https://doi.org/10.5194/essd-2026-685
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

A benchmark dataset of null-calibrated streamflow spectral-memory diagnostics for five CAMELS-family archives, with open rainfall-runoff model outputs for CAMELS-GB v2

Lijian Ren[1]

[1]School of Civil Engineering, Inner Mongolia University of Technology, Hohhot 010051, China

Correspondence: Lijian Ren (renlijian@imut.edu.cn)

Abstract. This article describes a reusable benchmark dataset, rather than a standalone methods paper, of streamflow spectral-memory diagnostics computed from daily discharge records in five CAMELS-family archives (CAMELS-US, CAMELS-GB v2, CAMELS-BR v1.2, CAMELS-AUS v2 and CAMELS-DK; 3,010 gauges in total). For each gauge the release provides the positive-integral autocorrelation time of discharge anomalies, local spectral-slope curves on the raw frequency axis and on the release-defined memory-normalized coordinate $\mathrm{De} = \tau_{\mathrm{acf}} f$, dispersion statistics with gauge- and region-cluster uncertainty, and matched stochastic reference products. For CAMELS-GB v2 it also provides the outputs of three open rainfall-runoff models and two seasonal references under a common chronological split. The De coordinate and the null-calibration ladder are operational fields and reference products defined for this release from standard signal processing and transparent simulations; they are not presented as previously validated, universal techniques. The archive-level characterization is intentionally bounded: memory scaling reduces cross-catchment spectral-curve dispersion in US, GB and BR (24.0%, 14.7% and 11.8%), is weak and uncertain in AUS (7.7%, region-cluster 95% interval -0.7–27.6%), and shows no reduction in the lowland DK archive (-4.7%, -20.2–11.1%). The matched references quantify how much alignment can arise from serial correlation alone. A CAMELS-GB worked reuse case, based on three independent models and 544 gauges (308 disagreement gauges), does not support spectral distance as a universal model-selection rule (median decision delta $+0.124$, bootstrap 95% CI $+0.085$ to $+0.163$). The release is intended for catchment-scale spectral characterization, frequency-specific model comparison and reproducible benchmark construction; it does not establish a storage mechanism or a cross-archive ranking rule. The curated dataset is available from Zenodo at https://doi.org/10.5281/zenodo.22055294 (Ren, 2026).

1 Introduction

Hydrological models are usually evaluated by how closely simulated and observed hydrographs agree, but aggregate scores such as the Nash–Sutcliffe and Kling–Gupta efficiencies do not locate errors in how streamflow variability is distributed across timescales (Yilmaz et al., 2008; Gupta et al., 2009; Knoben et al., 2019). A model can match the overall score while being too smooth at short timescales, too persistent at long timescales, or changing between spectral regimes at the wrong frequency. Spectral diagnostics can describe this structure, and they are widely applied: daily streamflow spectra show Lorentzian and $1/f^{\alpha}$ forms through multiple routes (Thompson and Katul, 2012; Dolgonosov et al., 2008), and large-sample studies have

https://doi.org/10.5194/essd-2026-685
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

mapped spectral characteristics across thousands of gauges (Tu et al., 2023). Relative to that work, this release contributes a data product with three linked layers that are not available together in the cited public spectral products: per-gauge local-slope curves on raw and memory-normalized axes, matched stochastic reference tables, and a same-split open model-output layer for CAMELS-GB v2. The coordinate $\mathrm{De} = \tau_{\mathrm{acf}} f$ and the null-calibration ladder are defined and implemented here as transparent derived fields and reference products. We do not claim that either is an established, independently validated technique from prior literature. The data-descriptor claim is therefore about the documented collection, provenance, uncertainty, accessibility and reuse of the resulting tables; any future methodological validation can be conducted using those tables and the supplied scripts. The benchmark is designed to (i) compare local spectral-slope curves across catchments with different discharge persistence using each catchment's measured memory as a frequency scale, (ii) report matched stochastic references so that serial-correlation effects can be separated from observed alignment, and (iii) provide every derived number as machine-readable data with its generating script and version.

We present the benchmark as a data descriptor for researchers who need a consistent, inspectable starting point rather than a new universal diagnostic. For each of 3,010 gauges in five CAMELS-family archives, the release contains the output-memory time τ_{acf}, local spectral-slope curves on the raw and De axes, dispersion statistics with gauge- and region-cluster uncertainty, shuffled-memory and matched-AR(1) reference statistics, and (for CAMELS-GB v2) open rainfall-runoff model outputs under a common chronological split. The archive results are deliberately reported as an applicability map: a positive three-archive core (US, GB and BR), a weak and interval-crossing Australian result, and no aggregate reduction in the lowland Danish archive. No precipitation-input response function, storage distribution or tracer transit-time distribution is inferred, and the product is not evidence of a storage mechanism. The model layer is a worked reuse case; its negative decision-test result is retained to prevent downstream users from treating spectral similarity as a standalone adequacy or ranking score.

