NSE = \frac{\sigma_{interannual}^2}{\sigma_o^2} NSE_{interannual} + \frac{\sigma_{seasonal}^2}{\sigma_o^2} NSE_{seasonal} + \frac{\sigma_{irregular}^2}{\sigma_o^2} NSE_{irregular} \qquad \text{[3S9]}$$

Equation S39 is the weighted sum of the component NSEs, so we are done.

S4: Comparing Goodness-of-fit statistics for models based on different thresholds and indices

The following figures show alternative versions of Figure 3 in the paper, using different thresholds (seasonal variance fraction of 0.4 and 0.6) and different indices (the streamflow concentration index QCI, the coefficient of variation of the streamflow (COV(Q)), the fraction of precipitation as snow (f_s) and the aridity seasonality index $I_{m,r}$.

f_s and $I_{m,r}$ are defined above. The streamflow concentration index is defined following Han et al (2024):

$$\mathrm{QCI} = \frac{\sum_{\mathrm{i}=1}^{12} \mathrm{Q}_{\mathrm{i}}^2}{\left(\sum_{\mathrm{i}=1}^{12} \mathrm{Q}_{\mathrm{i}}\right)^2} \times 100 \qquad \text{[S40]}$$

Where Q_{I} is the monthly climatological streamflow. QCI ranges from a theoretical minimum value of $8.\underline{3}$ (constant streamflow throughout the year) to a maximum of 100 (all streamflow occurs in one month).

We chose to use the coefficient of variation of the streamflow (COV(Q)) because the COV has been used to measure seasonality in precipitation (eg. Fick & Hijmans, 2017). We calculate the COV of the climatological streamflow Q_{d} (the interannual mean of each calendar day). For leap years both December 30 and 31 were used as the 365th day of the year.

$$COV(Q) = \frac{\sigma(\bar{Q}_d)}{\mu(\bar{Q}_d)}, d = 1,2,\ldots,365 \qquad \text{[S41]}$$

NSE calculated for each interannual signature metric

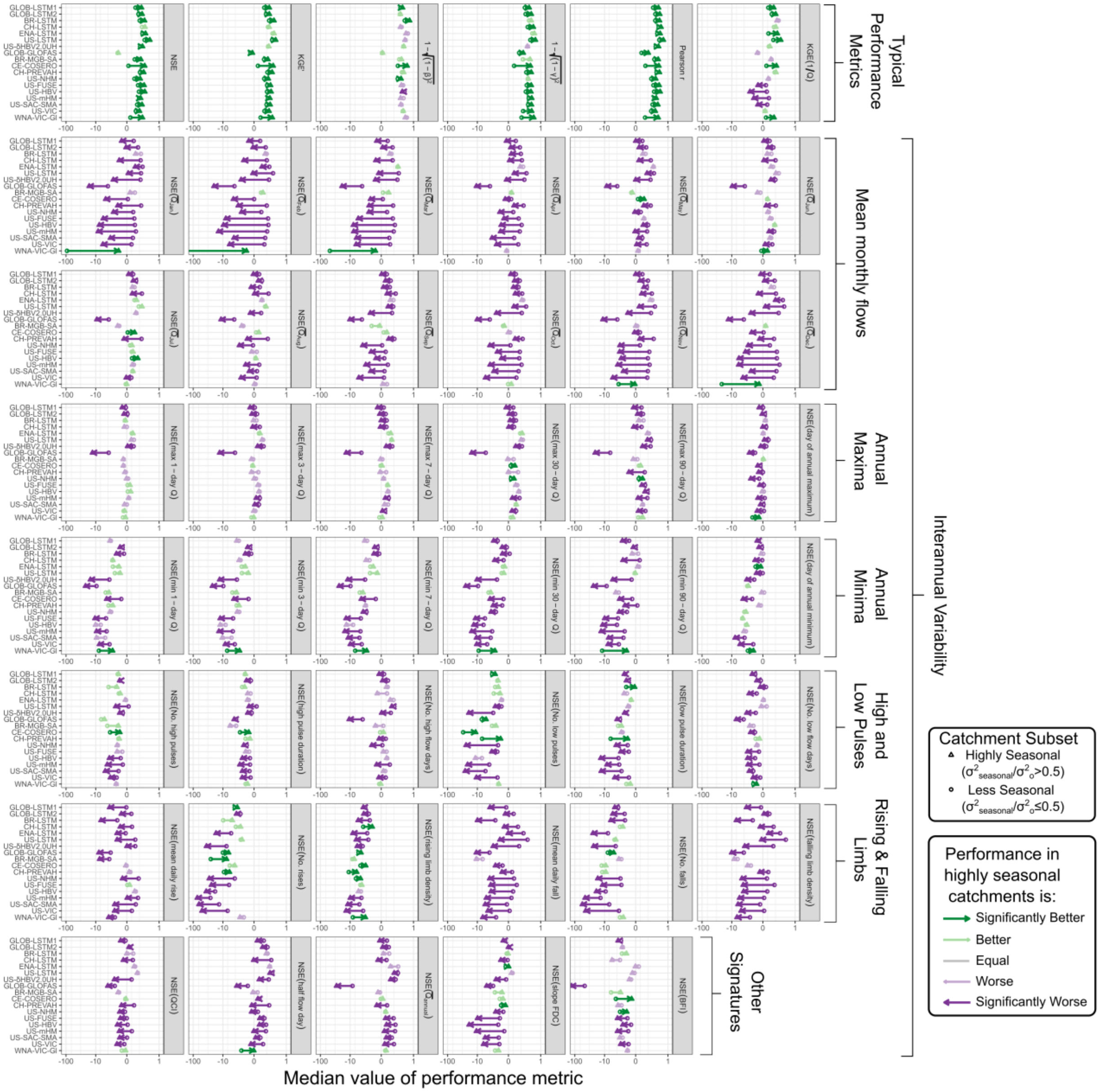

Figure S6: Equivalent to Figure 4 in text but using the NSE of each interannual metric, rather than the correlation. Note that many values are negative, but the overall pattern is similar.

Comparison: Seasonal Variance Fraction >/≤ 0.4

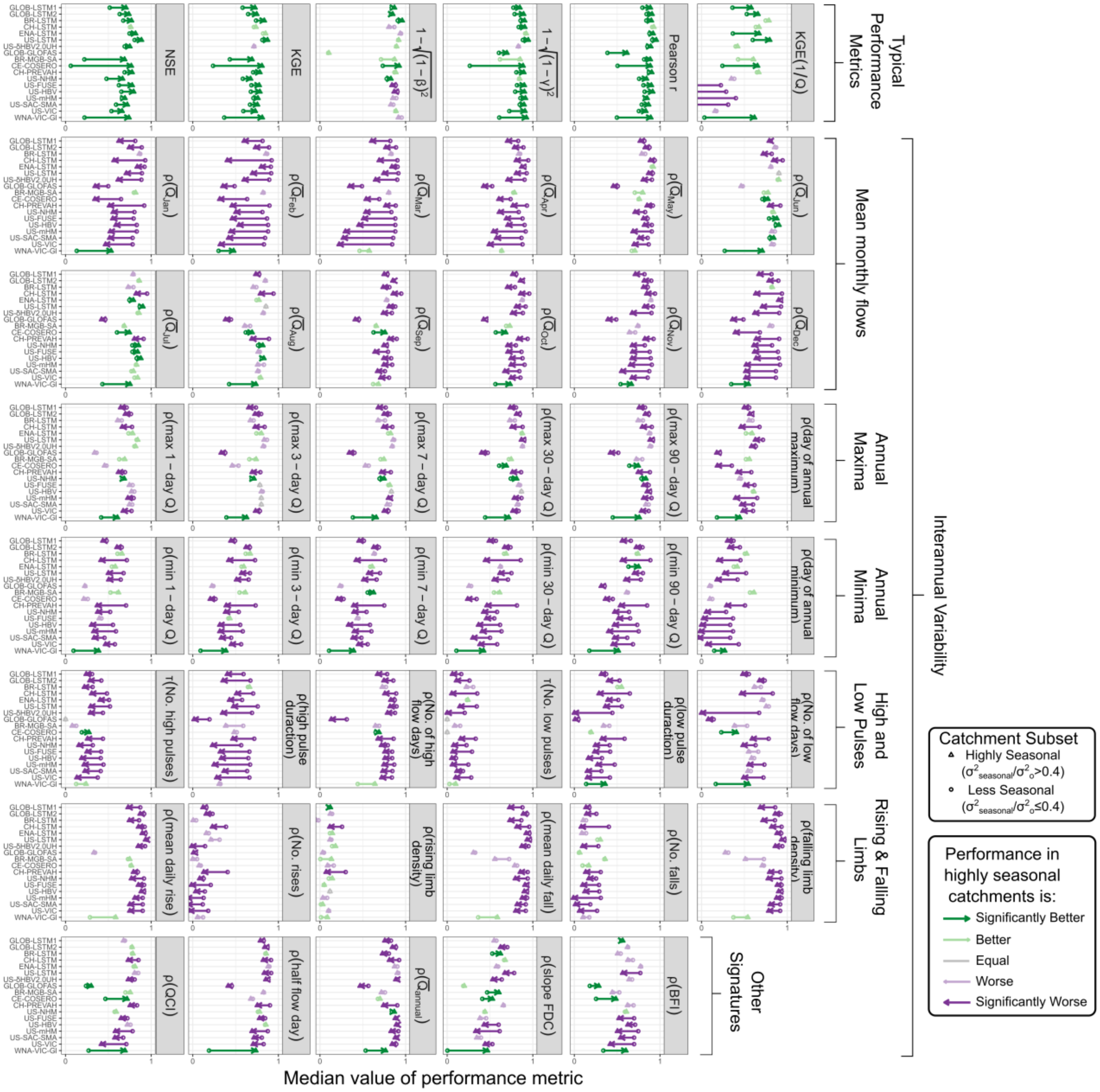

Figure S7: Equivalent to Figure 4 in text but using a threshold of 0.4 for the benchmark NSE to divide catchments into high-benchmark and low-benchmark groups.

