



Beyond instantaneous runoff generation: large-sample evaluation of temporal forcing kernels in mGRHyMoLAP

Sianou Ezéckiel Houénafa^{1,2} and Wolfgang Nowak^{3,4}

¹African Institute for Mathematical Sciences, Cape Town, South Africa

²Department of Applied Mathematics, Stellenbosch University, Stellenbosch, South Africa

³Institute for Modelling Hydraulic and Environmental Systems (IWS/LS3), University of Stuttgart, Pfaffenwaldring 5a, 70569 Stuttgart, Germany

⁴Stuttgart Center for Simulation Science (SimTech), University of Stuttgart, Stuttgart, Germany

Correspondence: Wolfgang Nowak (wolfgang.nowak@iws.uni-stuttgart.de)

Abstract. Conceptual rainfall–runoff models retain antecedent information through dynamic states, but differ in whether individual runoff-generation contributions are distributed over subsequent time steps. GRHyMoLAP represents antecedent influence through its production store and recession dynamics, yet applies its additive runoff-generation forcing instantaneously. We introduce mGRHyMoLAP, which replaces this instantaneous application with a normalized temporal convolution while retaining the original storage and recession structure. We represent different temporal convolution shapes by comparing a flexible discrete kernel with parsimonious exponential and log-normal kernels and testing them out of sample across 623 CAMELS-GB catchments against the original GRHyMoLAP and GR4J. The log-normal formulation improves NSE in 93.1% and KGE in 90.7% of catchments relative to GRHyMoLAP and outperforms GR4J in 68.1% and 72.6%, respectively. Beyond aggregate performance, mGRHyMoLAP maintains similar agreement during non-increasing-flow periods while more consistently reproducing the observed direction of daily flow increases. Extending the kernel from current-day forcing only to include one antecedent day captures most of the performance gain, with little further median improvement for longer memory windows. A non-zero kernel peak occurs in 42.9% of catchments, indicating a substantial contribution of antecedent runoff-generation forcing; 96.6% of these catchments peak at a lag of one day. All catchments with a non-zero lag peak improve relative to GRHyMoLAP, with a median NSE increase of 0.067, compared with improvements in 87.9% and a median increase of 0.010 among catchments with a lag-0 peak. Simulated performance is robust across calibration choices, whereas detailed kernel moments and tails are less identifiable, supporting peak lag as the primary diagnostic while cautioning against interpreting the calibrated kernel as a direct estimate of physical travel-time distributions. Non-zero peak lags are associated with larger catchments, longer high-flow durations, lower mean precipitation, higher aridity, and greater cropland fractions. These results demonstrate that temporal redistribution of runoff-generation forcing can improve GRHyMoLAP predictions and provide a useful model-based diagnostic of where such additional temporal flexibility is beneficial.



1 Introduction

Catchment runoff response is shaped by both precipitation inputs and antecedent catchment wetness, with the latter influencing the strength of the hydrological response to precipitation (Knapp et al., 2025). Field observations further show that the dominant controls on runoff generation can shift from antecedent soil water and storage to rainfall characteristics as catchment wetness increases (Farrick and Branfireun, 2014). As water is stored, transported, and released through the catchment, information from previous inputs is therefore carried forward through dynamic states and fluxes, creating temporal dependencies between antecedent conditions and subsequent streamflow. Conceptual models generally represent these dependencies through dynamic storage and state variables that retain the effects of previous inputs. In contrast, data-driven models commonly incorporate antecedent information through sequences of past meteorological inputs and/or hydrological states, allowing temporal dependencies to be learned directly from observations (Kratzert et al., 2018; Hashemi et al., 2022; Houénafa et al., 2025; Xu et al., 2021). These different representations provide alternative ways of accounting for the temporal influence of past inputs on runoff response, which is particularly important in catchments where precipitation does not translate into streamflow instantaneously.

Among conceptual rainfall–runoff models, the widely used Génie Rural à 4 Paramètres Journalier (GR4J) (Perrin et al., 2003) provides a well-established representation of antecedent forcing through both the evolution of its internal stores and the temporal routing of generated water. Precipitation and potential evapotranspiration inputs update the production store, while the resulting fluxes contribute to the routing functions. In addition, GR4J uses unit hydrographs to distribute generated water over successive time steps, thereby introducing a temporal delay between water generation and its contribution to streamflow (Payan et al., 2008; Perrin et al., 2003). Consequently, the influence of antecedent forcing in GR4J arises from both the evolving internal states and the temporal redistribution of generated water through the routing structure. This allows runoff at a given time to reflect the effects of inputs from previous time steps.