1.1 Scope, intended users and research questions

The primary users are (i) hydrologists comparing frequency-dependent streamflow variability among catchments, (ii) model developers evaluating whether a rainfall-runoff model reproduces observed variability at particular frequency bands, and (iii) researchers assembling reproducible multi-archive benchmarks. Typical questions are: how does a catchment's discharge memory locate the spectral transition on a common coordinate; which frequency bands differ between observed and simulated discharge after a matched null reference; and whether a model that is preferred by a spectral distance also has acceptable hydrograph and memory-time error. The release is not intended for operational forecasting, causal storage inference, or direct pooling of archives with different inclusion rules. Those boundaries are part of the data product and are restated in the Usage Notes and archive-provenance appendix.

https://doi.org/10.5194/essd-2026-685
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

2 Data and methods

2.1 Input archives

Daily discharge and catchment attributes are taken from five public archives: CAMELS-US (Addor et al., 2017), CAMELS-GB v2 (Coxon et al., 2025), CAMELS-BR v1.2 (Chagas et al., 2025), CAMELS-AUS v2 (Fowler et al., 2024) and CAMELS-DK (Liu et al., 2025). The release applies archive-specific quality gates rather than forcing a common sampling frame: finite anomaly, spectrum and local-slope support with at least 50 gauges per bin for US and GB; at least 20 years, at most 30% missingness and no gap requiring more than 30 days of linear interpolation for BR and AUS; and the public gauged lowland subset with finite windows and a 30-gauge bin minimum for DK. The resulting counts are 673, 664, 893, 560 and 220 gauges, respectively (3,010 in total). Exact thresholds, input versions and the resulting applicability boundaries are listed in Appendix B. Raw third-party files are not redistributed; the dataset contains only derived products, and the input files remain available from the archive providers.

2.2 Processing pipeline

For each gauge: (i) discharge anomalies are obtained by subtracting the climatological day-of-year mean and scaling by the median absolute deviation; (ii) the output-memory time τ_{acf} integrates the positive autocorrelation sequence of the daily anomalies (maximum lag 730 days, accumulated until the first non-positive lag; no significance test is applied, and this choice is documented in the code); (iii) Welch power spectra are estimated at a daily sampling rate with segment length 256 days, constant detrending and density scaling, and local log-log slopes β are computed over a 0.9-decade moving window (minimum eight points); (iv) each β curve is stored on the raw frequency axis and on the release-defined coordinate $\mathrm{De} = \tau_{\mathrm{acf}} f$; (v) dispersion is measured as the weighted cross-catchment variance of station-level median β values across log-spaced bins with archive-specific minima of 50 (US, GB, BR), 40 (AUS) and 30 (DK); (vi) gauge uncertainty uses 500 draws per archive (seed 20260604, resampling unit = gauge) and region-cluster uncertainty uses 1,000 draws per archive (seed 20260608, reproducible geographic clusters). The matched-AR(1) reference uses gauge-specific lag-one estimates, clipped to $[-0.99, 0.99]$, with matched mean, variance and innovation scale; the published ensemble has 100 draws for CAMELS-US and CAMELS-GB v2. Shuffled-memory reference tables use 300 draws with the preregistered seed 62. These are reference products for calibration, not independent validation of a universal method. The model layer follows a common chronological 70/30 split per gauge: RRMPG GR4J (Perrin et al., 2003; Kratzert, 2026) and HBV-Edu (Lindström et al., 1997; Kratzert, 2026) implementations pinned to Git commit `7de78c25acc1c255d2acaf739d65e9ce7bbd60c3` and an in-project global long short-term memory network (Kratzert et al., 2019) are calibrated or trained with fixed budgets, and seasonal-climatology and seasonal AR(1) references are computed on the same split; per-gauge evaluation metrics (NSE, KGE, β-curve distances, output-memory-time error) are stored for the held-out period. The full pipeline is provided as scripts (Sect. 4) and every derived table records its generating script and version.

https://doi.org/10.5194/essd-2026-685
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

2.3 Data-product design and method status

The release separates standard computational ingredients from release-specific fields. Day-of-year anomaly scaling, Welch spectra, autocorrelation estimates, bootstrap resampling and the open rainfall-runoff implementations are conventional or citable components. The memory-normalized coordinate and the collection of matched reference families are assembled here to make the machine-readable benchmark interpretable. They should be cited as properties of this release, not treated as independently validated hydrological laws. The validation outputs below are included so that users can audit sensitivity and reuse the data; they do not change the primary status of the article as a descriptor of the released tables.

2.4 Quality control and uncertainty

Quality control follows the same checks as the source archives plus dataset-specific gates: per-gauge curve validity, archive-specific bin population minima, support-matched binning variants, and a 12-configuration Welch/local-slope sensitivity grid. Uncertainties are reported at three levels: gauge bootstrap intervals from 500 gauge draws (US 17.0–29.0%, GB 11.9–20.2%, BR 9.1–15.9%, AUS -0.7–12.7%), region-cluster intervals from 1,000 geographic-cluster draws, and per-gauge residual distributions for the AR(1) reference layer. The product records the sampling frame beside each estimate because archive heterogeneity is scientifically consequential: the positive aggregate effect is concentrated in US, GB and BR, the Australian interval crosses zero, and the lowland Danish archive shows no reduction. The β curves inherit the spectral resolution and gap-handling choices of the Welch estimator; the memory time is a discharge-derived statistic rather than an independent storage measurement; and the model layer covers one national archive with open implementations rather than an operational multi-model ensemble.