Comparison: Seasonal Variance Fraction >/≤ 0.6

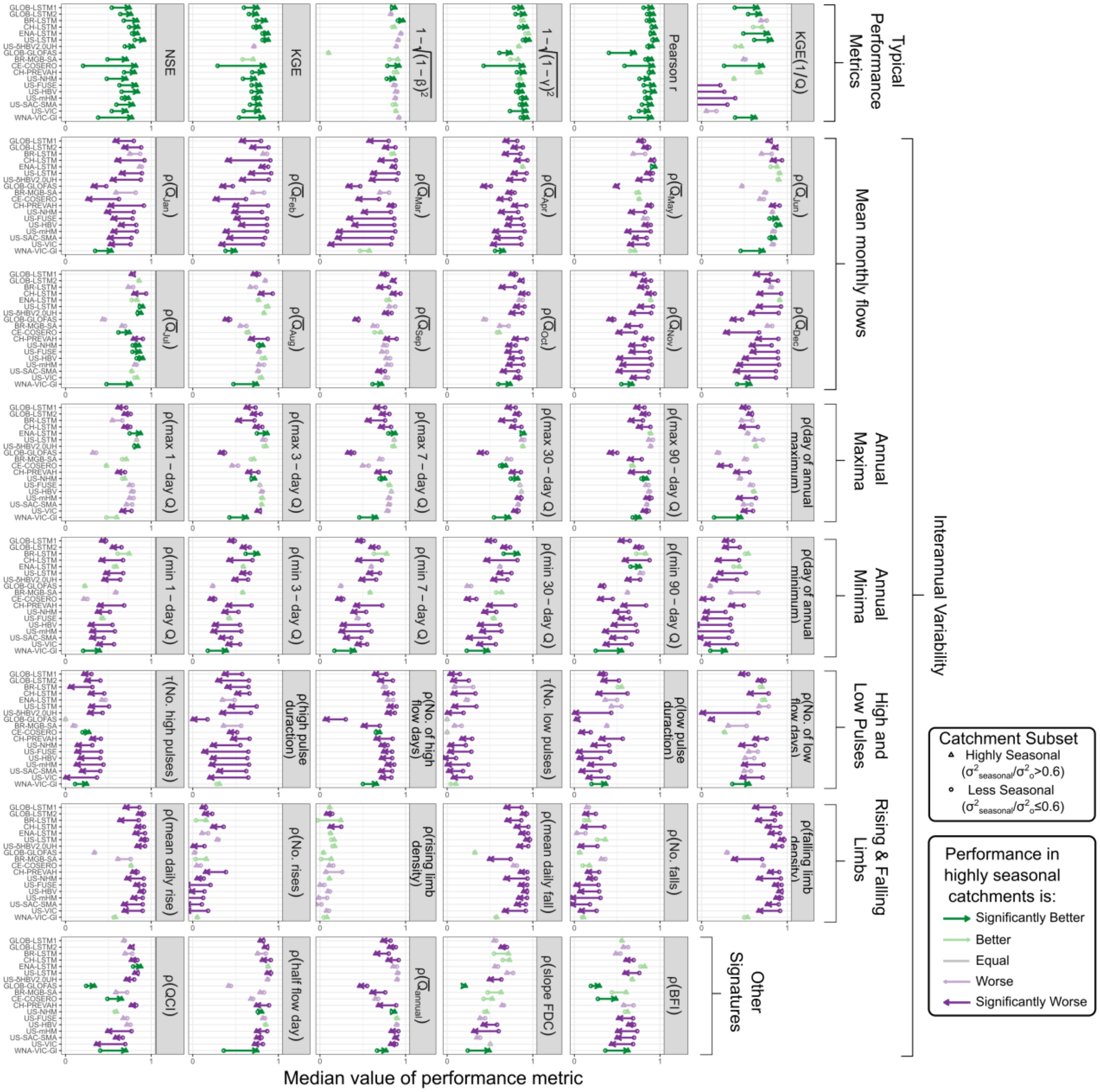

Figure S8: Equivalent to Figure 4 in text but using a threshold of 0.6 for the benchmark NSE to divide catchments into high-benchmark and low-benchmark groups.

Comparison: QCI >/≤ 15

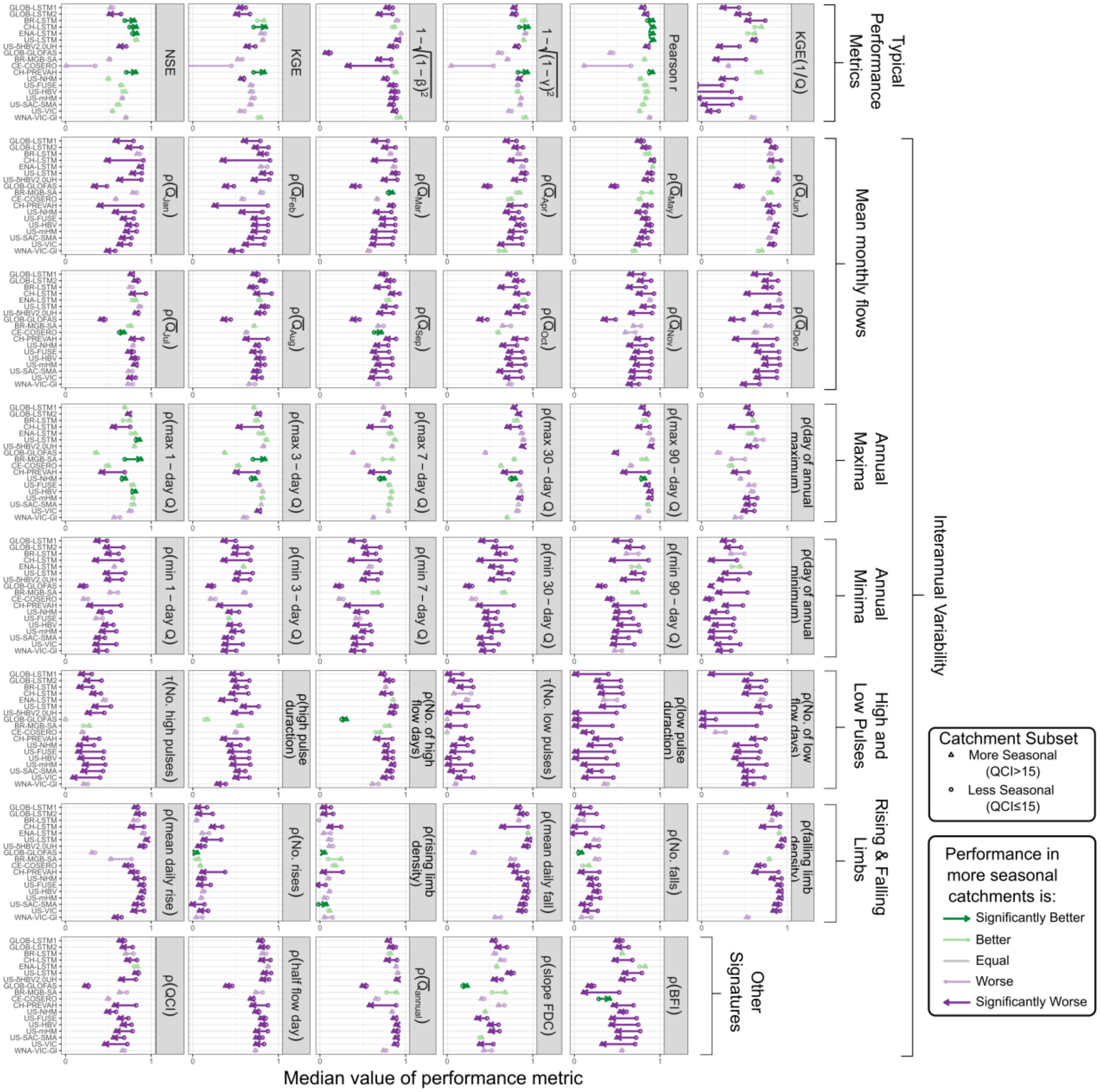

Figure S9: Equivalent to Figure 4 in text but catchments are divided by the streamflow concentration index (QCI) using a threshold of 15.

Comparison: COV(Q) >/≤ 1

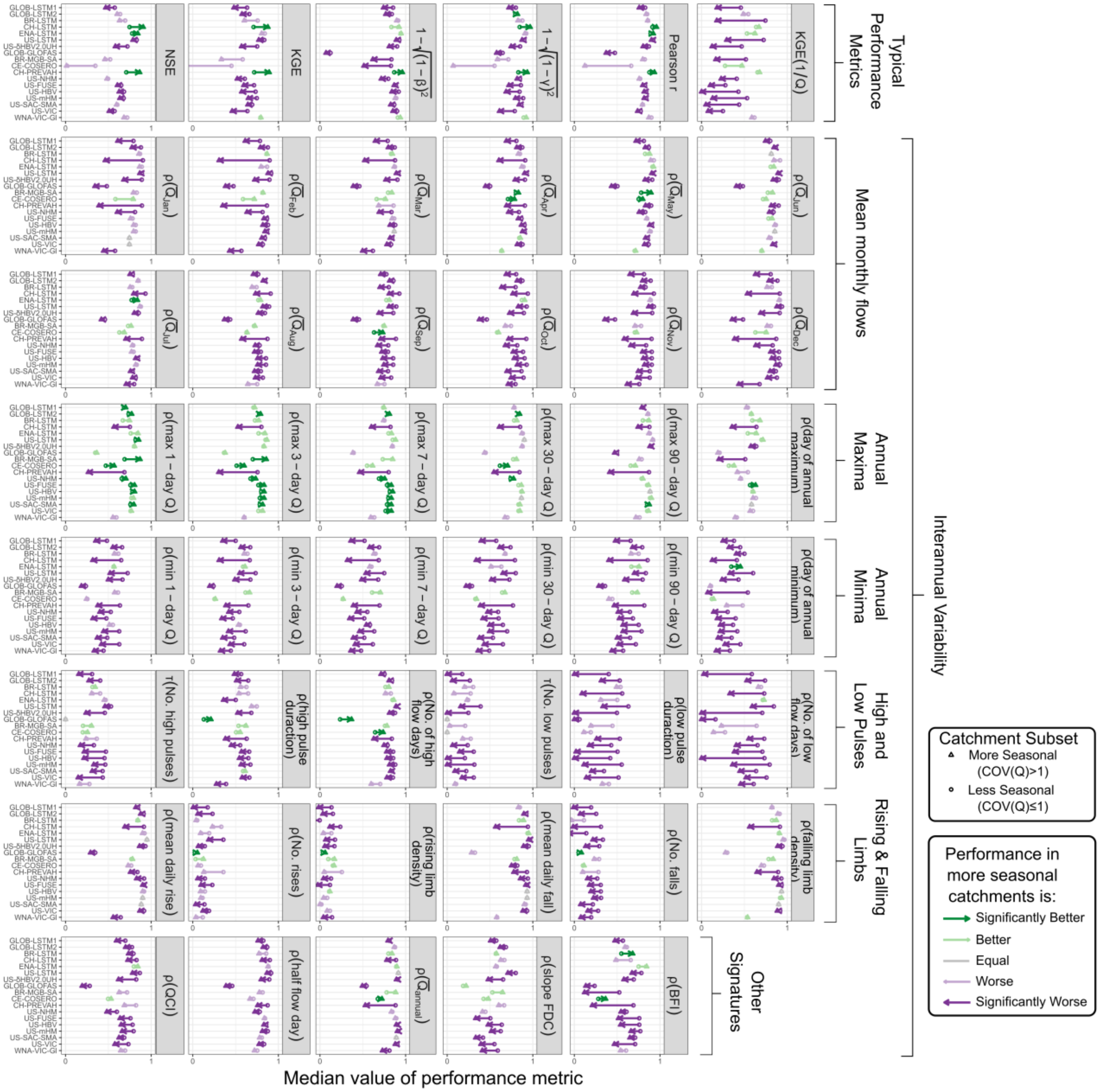

Figure S10: Equivalent to Figure 4 in text but catchments are divided by the coefficient of variation of the mean annual hydrograph (COV(Q)) using a threshold of 1.

