Purpose. This document preserves the detailed construction and sensitivity information that supports the data descriptor while keeping the main article focused on the contents, provenance and reuse of the released tables. The memory-normalized coordinate and the null-reference families are release-defined derived fields and calibration comparators. The material below does not claim independent validation of a universal hydrological method.

Section S1 Relation to the data descriptor

The main article reports a five-archive benchmark with 3,010 accepted gauges and a CAMELS-GB v2 model layer. This supplement records the numerical settings that a user needs to audit the derived fields: archive-specific support rules, resampling counts and seeds, null-series construction, metric definitions and the distinction between the primary three-model decision test and the larger reproduction table.

Section S2 Null-reference construction

Section S2.1 Matched AR(1) ensemble

For each observed anomaly series, the lag-one coefficient is estimated from consecutive centred observations and clipped to $[-0.99, 0.99]$. The innovation scale is derived from the residual variance, and the simulated recursion is started with the observed mean and standard deviation. The published ensemble contains 100 draws per gauge for CAMELS-US and CAMELS-GB v2, generated with the documented default seed 20260604. It is a gauge-matched stochastic reference, not a fitted physical model and not a claim that AR(1) is the only valid null.

The ensemble-level weighted β-variance reductions are 69.0% (US) and 65.7% (GB), whereas the corresponding observed reductions are 24.0% and 14.7%. Per-gauge analytical AR(1) residual summaries are retained in the release tables. The detailed ensemble summary is `reports/tables/r33_matched_ar1_ensemble_summary.csv`.

Section S2.2 Other reference families

The shuffled-memory tables contain 300 draws per archive and use the pre-registered seed 62. Two-timescale autoregressive and ARFIMA-like long-memory families are simulation comparators that test sensitivity to alternative serial structure. Their purpose is to show how the same data coordinate behaves under specified reference processes; they are not proposed as an exhaustive null library.

Section S3 Uncertainty and archive heterogeneity

Gauge-level uncertainty uses 500 independent resampling draws per archive, with seed 20260604 and gauge as the resampling unit. Region-cluster uncertainty uses 1,000 draws per archive with seed 20260608. The cluster schemes and minimum units are: HUC2 with 18 clusters and a 50-unit minimum for US; 12-cell geographic grids and a 50-unit minimum for GB and BR; a 12-division scheme and a 40-unit minimum for AUS; and a 12-cell lowland grid with a 30-unit minimum for DK. The corresponding summary is given in Table S1; the released table is `reports/tables/r42_hess_extreme_hardening_cluster_bootstrap_summary.csv`.

The positive aggregate result is concentrated in US, GB and BR. The AUS interval crosses zero, and the lowland DK estimate is negative with a cluster interval that also crosses zero. These are applicability boundaries of the released sampling frames rather than evidence that the same effect should appear in every CAMELS-family archive.

Section S4 Worked CAMELS-GB decision test

The primary comparison uses three independent rainfall-runoff models under the same chronological split: RRMPG GR4J, RRMPG HBV-Edu and the in-project global LSTM. It contains 544 gauges.

Table S1: Archive-level reductions and uncertainty summaries used in the main text.

Archive	Observed reduction (%)	Gauge 95% interval (%)	Cluster 95% interval (%)	Gauge draws
US	24.0	17.0–29.0	15.5–34.9	500
GB	14.7	11.9–20.2	5.6–27.3	500
BR	11.8	9.1–15.9	6.2–20.6	500
AUS	7.7	-0.7–12.7	-0.7–27.6	500
DK	-4.7	–	-20.2–11.1	500

For each gauge, the spectral criterion is the median absolute distance between observed and simulated local-slope curves on 18 pooled logarithmic bins; NSE is evaluated on the same held-out period. The primary outcome is

$$\Delta_\tau = |\log_{10}(\tau_{\text{pred}}/\tau_{\text{obs}})|_{\text{spectral choice}} - |\log_{10}(\tau_{\text{pred}}/\tau_{\text{obs}})|_{\text{NSE choice}}.$$

On the 308 gauges for which the choices disagree, the median is +0.124, the 5,000-draw decision bootstrap interval is $[+0.085, +0.163]$, the Wilcoxon test gives $p = 4.2 \times 10^{-13}$, and the spectral choice reduces timescale error for 0.32 of gauges. The pre-registered threshold grid (0.50, 0.55 and 0.60) therefore fails at every level.

The larger 586-row reproduction table includes seasonal references: 354 rows contain four model entries and 232 contain five. It is used to reproduce the published ranking tables (100% agreement), not to redefine the primary three-model sample. The ablations are retained in the release: raw-frequency distance (median +0.019, CI $[-0.012, +0.051]$), a global scalar spectral slope (median +0.095, CI $[+0.069, +0.145]$) and per-series shuffled memory (median +0.087, CI $[+0.059, +0.131]$). These results are a worked reuse case and a caution against standalone spectral model selection.

Section S5 Reproducibility manifest

The model-intercomparison manifest records Python 3.11.15, CPU PyTorch 2.12.0, seed 20260608, four workers, a six-year minimum evaluation period, 128 RRMPG samples, LSTM hidden size 32, sequence length 180, five training epochs, 260 batches and batch size 64. The RRMPG implementations are pinned to commit `7de78c25acc1c255d2acaf739d65e9ce7bbd60c3`. The primary summary tables are generated by the scripts listed in the main repository runbook; the release package stores script names and hashes beside each machine-readable output.

Section S6 Use boundaries

The release is suitable for frequency-specific characterization, null-calibrated comparison and reproducible benchmark construction. It is not a causal storage inference product, an operational forecasting tool or a universal model-ranking rule. New archives should be audited for record length, missingness, gap handling, finite spectral support, bin minima, input version and geographic identifiers before their rows are compared with the five supplied sampling frames. The public scripts are portable to daily discharge series after those checks; portability does not remove archive heterogeneity.