2.5 Dataset contents (summary)

The curated product contains: (i) per-gauge memory-time tables for all five archives with attribute correlations; (ii) per-gauge β-curve tables on raw and De axes; (iii) archive-level dispersion tables with gauge- and region-cluster bootstrap intervals; (iv) null-reference tables (shuffled memory, gauge-matched AR(1) ensembles, two-timescale AR families, long-memory families, per-gauge analytical AR(1) residuals); (v) the CAMELS-GB model-evaluation tables; (vi) the decision-test tables and the pre-registration file with timestamps; (vii) figure sources. The full table inventory with generating scripts is given in Appendix A, the archive-specific provenance and applicability table in Appendix B, and the field dictionary in Appendix C.

2.6 Use of artificial-intelligence tools

The author used DeepSeek for code management and checking and for English-language editing to improve grammar and readability. OpenAI Codex assisted with technical consistency checks, corrections to data-access metadata, supplementary numbering and cross-references, and preparation of the data and code release. The author is responsible for the accuracy and integrity of the manuscript, code and dataset and for all scientific interpretations.

https://doi.org/10.5194/essd-2026-685
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

a) Dataset coverage

Archive	Gauges	tau_acf median (d)
CAMELS-US	673	8.7
CAMELS-GB v2	664	9.8
CAMELS-BR v1.2	893	38.2
CAMELS-AUS v2	560	8.6
CAMELS-DK	220	34.6

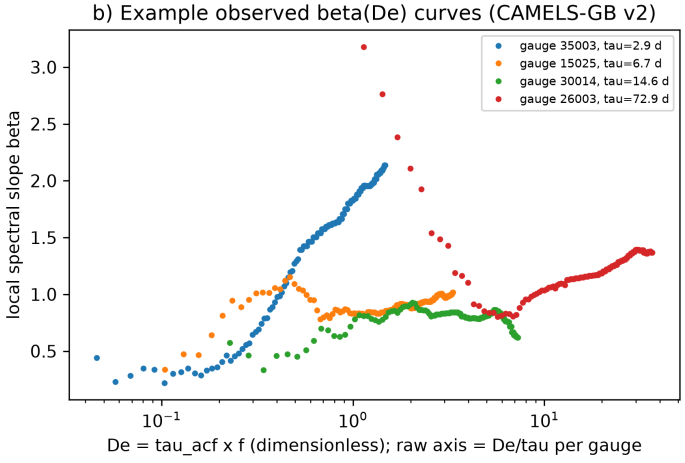


Figure 1. Dataset overview. (a) Coverage of the five CAMELS-family archives: number of gauges passing the archive-specific filters and the median output-memory time τ_{acf}. **(b)** Example observed $\beta(\mathrm{De})$ curves for four CAMELS-GB v2 gauges spanning the memory-time distribution; on the raw frequency axis the same curves are shifted by $\log_{10}\tau_{\mathrm{acf}}$ per gauge.

3 Dataset characterization and worked reuse case

This section gives the minimum characterization needed to use the released tables. Detailed null-family construction, ablations and the full decision-test statistics are retained in the Supplement (Sections S2–S5) so that the main article remains a data descriptor rather than a methods paper.

3.1 Discharge memory varies widely and is associated with archive attributes

Across the 673 CAMELS-US gauges, τ_{acf} spans 0.8–582 days (median 8.7 days), showing that discharge memory is not a fixed archive constant. τ_{acf} is strongly associated with baseflow index (Spearman $\rho = 0.654$, $n = 671$) and weakly associated with drainage area ($\rho = 0.023$); CAMELS-GB v2 shows the same contrast (0.827 and 0.064, $n = 664$), and the baseflow association survives rank-residual controls for area, aridity, snow fraction, elevation, slope and runoff ratio. These associations provide an empirical covariate description for the released tables; they do not identify a storage mechanism. The memory time remains a discharge-derived statistic, which the dataset records explicitly.

3.2 Memory scaling reduces spectral-curve dispersion, with archive boundaries

De reduces the weighted cross-catchment β-curve variance by 24.0% in CAMELS-US, 14.7% in CAMELS-GB v2 and 11.8% in CAMELS-BR v1.2 (region-cluster bootstrap 95% intervals 15.5–34.9%, 5.6–27.3% and 6.2–20.6%), is weakly positive in CAMELS-AUS v2 (7.7%, -0.7–27.6%) and absent in the lowland CAMELS-DK archive (-4.7%, -20.2–11.1%). Randomly permuting the memory assignments gives negative median reductions in all five archives (-21.7%, -14.2%, -10.5%, -37.6%,

https://doi.org/10.5194/essd-2026-685
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