Comparison: f_s >/≤ 0.5

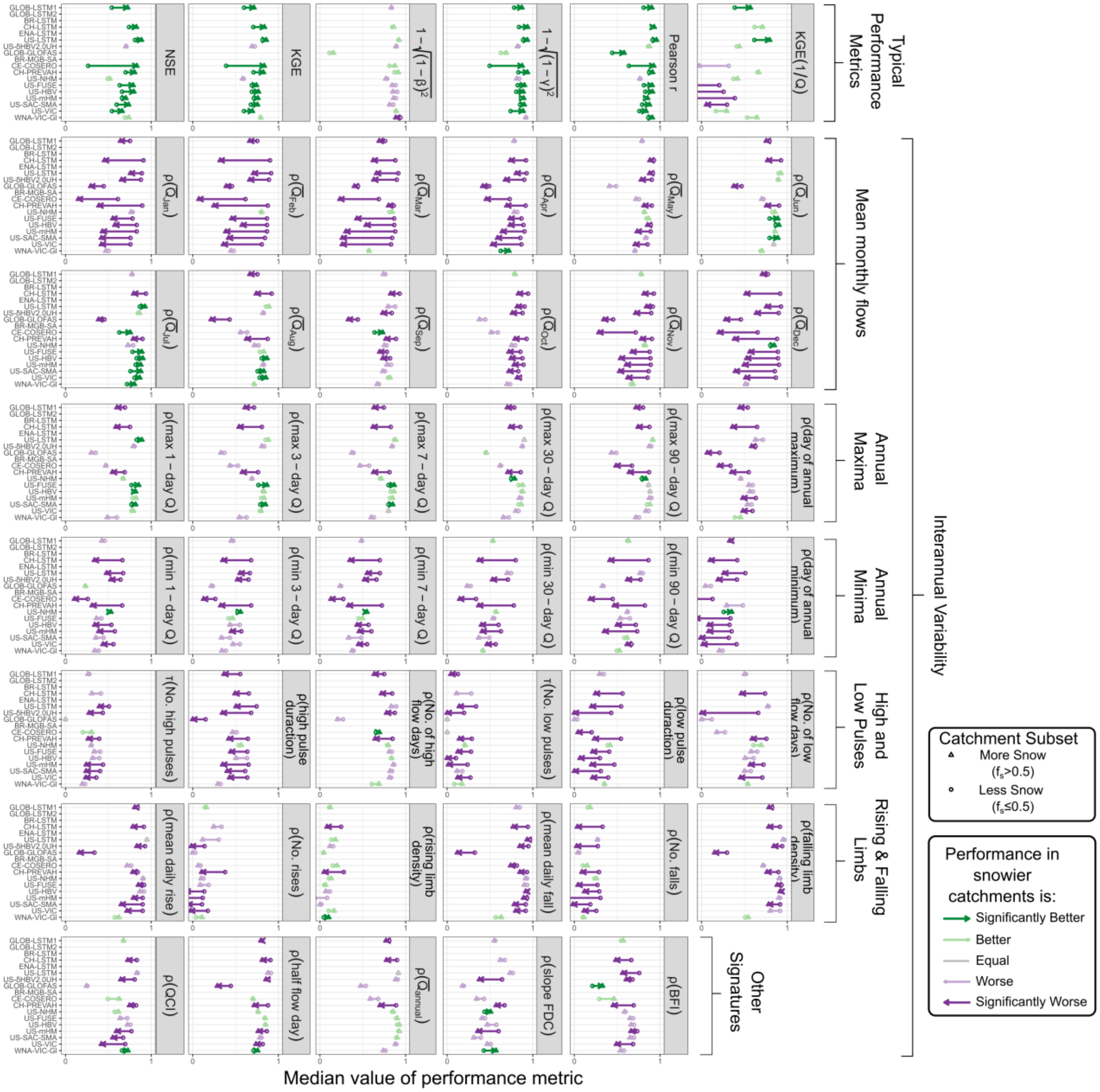

Figure S11: Equivalent to Figure 4 in text but catchments are divided into snowier and less snowy groups using a threshold of 0.5 for the snow fraction (f_s).

Comparison: $I_{m,r}$>/≤ 1

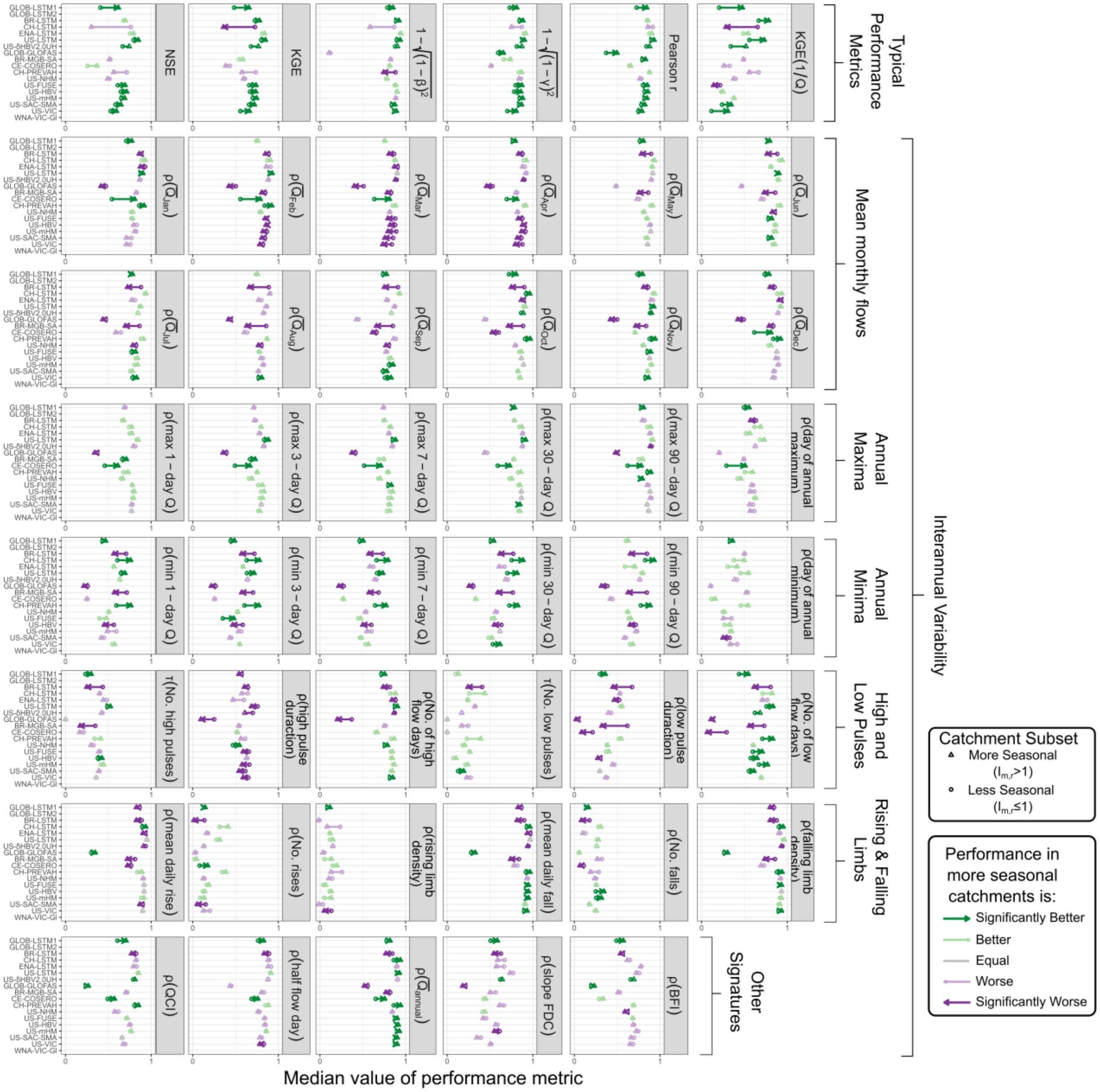

Figure S12: Equivalent to Figure 4 in text but catchments are divided more seasonal and less seasonal groups using a threshold of 1 for the aridity seasonality index ($I_{m,r}$).

S5: Long Short-Term Memory model for Brazil

Since we identified Brazil as a location with very high benchmark NSE values, as well as a large amount of good-quality data, we wanted to include some models from Brazil in our analysis. Unfortunately we were unable to find any freely accessible machine learning hydrologic models that included catchments from across Brazil, so we created a Long Short-Term Memory (LSTM)model using the Camels-BR dataset, version 1.2 (Chagas et al., 2020, 2025)

We created an LSTM using the neuralhydrology package for Python (Kratzert et al., 2022). We used data from the period 01/01/1980 to 30/12/2020, which represents a compromise between maximizing record length and maximizing the number of available input datasets. The Camels-BR dataset includes streamflow and meteorological data for 897 catchments, and we used all catchments for training, validation, and testing. We trained on data from 2010-2020, validated on data from 1980-1989, and tested on data from 1990-2009. We reserved a long period (20 years) for testing because the objective here is to analyse differences in testing performance across catchment types, and not necessarily to maximize the model performance overall.

We included all available static attributes in the model, in addition to one-hot encoding for the basin ID.

For dynamic attributes we included all variables that were available for the full 41-year period. These are summarized in Table S1.

Table S1: Dynamic Variables used in the LSTM

Variable	**Source(s)**
Precipitation	CHIRPS[1], CPC[2], ERA5-Land[3], MSWEP[4]
Minimum Temperature	CPC[2], ERA5-Land[3]
Maximum Temperature	CPC[2], ERA5-Land[3]
Mean Temperature	ERA5-Land[3]
Actual Evapotranspiration	Gleam[5], ERA5-Land[3]
Potential Evapotranspiration	Gleam, ERA5-Land[3]
Soil Moisture (surface)	Gleam[5]
Soil Moisture (root zone)	Gleam[5]
Soil Moisture (layers 1-4)	ERA5-Land[3]

[1](Funk et al., 2015), [2](Chen & Xie, 2008; NOAA Physical Sciences Laboratory, 2025), [3](Muñoz Sabater, 2019), [4](Beck et al., 2019) [5]

The most important hyperparameters are summarized below in table S2.

Table S2: Hyperparameters used in the LSTM

Hyperparameter	**Value**
Hidden size	256
Batch size	256
Sequence length	365
Initial forget bias	3
Output dropout	0.4
Output activation	Linear
Optimizer	Adam
Loss	NSE
Epochs	50
Learning rate	1e-4 (epochs 1-30) 1e-5 (epochs 31-40) 5e-6 (epochs 41-50

These values are typical for LSTMs (eg. Kratzert et al., 2024). We did not tune the hyperparameters except for the learning rate, which we reduced because with a typical learning rate of 1e-3, the maximum validation NSE occurred on the first epoch. Even with the reduced learning rate the maximum validation NSE tended to occur within the first ten epochs. Further reductions to the learning rate resulted in a lower maximum validation NSE.

We generated an ensemble of five models with the same hyperparameters, and averaged the predictions.

The validation and testing NSE or the ensemble model are shown in Figure S10. The median validation NSE is 0.75, while the median test NSE is 0.72.

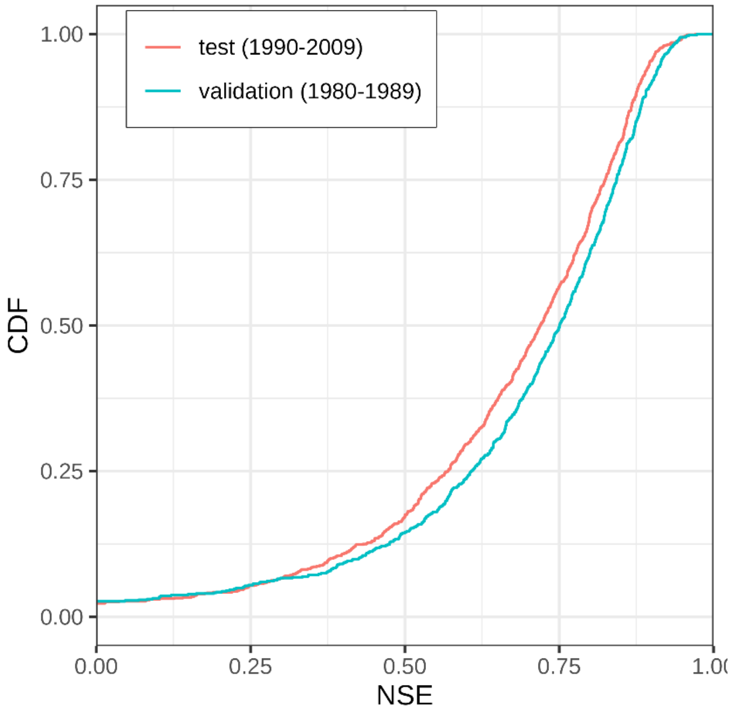

Figure S13: The empirical cumulative density function of NSE values for the ensemble LSTM model.