GRHyMoLAP (Houénafa et al., 2026b) was recently introduced as a quasi-physical daily rainfall–runoff model that combines the production-store concept of GR4J with the power-law recession formulation of the Hydrological Model based on the Least Action Principle (HyMoLAP) (Afouda and Alamou, 2010). Motivated by the joint role of precipitation inputs and antecedent catchment wetness in shaping hydrological response (Knapp et al., 2025), GRHyMoLAP employs a runoff-generation mechanism in which percolation modulates current effective precipitation within a parsimonious four-parameter ODE formulation. Compared with GR4J and the Hydrologiska Byråns Vattenbalansavdelning (HBV) model (Bergström and Forsman, 1973), its unified formulation represents runoff recession explicitly without requiring a dedicated routing store or unit-hydrograph component, resulting in a streamlined description of runoff-generation and recession dynamics. GRHyMoLAP has demonstrated its effectiveness in large-sample hydrological modeling. Across 222 CAMELS-AUS (Fowler et al., 2021), 671 CAMELS-US (Newman et al., 2015), and 526 CAMELS-FR (Delaigue et al., 2025), GRHyMoLAP showed greater robustness, particularly in low-performing catchments (Houénafa et al., 2026b, a). In CAMELS-AUS, it also performed well under extreme runoff-ratio conditions, mainly due to its enhanced representation of recession dynamics (Houénafa et al., 2026b).



Although the production store of GRHyMoLAP, like that of GR4J, carries information from antecedent inputs through its
 55 state dynamics, the additive runoff-generation forcing is evaluated only from current percolation-modulated effective precipita-
 tion. When current effective precipitation is zero, this generation term vanishes, and simulated discharge evolves solely through
 the existing discharge state and recession dynamics. GRHyMoLAP can therefore represent antecedent influence through stor-
 age and recession, but it cannot distribute the effect of a given runoff-generation forcing contribution over subsequent days.
 This instantaneous-forcing assumption may be restrictive where present streamflow reflects runoff-generation contributions
 60 associated with both current and antecedent effective forcing. We therefore introduce mGRHyMoLAP, which replaces the
 instantaneous application of runoff-generation forcing with a normalized temporal convolution while retaining the production-
 store and recession structure of GRHyMoLAP. We hypothesize that this additional temporal flexibility improves rainfall–runoff
 prediction where the instantaneous allocation of runoff-generation forcing is insufficient.

The objective of this study is to determine whether, and under which catchment conditions, relaxing the instantaneous-
 65 forcing assumption improves daily rainfall–runoff simulation with GRHyMoLAP. We evaluate a fully flexible normalized
 discrete kernel together with parsimonious exponential and log-normal kernels, allowing the predictive value of temporal
 redistribution to be examined across alternative kernel structures. The resulting mGRHyMoLAP formulations are evaluated
 out of sample across 623 CAMELS-GB catchments (Coxon et al., 2020) and compared with the original GRHyMoLAP and
 with GR4J as a widely used benchmark (Kunnath-Poovakka and Eldho, 2019; Sezen and Partal, 2019; Nguyen et al., 2026). We
 70 further assess the robustness of the inferred forcing-memory characteristics to memory-window length and calibration choices.
 Using peak memory lag as the primary diagnostic, we investigate whether catchments with a delayed kernel peak, defined as a
 non-zero peak lag, obtain greater predictive benefits and exhibit distinct catchment characteristics. The study thereby provides
 a targeted extension of GRHyMoLAP and a large-sample test of whether temporal redistribution at the runoff-generation stage
 adds predictive and diagnostic value beyond the model’s existing storage and recession dynamics.

75 The paper is organized as follows. Section 2 presents the methodology, including the GRHyMoLAP and GR4J models, the
 proposed memory formulations mGRHyMoLAP, and the performance metrics. Section 3 describes the dataset used. Section 4
 presents the results, which are discussed in Section 5. Section 6 concludes the paper.

2 Methodology

2.1 GR4J and GRHyMoLAP models

80 The GR4J model, introduced by Perrin et al. (2003), is one of the most widely used lumped conceptual rainfall–runoff models.
 It is governed by four parameters: (i) the production store capacity (X_1), (ii) the intercatchment exchange coefficient (X_2),
 (iii) the routing store capacity (X_3), and (iv) the unit hydrograph time constant (X_4). The model combines a production store,
 a routing store, and two unit hydrographs to represent the transformation and transfer of water from rainfall to catchment
 discharge.

85 GRHyMoLAP is a quasi-physical daily rainfall-runoff model developed by Houénafa et al. (2026b), which combines the
 storage-based formulation of GR4J with the recession-oriented representation of the HyMoLAP. The model is also character-



ized by four parameters: (i) the production store capacity X_1 , (ii) the input scaling parameter γ , which controls the contribution of the percolation-modulated effective precipitation forcing to runoff generation; (iii) the non-linearity parameter μ , which controls the shape of the flow recession; and (iv) the macroscopic parameter λ , which scales the recession dynamics of the system.