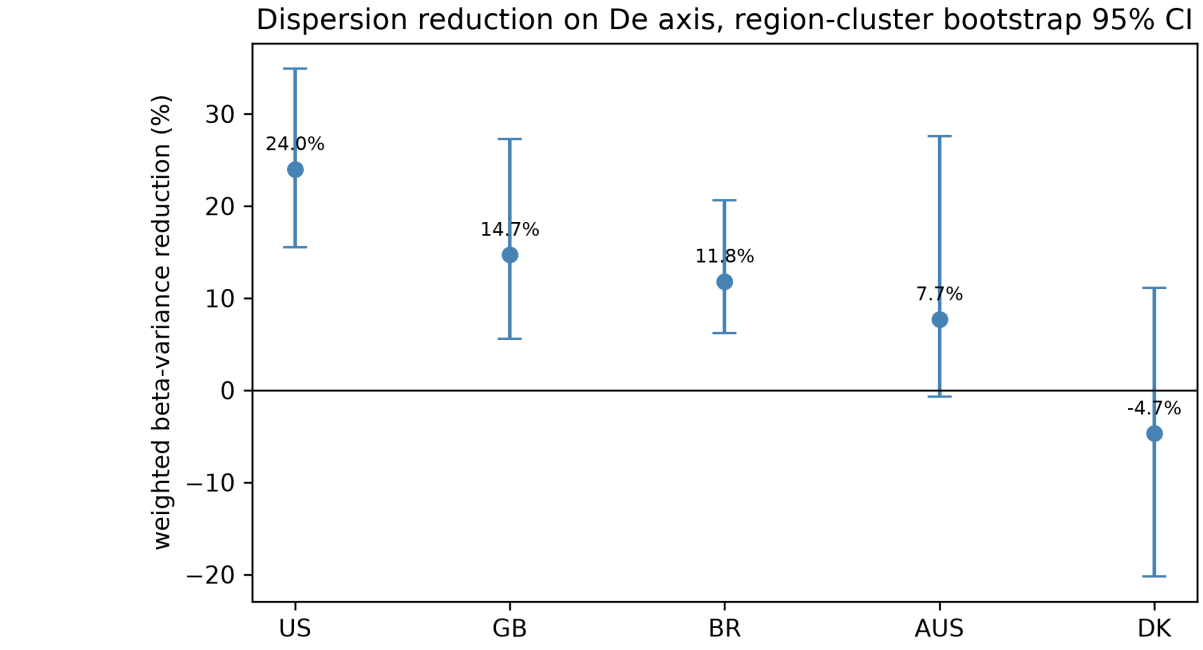


Figure 2. Memory scaling reduces spectral-curve dispersion, with archive boundaries. Weighted cross-catchment β-variance reduction on the De axis relative to the raw frequency axis for the five archives; error bars are region-cluster bootstrap 95% intervals.

-18.5%; empirical probability 0.0033 per archive). Proxy coordinates do not approach the measured-memory effect (CAMELS-US: recession-time 3.8%, drainage-area rank 1.5%, baseflow-index rank -3.7% and constant memory 0.0% dispersion reduction, against 23.7% for the measured memory on the same subset). The dispersion result persists under temporal cross-fitting, across the 12 tested spectral-estimator configurations and under support-matched binning. The dataset reports the heterogeneity as a feature: the positive three-archive core, the weak Australian effect and the negative Danish result define where the coordinate is useful.

3.3 The matched stochastic reference is part of the product

The null-reference tables are supplied beside the observed curves so that users can distinguish data alignment from alignment expected from serial correlation. In the two archives with the published matched-AR(1) ensemble, the weighted reductions are 69.0% (US) and 65.7% (GB), compared with 24.0% and 14.7% for the observed records. Shuffled-memory, two-timescale AR and long-memory reference families are also included, together with per-gauge analytical AR(1) residuals. These outputs are calibration comparators and sensitivity material; they do not establish a mechanistic ceiling or a generally valid diagnostic. Detailed generation settings and the complete numerical comparisons are in the Supplement (Sections S2–S3 and Table S1).

https://doi.org/10.5194/essd-2026-685
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