S6: Variance Components

Figure S14 shows histograms of the three variance components for 17,245 catchments. Figures S15 to S32 show examples of decomposed time series for 18 example catchments from around the world. Figures S15-S20 show catchments with high interannual variance, Figs. S21-S26 show catchments with high irregular variance, and Figs. S27-S32 show catchments with high seasonal variance. The examples were not chosen systematically but are intended to represent a broad geographical range.

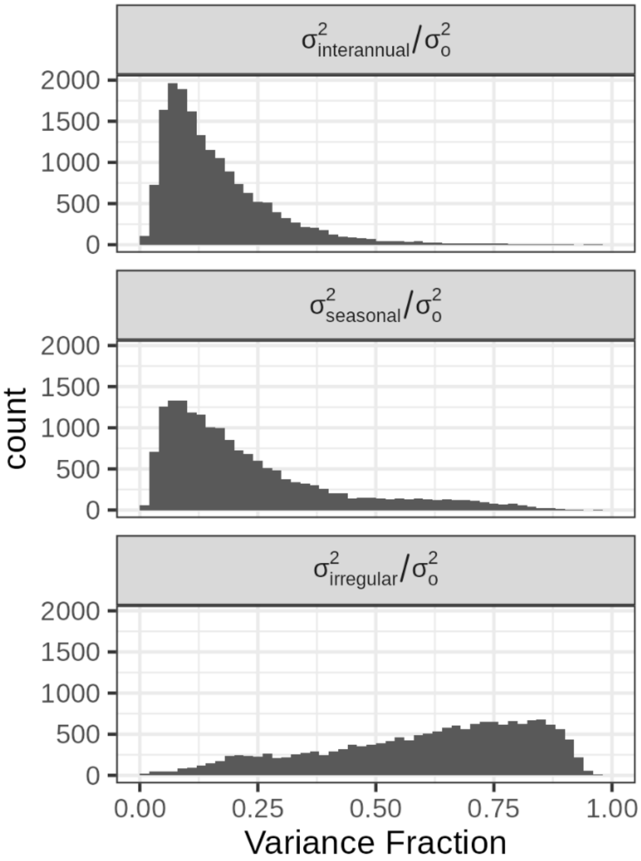

Figure S14: The distribution of variance fractions for all 17,245 catchments plotted in Figure 1 of the manuscript.

S6.1: Examples of interannually variable streams

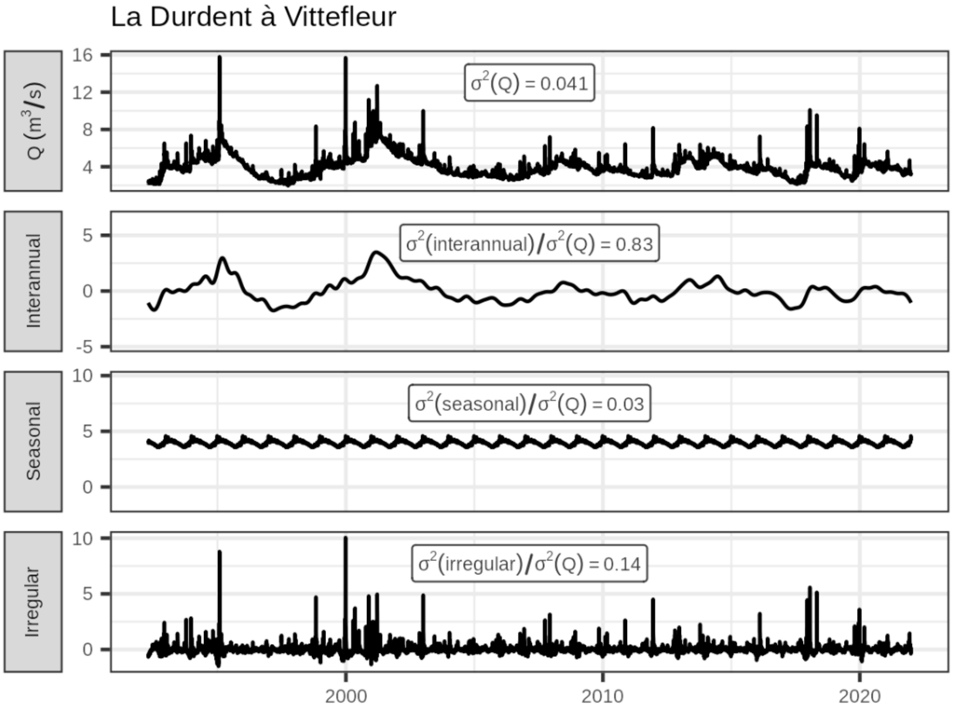

Figure S15: Decomposed time series for La Durdent à Vittefleur (Sandre G600061010), an interannually variable stream in Normandy, France.

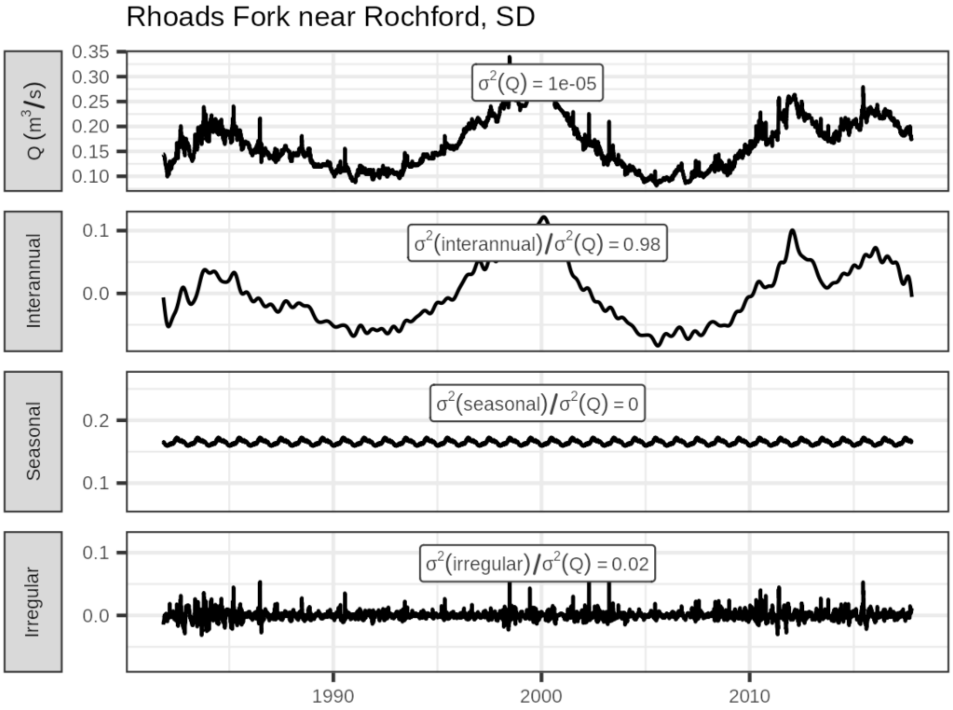

Figure S16: Decomposed time series for Rhoads Fork Near Rochford, SD (USGS 06408700), an interannually variable stream in South Dakota, United States.

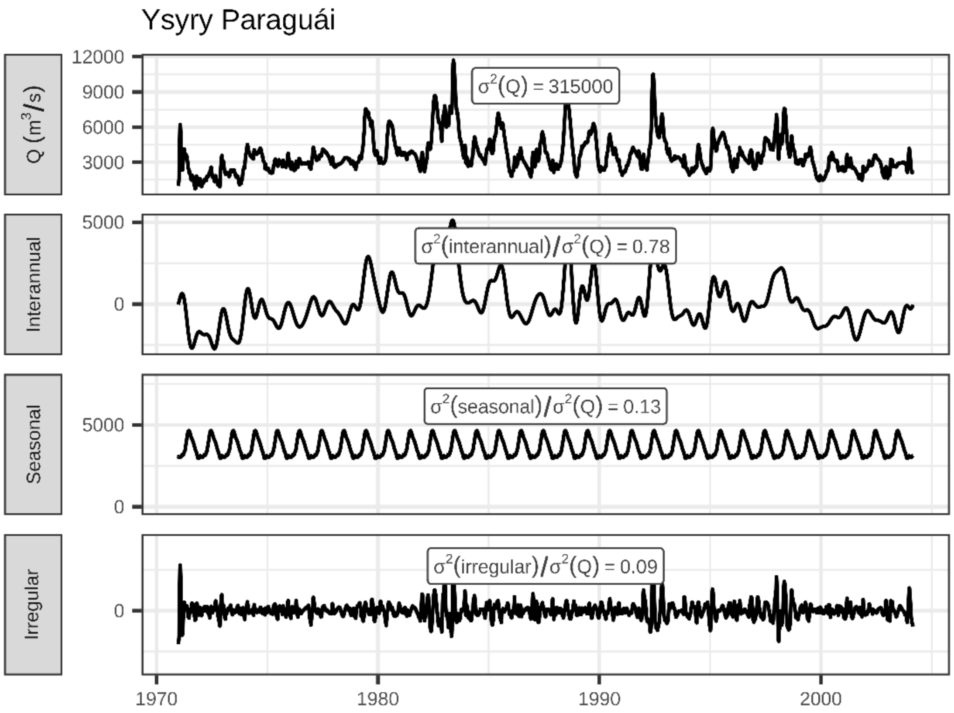

Figure S17: Decomposed time series for Ysyry Paraguái (Paraguay River) at Asunción, Paraguay, (GRDC 3368100) an interannually variable river.

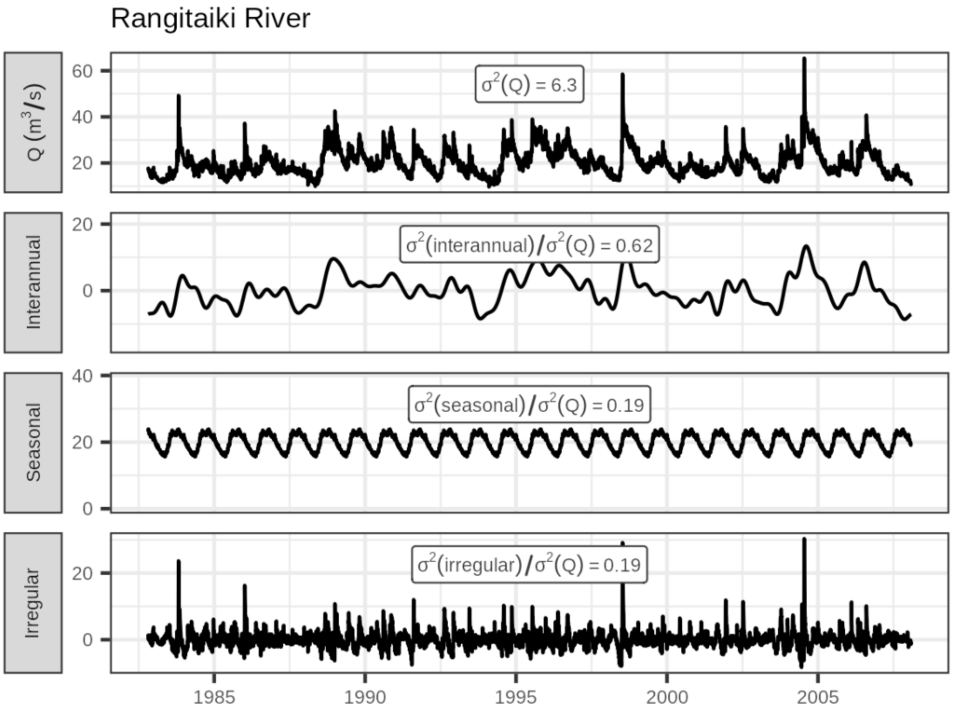

Figure S18: Decomposed time series for the Rangitaiki River at Murupara, Aotearoa (New Zealand), (GRDC 5863120) a river with 62% interannual variance.

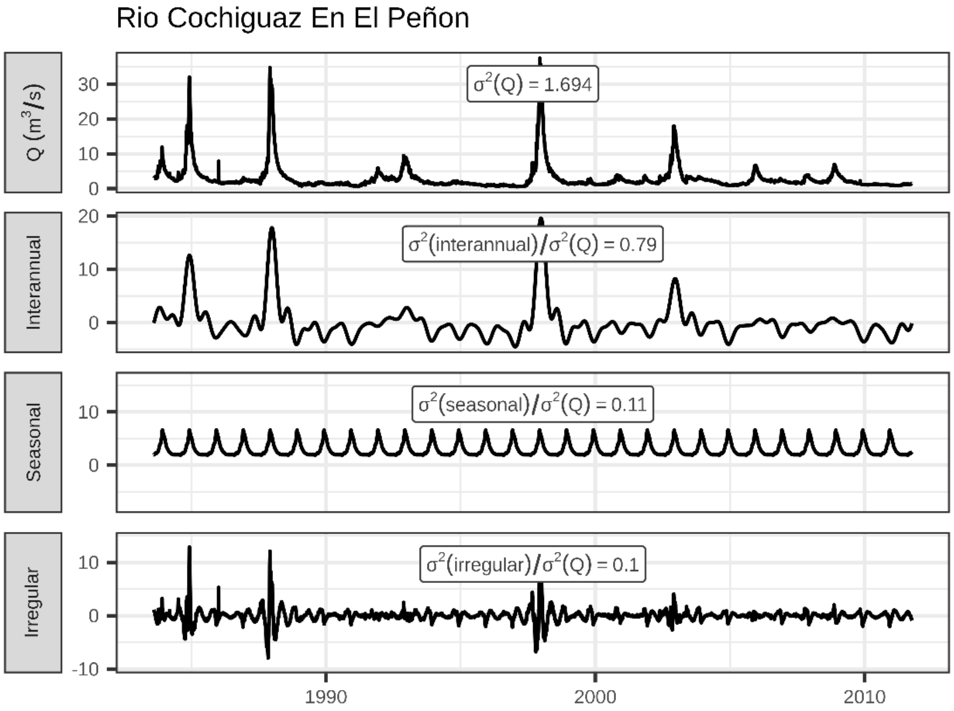

Figure S19: Decomposed time series for the Cochiguaz River (El Peñon, Chile), (Dirección General de Aguas 4313001) an interannually variable stream.

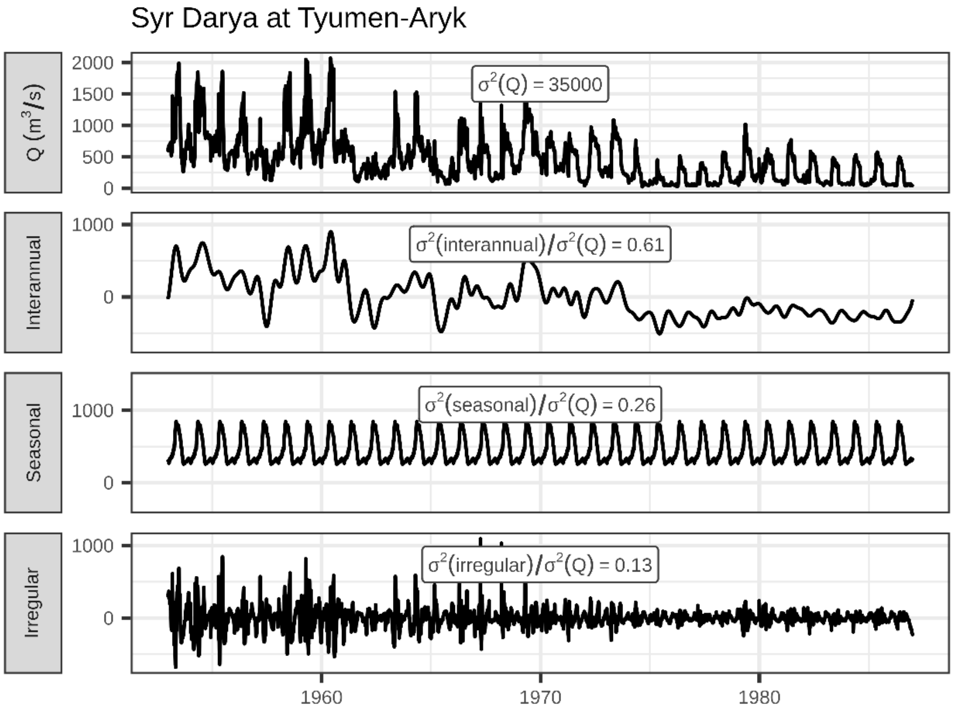

Figure S20: Decomposed time series for the Syr Darya (Tyumen-Aryk, Kazakhstan), (GRDC 2316200) an interannually variable stream, where interannual variability has been driven largely by water withdrawals for irrigation beginning in 1973 (Zou et al., 2019)

S6.2: Examples of highly irregular streams

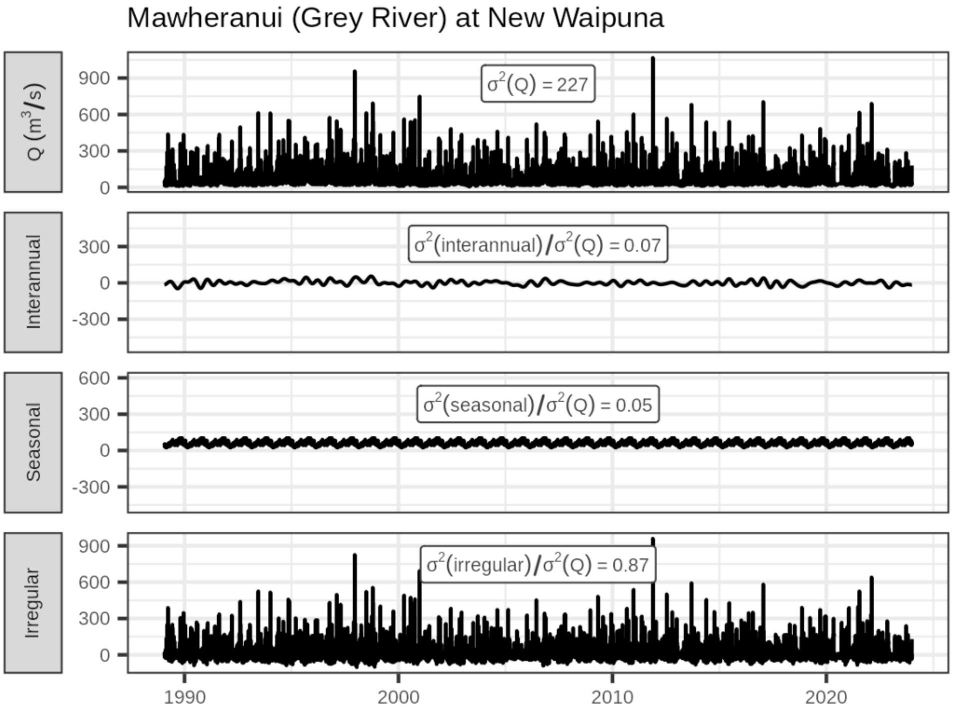

Figure S21: Decomposed time series for the Māwheranui River (New Zealand), (GRDC 5867710), a stream with highly irregular variance.

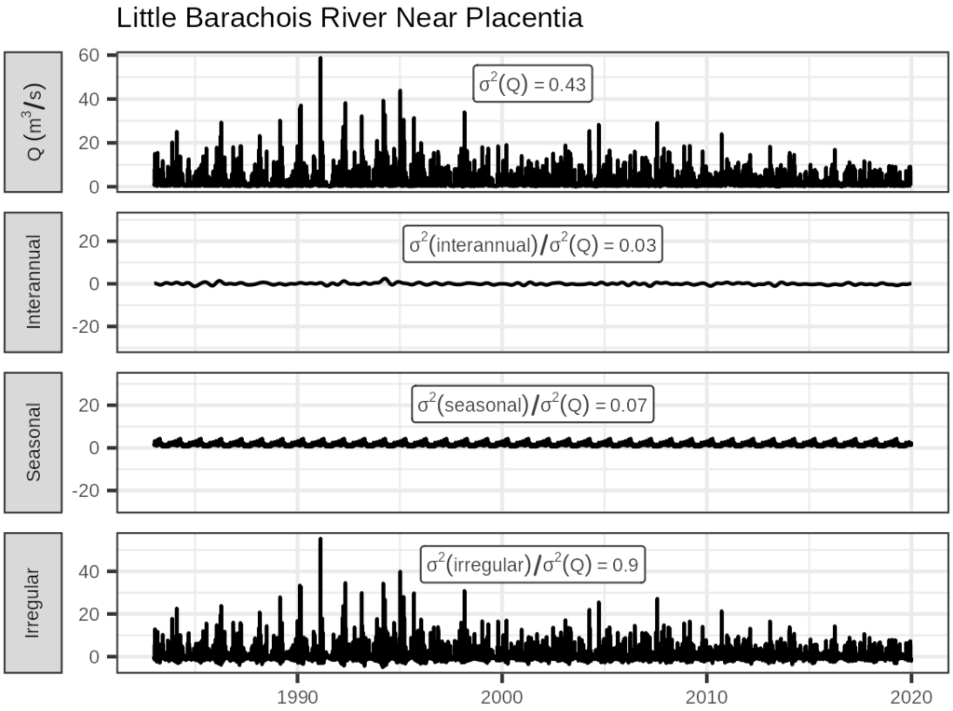

Figure S22: Decomposed time series for the Little Barachois River (Newfoundland, Canada), (Water Survey of Canada 02ZK003), a stream with highly irregular variance.