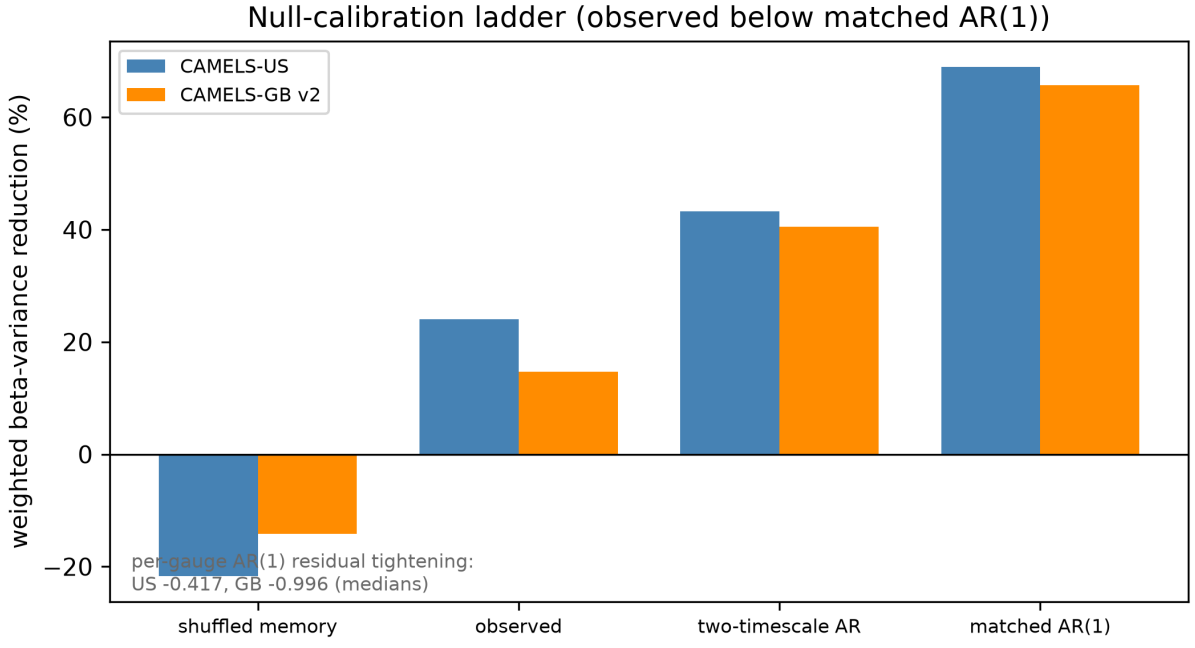


Figure 3. The null-calibration ladder. Weighted β-variance reduction for shuffled-memory assignments, the observed records, analytical two-timescale AR families and gauge-matched AR(1) ensembles in CAMELS-US and CAMELS-GB v2. The observed reduction sits below the matched AR(1) reference, which supplies a reference level for this release; it is not a mechanistic ceiling. The inset notes the per-gauge analytical AR(1) residual medians.

3.4 A worked model-comparison case bounds reuse

The CAMELS-GB model layer illustrates how the data can be reused without turning the benchmark into a ranking rule. The primary pre-registered test uses the three independent rainfall-runoff models on 544 gauges; the NSE- and $\beta(\mathrm{De})$-preferred models disagree on 308 gauges. On those disagreement gauges, the spectral choice has a higher held-out output-memory error (median decision delta $+0.124$, bootstrap 95% CI $[+0.085, +0.163]$, Wilcoxon $p = 4.2 \times 10^{-13}$), and the spectral choice reduces timescale error for only 0.32 of gauges. The separate 586-row reproduction table includes the seasonal references (354 rows with four model entries and 232 with five); it is not the sample size of the primary three-model test. The negative result is retained as a use boundary: spectral curve similarity should be reported with NSE/KGE and memory-time error, not used as a standalone model-selection criterion. The full ablation grid and rank reproduction details are in the Supplement (Sections S4–S5).

https://doi.org/10.5194/essd-2026-685
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