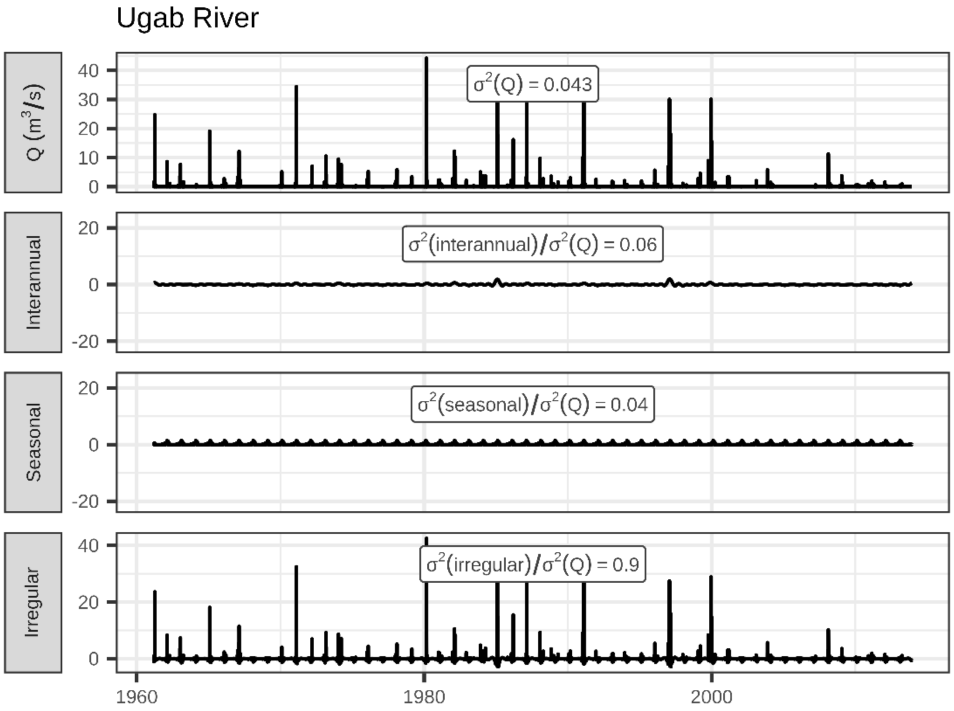

Figure S23: Decomposed time series for the Ugab River (Namibia), (GRDC 1258202), a stream with highly irregular variance.

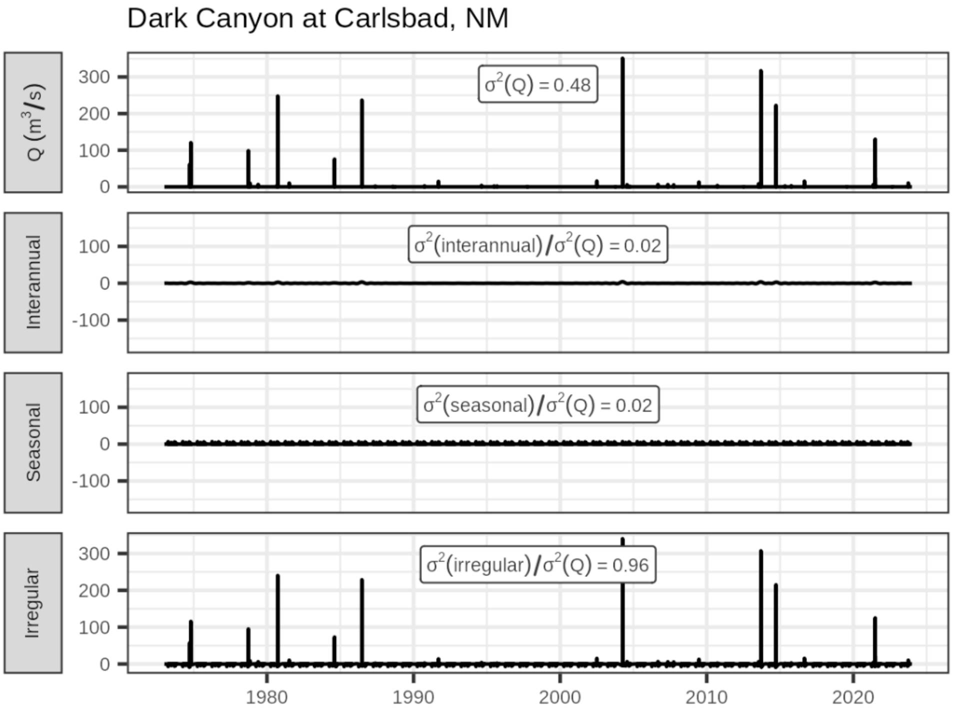

Figure S24: Decomposed time series for Dark Canyon at Carlsbad (New Mexico, United States), (USGS 08405150), a stream with highly irregular variance.

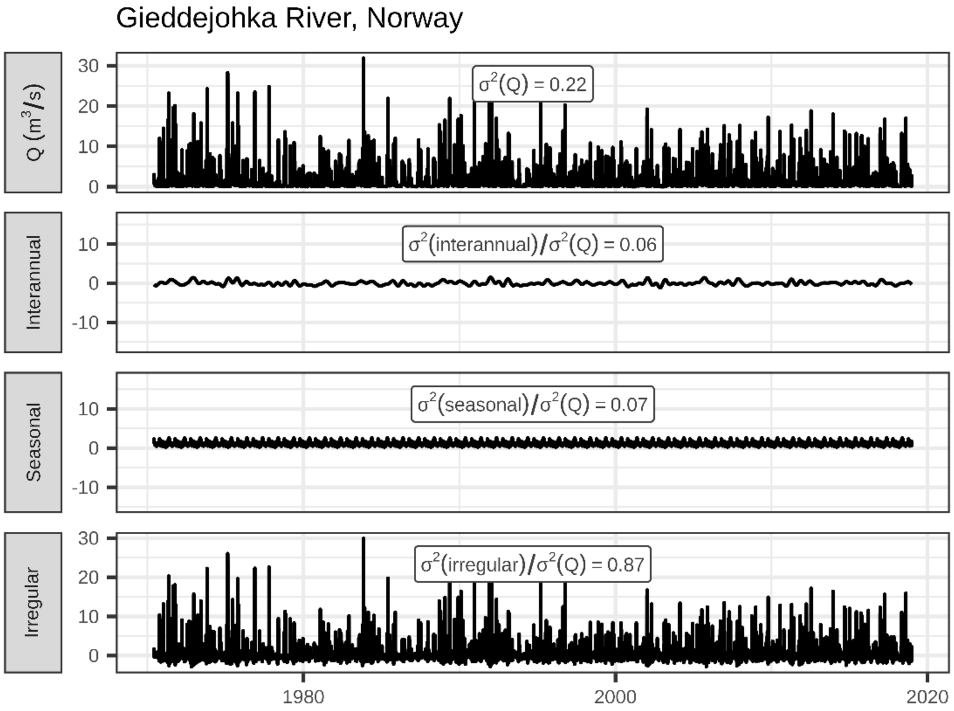

Figure S25: Decomposed time series for the Gieddejohka River (Leirpoldvatn, Norway), (GRDC 6731750), a stream with highly irregular variance.

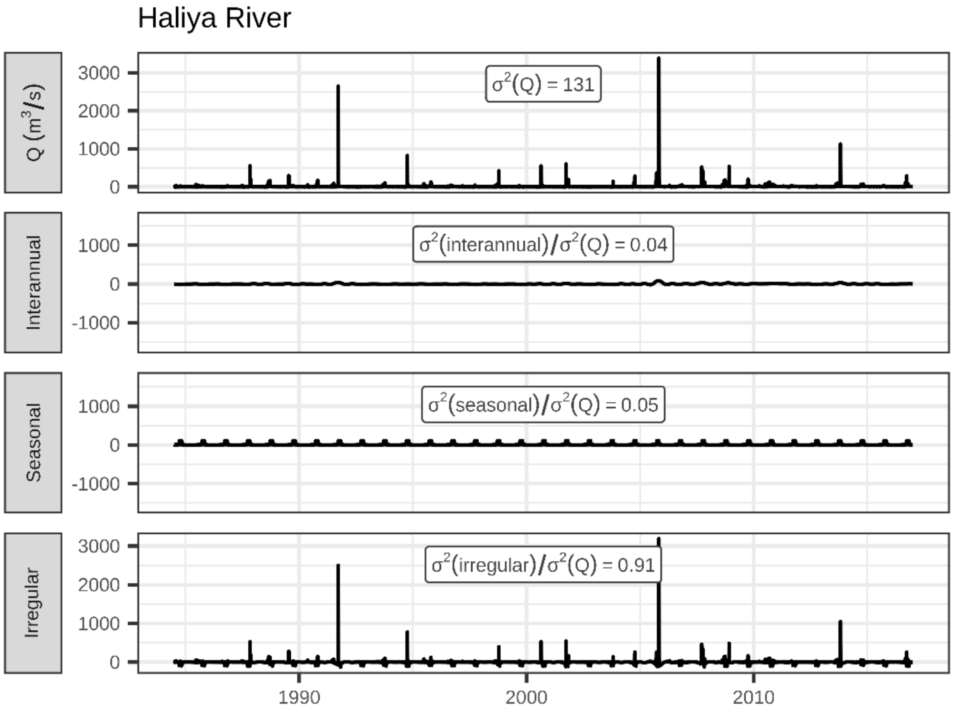

Figure S26: Decomposed time series for the Haliya River (Telangana, India), (Camels-IND 04012), a stream with highly irregular variance.

S6.3: Examples of highly seasonal streams

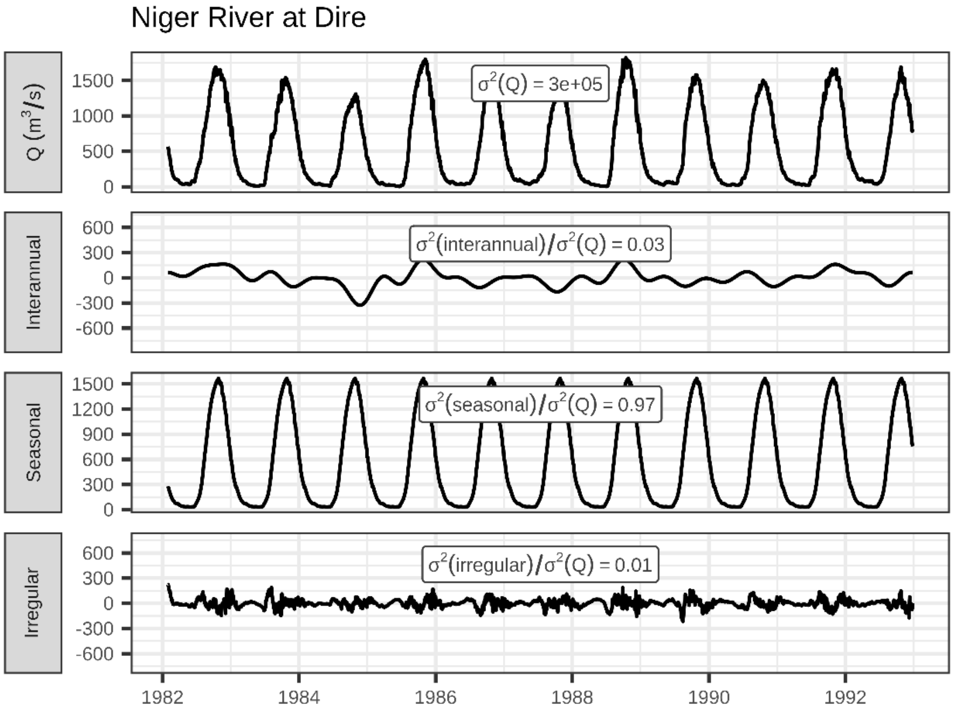

Figure S27: Decomposed time series for the Niger River at Dire (Mali), (GRDC 1134700), a highly seasonal river.