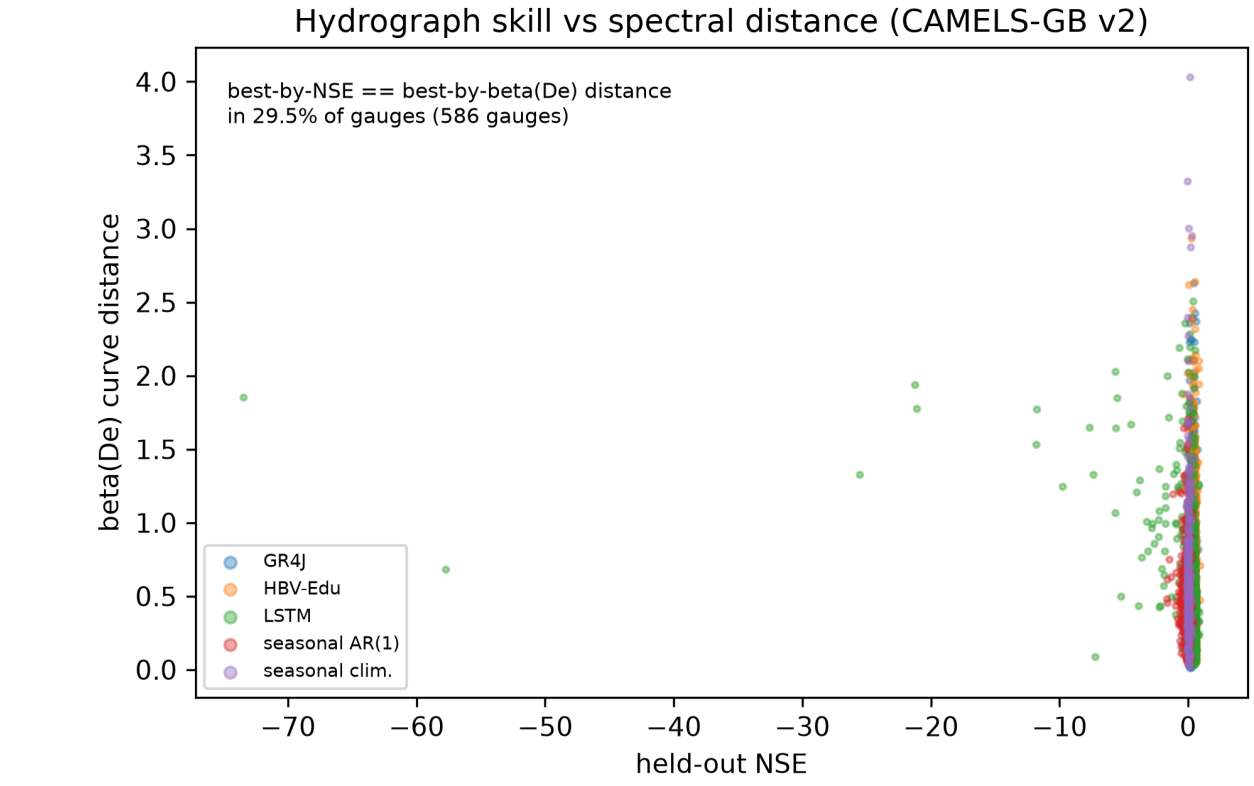


Figure 4. The model layer. Held-out NSE against $\beta(\mathrm{De})$ curve distance for the three independent rainfall-runoff models and the two references in CAMELS-GB v2; the annotation reports the fraction of gauges in which the best model by NSE and the best model by spectral distance agree.

4 Usage notes

The benchmark is intended for catchment-scale diagnosis and model comparison by hydrologists, model developers and researchers building reproducible archive benchmarks. It should be used together with hydrograph skill scores and the archive-specific provenance fields.

A worked CAMELS-GB use case is: (1) select a gauge by `gauge_id` and read its `tau_acf`, raw-axis `frequency_cpd` and De coordinates; (2) apply the same daily anomaly, gap and Welch settings to an observed or simulated series; (3) compare the observed and simulated `beta` curves in the frequency bands that are populated for that gauge; (4) inspect the matched-AR(1) and shuffled-memory rows before interpreting an apparent alignment; and (5) report the curve distance together with `eval_nse`, `eval_kge` and the output-memory-time error. In the model tables, a positive `delta_tau_err` means that the spectral-distance choice has larger timescale error than the NSE choice in the reported decision comparison. This workflow locates frequency-specific mismatch; it does not convert a spectral distance into a universal ranking.

https://doi.org/10.5194/essd-2026-685
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

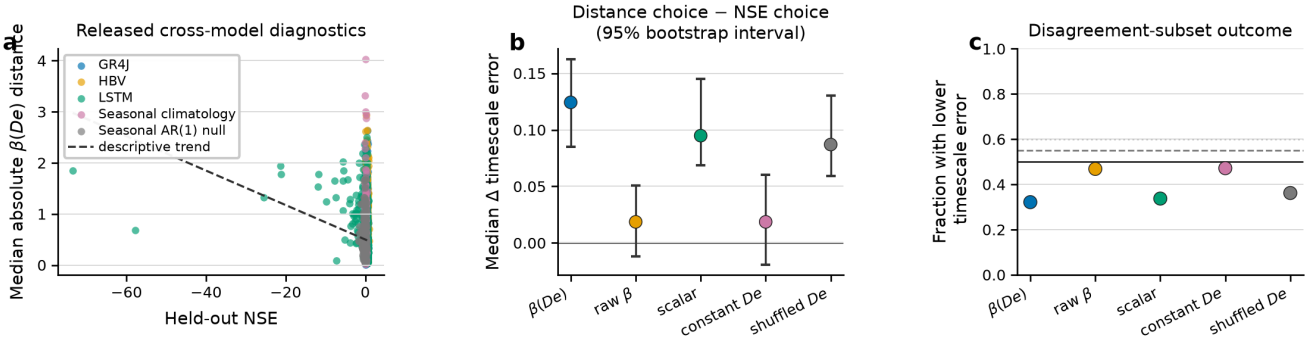


Figure 5. The pre-registered decision test. **(a)** Held-out NSE against $\beta(\mathrm{De})$ distance per model. **(b)** Median decision delta (timescale error of the distance-preferred model minus that of the NSE-preferred model) with bootstrap 95% intervals for the benchmark coordinate and its ablations, on disagreement gauges and the three independent models. **(c)** Fraction of disagreement gauges in which each criterion's preferred model has the lower timescale error, with the pre-registered threshold grid (0.50/0.55/0.60) shown as dashed lines.

For a new catchment or archive, users should first reproduce the archive audit: record length, missingness, gap handling, finite spectral-window support, archive-specific bin minima, input version and gauge/region identifiers. They should then store the derived rows with the same field names and provenance metadata, and keep the new archive separate from the five supplied sampling frames unless a harmonized inclusion rule has been justified. The public scripts are portable to daily discharge series after those checks, but daily sampling alone does not make two archives comparable. The Australian interval crosses zero and the lowland Danish result is negative; these are applicability boundaries, not missing positive findings. The product should not be used for operational forecasting, causal storage inference, or standalone model selection.

Data availability

The benchmark dataset (version 1.0.0) is openly available from Zenodo at https://doi.org/10.5281/zenodo.22055294 (Ren, 2026). The record and files can be accessed without registration or login. Author-created derived data and documentation are provided under CC BY 4.0, while source-specific rights and attribution statements are retained in the record metadata. Raw third-party archive files are excluded from the curated product and remain available from the respective archive providers (Sect. 2.1).

Code availability

The accompanying analysis scripts are available in the Dataset and the companion GitHub repository https://github.com/Johnsonlijian/earth-deborah-benchmark (commit `990843066c7ed05c99ddb35ef29ec163616325f6`), under

https://doi.org/10.5194/essd-2026-685
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

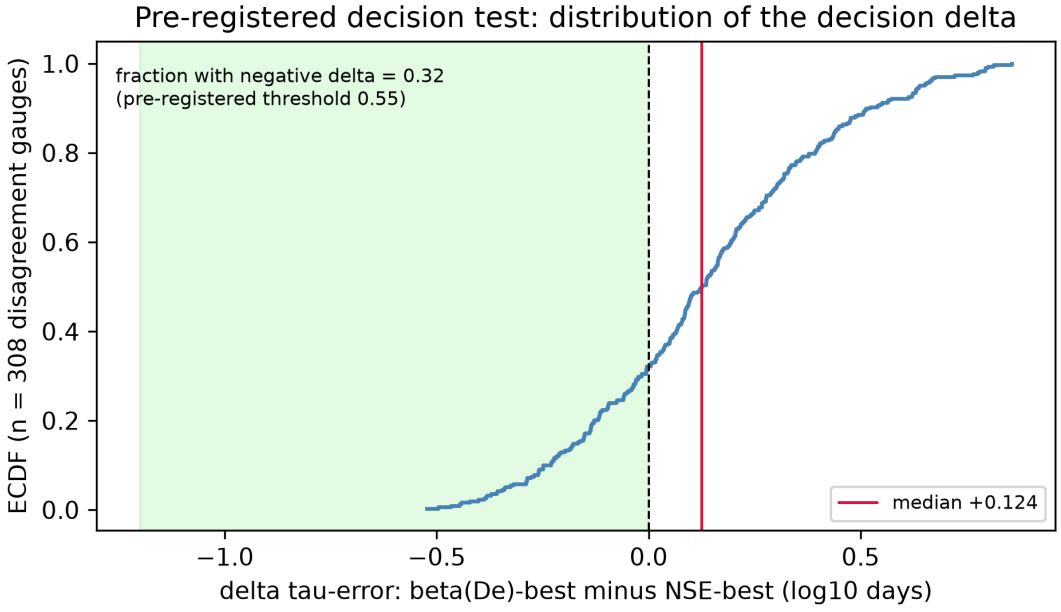


Figure A1. Decision-delta distribution. Empirical cumulative distribution of the decision delta (timescale error of the distance-preferred model minus that of the NSE-preferred model) across the 308 disagreement gauges of the three-model set. The vertical line marks the median (+0.124); the green region marks deltas in which the spectral choice reduces timescale error (fraction 0.32, below the pre-registered threshold of 0.55).

`dataset_release/v1.0.0/code/`. The reproducible runbook specifies the original execution layout, analysis order and expected outputs.

Appendix A: Table inventory (summary)

Grouped by product; each table records its generating script and version in the dataset metadata. (i) memory times: τ_{acf} per gauge, five archives, with attribute correlation tables; (ii) spectral curves: β on raw and De axes, per gauge, five archives; regenerated observed curves for the decision test; (iii) dispersion: archive-level reductions, gauge-bootstrap and region-cluster bootstrap intervals, support-matched variants, 12-configuration estimator grid; (iv) null references: shuffled memory, 100-draw gauge-matched AR(1) ensembles (US, GB), two-timescale AR families, long-memory families, per-gauge analytical AR(1) residuals; (v) model layer: per-gauge metrics, pairwise tests, ranking-disagreement table, per-model β curves (CAMELS-GB v2); (vi) decision test: pre-registration (v1 + addenda), gauge-variant tables, variant summaries, threshold grid, rank-reproduction check; (vii) figure sources.

https://doi.org/10.5194/essd-2026-685
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

Table B1. Archive-specific provenance and applicability boundaries.

Archive	Gauges	Release inclusion gates	Interpretation boundary
CAMELS-US	673	Finite anomaly, PSD and local-slope windows; at least 50 gauges per bin; archive version recorded in the audit table.	Positive aggregate memory-scaling result; use only with the supplied US filter and uncertainty tables.
CAMELS-GB v2	664	Replication filter with short-gap handling, finite windows and at least 50 gauges per bin; model layer uses the common chronological split.	Positive aggregate result and worked model case; do not treat the model comparison as a universal ranking test.
CAMELS-BR v1.2	893	At least 20 years, at most 30% missingness, no gap requiring more than 30 days of linear interpolation, and a 50-gauge bin minimum.	Positive aggregate result under the BR sampling frame; transfer requires a new archive audit.
CAMELS-AUS v2	560	At least 20 years, at most 30% missingness, no gap requiring more than 30 days of interpolation, and a 40-gauge bin minimum.	Weak, interval-crossing aggregate result; treat as uncertain rather than as confirmation of the US/GB pattern.
CAMELS-DK	220	Public gauged lowland subset, common anomaly/interpolation/coverage gates, finite windows and a 30-gauge bin minimum.	No aggregate reduction in this lowland subset; do not extrapolate to all Danish catchments.