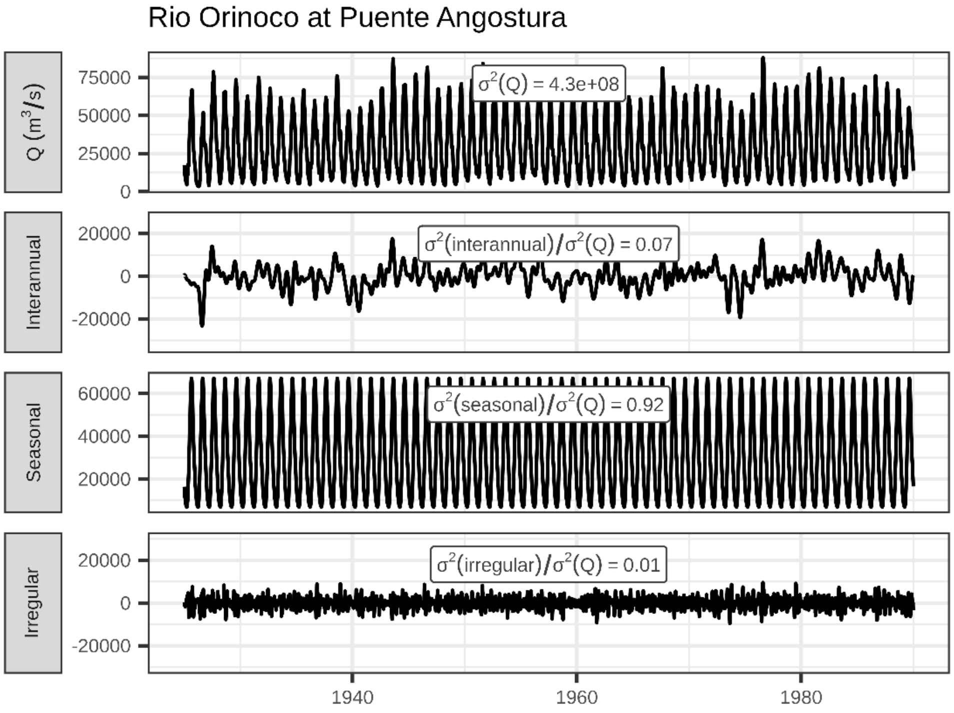

Figure S28: Decomposed time series for the Orinoco River at Puente Angostura (Venezuela), (GRDC 3206720), a highly seasonal river.

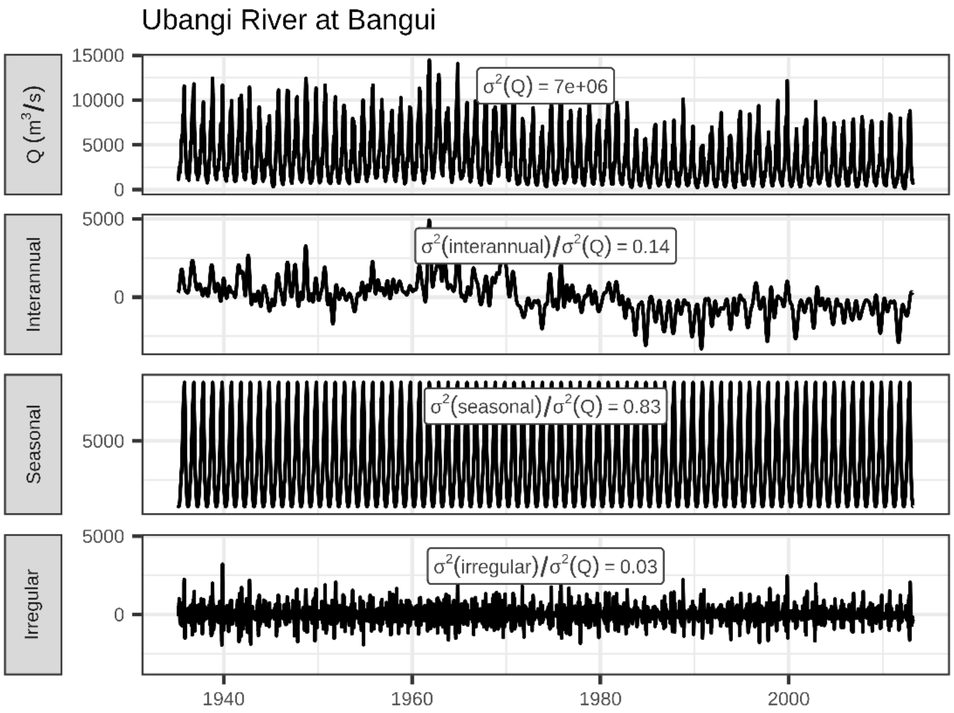

Figure S29: Decomposed time series for the Ubangi River at Bangui (Central African Republic), (GRDC 1749100), a highly seasonal river.

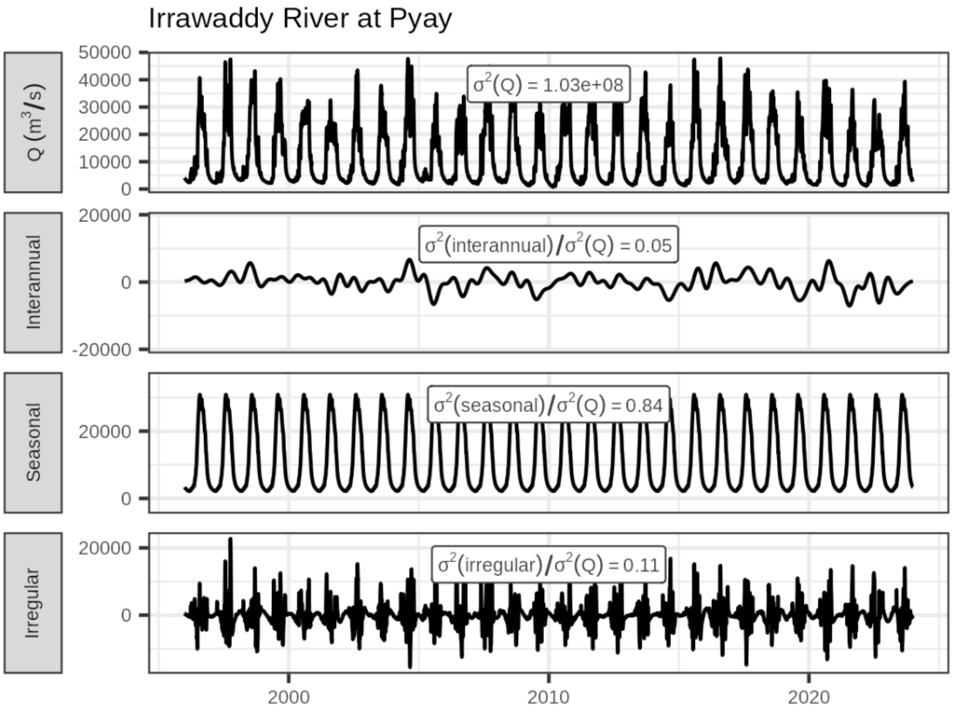

Figure S30: Decomposed time series for the Irrawaddy River at Pyay, Myanmar, (GRDC 2260700), a highly seasonal river.

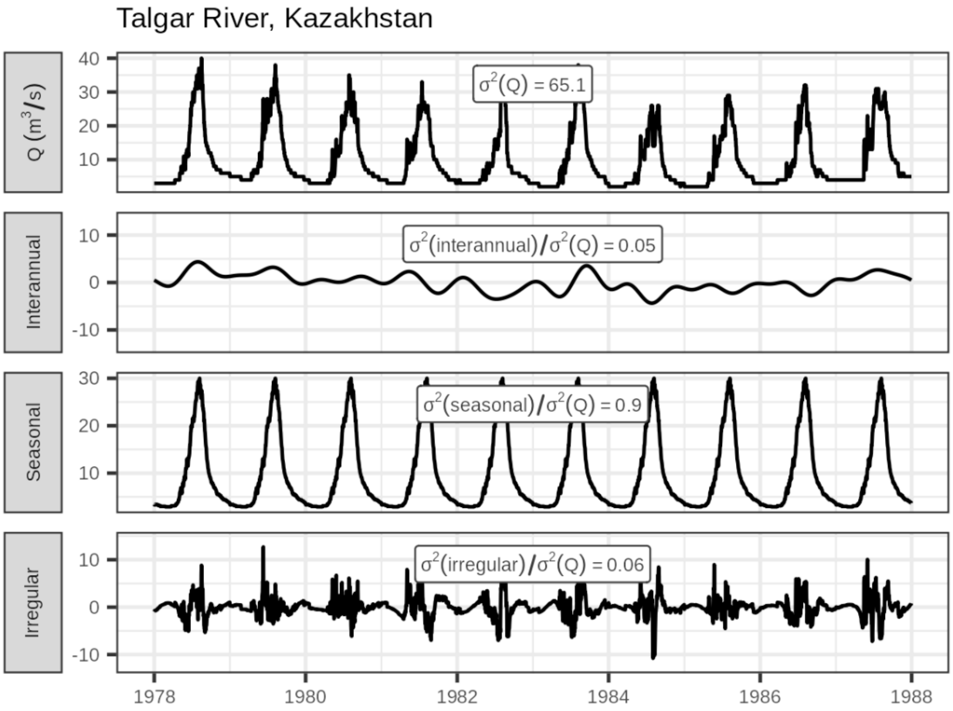

Figure S31: Decomposed time series for the Talgar River at Talgar (Kazakhstan), (GRDC 2314400), a highly seasonal river.

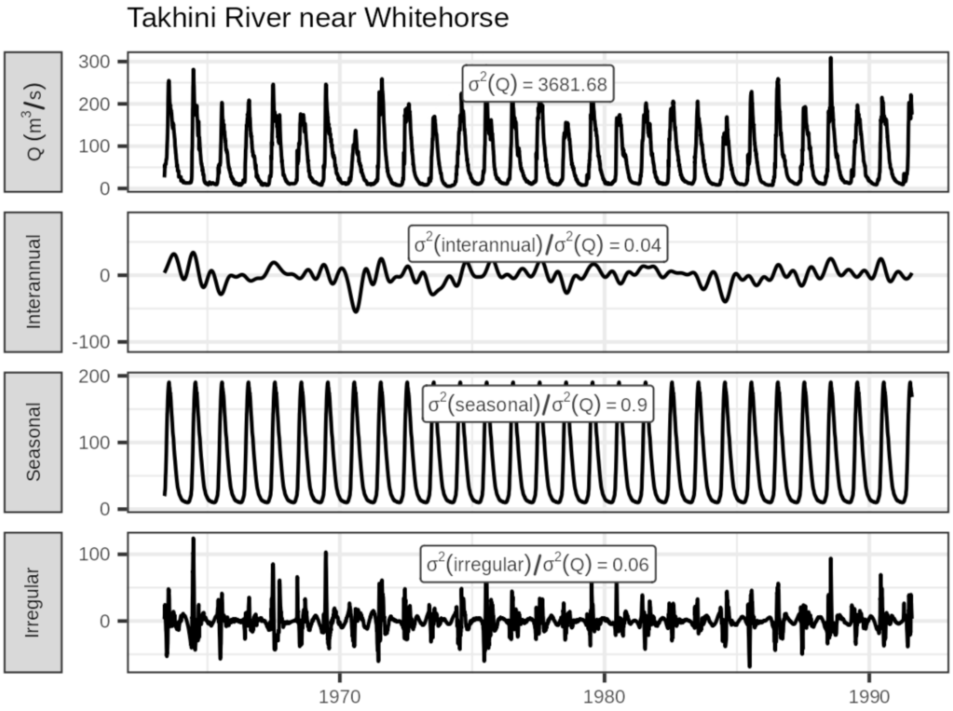

Figure S32: Decomposed time series for the Takhini River near Whitehorse (Yukon, Canada), (Water Survey of Canada 09AC001), a highly seasonal river.

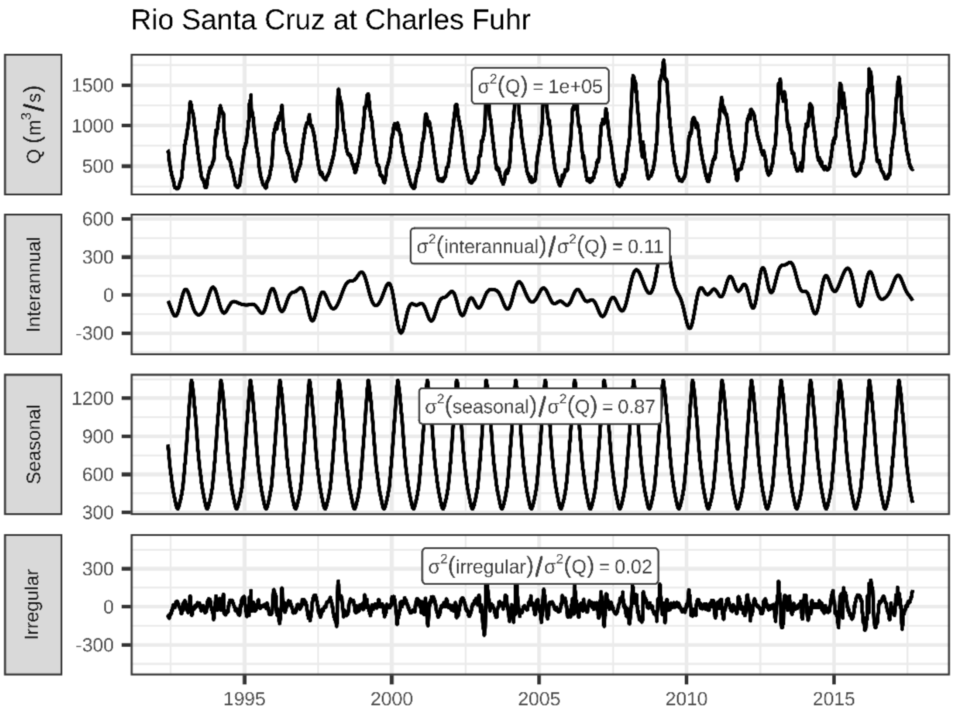

Figure S33: Decomposed time series for the Santa Cruz River at Charles Fuhr station (Santa Cruz, Argentina), (GRDC 3276800), a highly seasonal river.

S7: Climatological NSE_{cb} based on differential split samples

When hydrologic models are intended to be used for climate change projection, a popular technique is the differential split sample, where the dataset is split to maximize the difference in some climate variable between the training and testing periods (Klemeš, 1986). If the model achieves a high NSE when evaluated on a climate that is warmer, colder, wetter, or drier than it was trained on, then it is assumed to be good at extrapolating to a future climate.

We tested for split sample robustness using catchments in Brazil, Switzerland, and North America. For all catchments with at least 20 years of data, we split the years into warm/cold and wet/dry differential split samples. We used water years beginning October 1st, which is consistent with standard practices in each location (Almagro et al., 2021; Höge et al., 2023; United States Geologic Survey, 2016).

To determine the warm/cold and wet/dry splits we used ERA5-Land data for Brazil and North America and gridded daily precipitation and temperature products from Meteo-Swiss for Switzerland (Höge et al., 2023). We evaluated the benchmark NSE when 'trained' on one half of each differential split sample and tested on the other half, and averaged the NSE across the two splits.

For comparison, we also randomly split the years into two equal sets, repeated the random split 10 times, and took the median benchmark NSE across the 10 splits.

Figure S34 shows the benchmark NSE for three sample splitting routines: random, a warm/cold differential split, and wet/dry differential split. These benchmark NSE values are shown for three datasets, covering Brazil, Switzerland, and North America. We find that in general, differential splitting of the sample reduces the benchmark NSE, as expected.

However, the reduction in benchmark NSE is smallest for the arctic, alpine, and tropical regions that have the highest benchmark NSEs to begin with. In other words, in these regions it is not necessary to accurately account for interannual climatic variability to achieve a 'high' NSE under a differential split sample. Since changes to temperature and precipitation in these regions over the next century may be much larger than historical climate variability, the NSE is unreliable judge of a model's suitability for climate change projection.

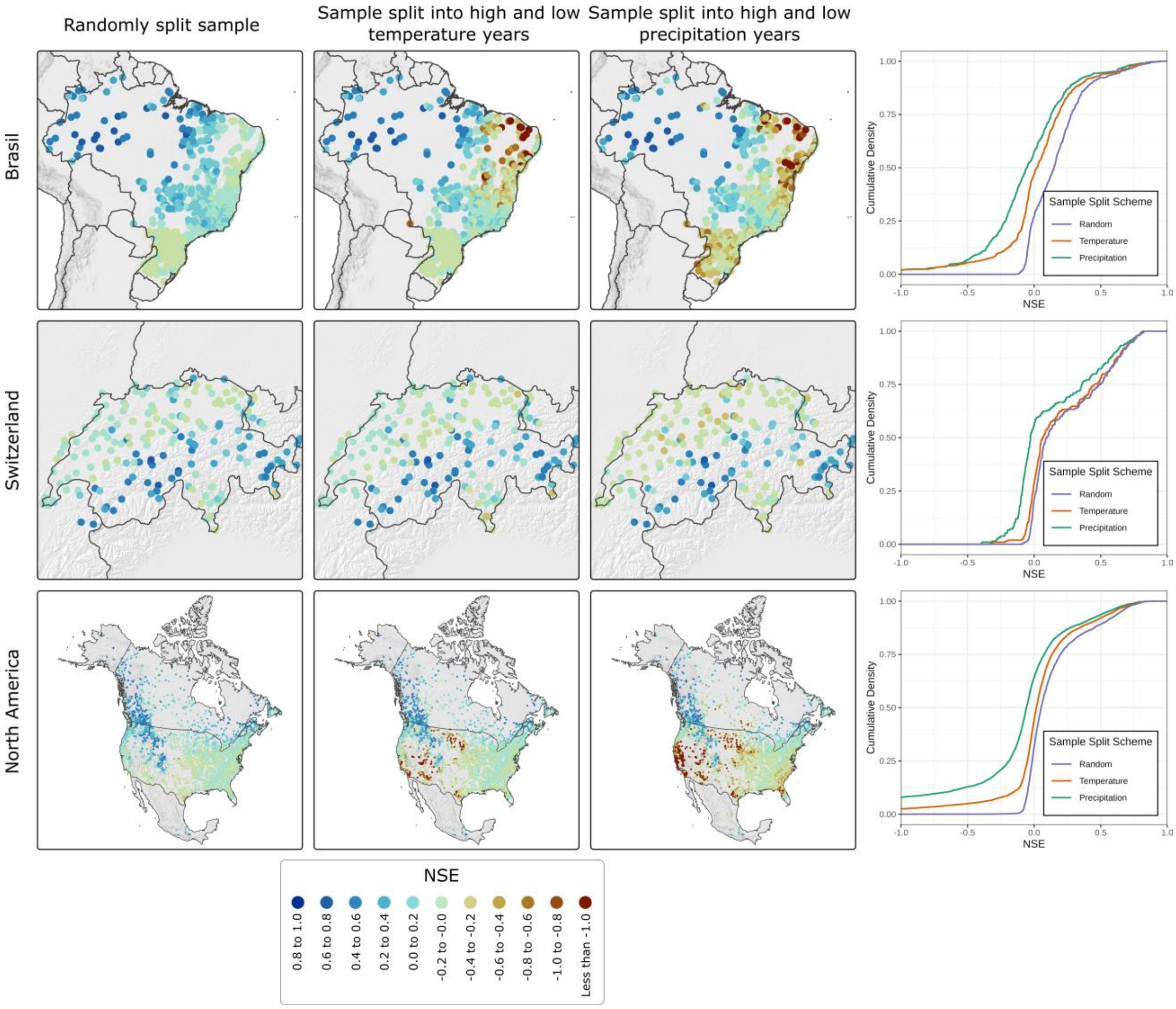

Figure S34: Splitting samples into warm/cold and wet/dry years reduces the performance of the climatological benchmark model, as expected. However, the reduction is smallest for the catchments that have the highest benchmark NSE under a random split.

References

Almagro, A., Oliveira, P. T. S., Meira Neto, A. A., Roy, T., & Troch, P. (2021). CABra: A novel large-sample dataset for Brazilian catchments. *Hydrology and Earth System Sciences*, *25*(6), 3105–3135. https://doi.org/10.5194/hess-25-3105-2021

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., Dijk, A. I. J. M. van, McVicar, T. R., & Adler, R. F. (2019). *MSWEP V2 Global 3-Hourly 0.1° Precipitation: Methodology and Quantitative Assessment*. https://doi.org/10.1175/BAMS-D-17-0138.1

Chagas, V. B. P., Chaffe, P. L. B., Addor, N., Fan, F. M., Fleischmann, A. S., Paiva, R. C. D., & Siqueira, V. A. (2020). CAMELS-BR: Hydrometeorological time series and landscape attributes for 897 catchments in Brazil. *Earth System Science Data*, *12*(3), 2075–2096. https://doi.org/10.5194/essd-12-2075-2020

Chagas, V. B. P., Chaffe, P. L. B., Addor, N., Fan, F. M., Fleischmann, A. S., Paiva, R. C. D., & Siqueira, V. A. (2025). *CAMELS-BR: Hydrometeorological time series and landscape attributes for 897 catchments in Brazil - link to files*. (Version 1.2) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.15025488

Chen, M., & Xie, P. (2008, January 8). *CPC unified gauge-based analysis of global daily precipitation*. Western Pacific Geophysics Meeting, Cairns, Australia.

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302–4315. https://doi.org/10.1002/joc.5086

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., & Michaelsen, J. (2015). The climate hazards infrared precipitation with stations—A new environmental record for monitoring extremes. *Scientific Data*, *2*(1), 150066. https://doi.org/10.1038/sdata.2015.66

Han, J., Liu, Z., Woods, R., McVicar, T. R., Yang, D., Wang, T., Hou, Y., Guo, Y., Li, C., & Yang, Y. (2024). Streamflow seasonality in a snow-dwindling world. *Nature*, *629*(8014), 1075–1081. https://doi.org/10.1038/s41586-024-07299-y

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., & Fenicia, F. (2023). CAMELS-CH: Hydro-meteorological time series and landscape attributes for 331 catchments in hydrologic Switzerland. *Earth System Science Data*, *15*(12), 5755–5784. https://doi.org/10.5194/essd-15-5755-2023

Klemeš, V. (1986). Operational testing of hydrological simulation models. *Hydrological Sciences Journal*, *31*(1), 13–24. https://doi.org/10.1080/02626668609491024

Knoben, W. J. M., Woods, R. A., & Freer, J. E. (2018). A Quantitative Hydrological Climate Classification Evaluated With Independent Streamflow Data. *Water Resources Research*, *54*(7), 5088–5109. https://doi.org/10.1029/2018WR022913

Kratzert, F., Gauch, M., Klotz, D., & Nearing, G. (2024). HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin. *Hydrology and Earth System Sciences*, *28*(17), 4187–4201. https://doi.org/10.5194/hess-28-4187-2024

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

Muñoz Sabater, J. (2019). *ERA5-Land monthly averaged data from 1950 to present* [Dataset]. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). https://doi.org/10.24381/cds.68d2bb30

NOAA Physical Sciences Laboratory. (2025). *CPC Global Unified Temperature*. https://psl.noaa.gov/data/gridded/data.cpc.globaltemp.html

United States Geologic Survey. (2016). *What is a Water Year?* Water Resources of the United States. https://water.usgs.gov/nwc/explain_data.html

Zomer, R. J., Xu, J., & Trabucco, A. (2022). Version 3 of the Global Aridity Index and Potential Evapotranspiration Database. *Scientific Data*, *9*(1), Article 1. https://doi.org/10.1038/s41597-022-01493-1

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