Table C1. Core fields, units and definitions.

Field	Unit / format	Definition
gauge_id	string	archive gauge identifier
series	string	"observed" or model identifier
frequency_cpd	cycles/day	frequency (raw axis)
de	dimensionless	$\mathrm{De} = \tau_{\mathrm{acf}} f$
beta	dimensionless	local log-log spectral slope
tau_acf	days	positive-integral autocorrelation time (max lag 730 d)
reduction_pct	%	weighted beta-variance reduction, De vs raw
ci_lo / ci_hi	%	95% interval; procedure and draw count are given in metadata
eval_nse / eval_kge	dimensionless	NSE / KGE on the held-out period
beta_de_median_abs_distance	dimensionless	median absolute curve distance on the De axis
abs_log10_tau_error	log10 days	absolute held-out output-memory-time error
delta_tau_err	log10 days	spectral-choice error minus NSE-choice error

Appendix B: Archive-specific provenance and applicability

Table B1 records the inclusion rules that define the five sampling frames. The rules are intentionally kept beside the derived statistics because the archive result is not transferable without its filter context.

Appendix C: Field dictionary (core tables)

The core tables and their fields are documented in the dataset README and summarized in Table C1. NaN policy: a field is empty when the corresponding filter or estimation step was not applicable for that gauge; no value is imputed.

https://doi.org/10.5194/essd-2026-685
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

Author contributions. LR designed the benchmark, carried out the analysis, drafted the manuscript and prepared the dataset.

Competing interests. The author declares that he has 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, Hydrol. Earth Syst. Sci., 21, 5293–5313, https://doi.org/10.5194/hess-21-5293-2017, 2017.

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 – link to files, Zenodo [data set], https://doi.org/10.5281/zenodo.15025488, 2025.

Coxon, G., Zheng, Y., Barbedo, R., Cooper, H., Fileni, F., Fowler, H., Fry, M., Green, A., Gribbin, T., Harfoot, H., Lewis, E., Qiu, X., Salwey, S., Wendt, D. E., and Ribeiro Neto, G.: Catchment boundaries, daily and sub-daily hydrometeorological time series, groundwater level time series and attributes for 671 catchments in Great Britain (CAMELS-GB v2), NERC EDS Environmental Information Data Centre [data set], https://doi.org/10.5285/9a46d428-958f-4ac1-86eb-94eee70c0955, 2025.

Dolgonosov, B. M., Korchagin, K. A., and Kirpichnikova, N. V.: Modeling of annual oscillations and 1/f-noise of daily river discharges, J. Hydrol., 357, 174–187, https://doi.org/10.1016/j.jhydrol.2008.04.023, 2008.

Fowler, K., Zhang, Z., and Hou, X.: CAMELS-AUS v2: updated hydrometeorological time series and landscape attributes for an enlarged set of catchments in Australia, Zenodo [data set], https://doi.org/10.5281/zenodo.14289037, 2024.

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, J. Hydrol., 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Knoben, W. J. M., Freer, J. E., and Woods, R. A.: Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, Hydrol. Earth Syst. Sci., 23, 4323–4331, https://doi.org/10.5194/hess-23-4323-2019, 2019.

Kratzert, F.: RRMPG: Rainfall-runoff modelling playground, GitHub repository, commit 7de78c25acc1c255d2acaf739d65e9ce7bbd60c3, https://github.com/kratzert/RRMPG/tree/7de78c25acc1c255d2acaf739d65e9ce7bbd60c3, last access: 23 August 2026, 2026.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrol. Earth Syst. Sci., 23, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019, 2019.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S.: Development and test of the distributed HBV-96 hydrological model, J. Hydrol., 201, 272–288, https://doi.org/10.1016/S0022-1694(97)00041-3, 1997.

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 Danish catchments with streamflow observations from 304 gauged stations, Earth Syst. Sci. Data, 17, 1551–1572, https://doi.org/10.5194/essd-17-1551-2025, 2025.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, J. Hydrol., 279, 275–289, https://doi.org/10.1016/S0022-1694(03)00225-7, 2003.

Ren, L.: A benchmark dataset of null-calibrated streamflow spectral-memory diagnostics for five CAMELS-family archives, with open rainfall-runoff model outputs for CAMELS-GB v2, Zenodo [data set], version 1.0.0, https://doi.org/10.5281/zenodo.22055294, 2026.

https://doi.org/10.5194/essd-2026-685
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.



Thompson, S. E. and Katul, G. G.: Multiple mechanisms generate Lorentzian and $1/f^{\alpha}$ power spectra in daily stream-flow time series, Adv. Water Resour., 37, 94–103, https://doi.org/10.1016/j.advwatres.2011.10.010, 2012.

Tu, T., Comte, L., and Ruhi, A.: The color of environmental noise in river networks, Nat. Commun., 14, 1728, https://doi.org/10.1038/s41467-023-37062-2, 2023.

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resour. Res., 44, W09 417, https://doi.org/10.1029/2007WR006716, 2008.