90 In both GR4J and GRHyMoLAP models, the production store dynamics follow the same formulation. At a given time t , the net precipitation $P_n(t)$ and the evapotranspiration $E_n(t)$ are first calculated as:

$$P_n(t) = \max(P(t) - E(t), 0),$$

$$E_n(t) = \max(E(t) - P(t), 0).$$

95 The portions of P_n and E_n , denoted by P_s and E_s , respectively, which are directed to the production store, are computed as:

$$P_s(t) = \frac{X_1 \left[1 - \left(\frac{S(t-1)}{X_1} \right)^2 \right] \tanh \left(\frac{P_n(t)}{X_1} \right)}{1 + \frac{S(t-1)}{X_1} \tanh \left(\frac{P_n(t)}{X_1} \right)},$$

$$E_s(t) = \frac{S(t-1) \left[2 - \frac{S(t-1)}{X_1} \right] \tanh \left(\frac{E_n(t)}{X_1} \right)}{1 + \left[1 - \frac{S(t-1)}{X_1} \right] \tanh \left(\frac{E_n(t)}{X_1} \right)},$$

where the production store level at time t is updated through:

$$\hat{S}(t) = S(t-1) + P_s(t) - E_s(t).$$

100 A percolation leakage $Perc$ from the production store is then calculated as a power function of the reservoir content:

$$Perc(t) = \hat{S}(t) \left[1 - \left\{ 1 + \left(\frac{4}{9} \frac{\hat{S}(t)}{X_1} \right)^4 \right\}^{-0.25} \right],$$

while the reservoir content S is obtained by:

$$S(t) = \hat{S}(t) - Perc(t).$$

The above operations remain identical for both GR4J and GRHyMoLAP models. However, in the next step, the GRHy-
105 MoLAP approach dynamically modulates the effective precipitation (P_n) through the percolation flux ($Perc$) according to catchment moisture conditions, and the runoff evolution is governed by an ODE defined as:

$$\frac{dQ(t)}{dt} = -\frac{\mu}{\lambda} Q(t)^{2\mu-1} + \gamma Perc(t) P_n(t). \quad (1)$$

The first term on the right-hand side represents the recession process, whereas the second term governs runoff generation. The
simulated runoff Q , net precipitation P_n , and percolation $Perc$ are expressed in [mm/d], while γ has units of [mm⁻¹], ensuring
110 dimensional consistency of the forcing term $\gamma Perc(t) P_n(t)$.

With a daily time step ($\Delta t = 1$ day) and using an explicit Euler method with forcing term evaluated at the current state, the fully discrete GRHyMoLAP model for simulating daily runoff is given by:

$$Q_t = Q_{t-1} - \frac{\mu}{\lambda} Q_{t-1}^{2\mu-1} + \gamma Perc_t P_{n_t}, \quad (2)$$

where Q_{t-1} denotes the simulated runoff at the previous time step.



115 2.2 mGRHyMoLAP: Extending GRHyMoLAP with temporal memory in runoff-generation forcing

2.2.1 General framework

Runoff responses are distributed over time because water generated by effective hydrological inputs may influence streamflow across different timescales. Consequently, streamflow at time t may reflect the combined influence of current and antecedent runoff-generation forcing.

120 To represent this temporal dependence, we introduce a forcing-memory formulation in which runoff-generation forcing is distributed across lags. Let $G(t)$ denote the runoff-generation forcing. Rather than being represented only at its time of occurrence, its influence on runoff generation is distributed across subsequent lags through the convolution

$$\int_0^t G(t-\tau)w(\tau) d\tau, \quad (3)$$

where $w(\tau)$ is a memory kernel that assigns a temporal weight to runoff-generation forcing occurring at time $t-\tau$ in the representation of runoff generation at time t . These kernel weights represent the relative temporal emphasis assigned by the model to runoff-generation forcing across different lags. The kernel therefore controls the temporal structure of current and antecedent runoff-generation forcing while preserving the general structure of the original rainfall-runoff model.

The proposed framework does not impose a particular functional form on $w(\tau)$. Different kernel representations can therefore be adopted to characterize forcing memory and accommodate different temporal structures of the runoff-generation response. The kernel representations considered in this study are described in the following subsection.

2.2.2 Formulation of mGRHyMoLAP

The effective runoff-generation forcing in the GRHyMoLAP model is represented by

$$G(t) = Perc(t)P_n(t),$$

where $Perc(t)$ denotes the percolation flux and $P_n(t)$ the effective precipitation. The parameter γ scales this forcing in the runoff-generation term of the model.

To facilitate the use of an integral formulation, we assume that the above input term is represented as a continuous-time process. Substituting this expression into the general convolution formulation, Eq. (3), yields

$$\int_0^t Perc(t-\tau)P_n(t-\tau)w(\tau) d\tau.$$

Accordingly, the GRHyMoLAP runoff equation becomes

$$140 \quad \frac{dQ(t)}{dt} = -\frac{\mu}{\lambda}Q(t)^{2\mu-1} + \gamma \int_0^t Perc(t-\tau)P_n(t-\tau)w(\tau) d\tau. \quad (4)$$



Eq. (4) defines the formulated mGRHyMoLAP model, in which runoff generation depends on a weighted history of the runoff-generation forcing, represented by the interaction between percolation and effective precipitation, rather than on a single instantaneous contribution.

In the discrete formulation, let t denote the time-step index. Using the rectangle quadrature rule, the convolution integral is approximated by

$$\int_0^t Perc(t-\tau)P_n(t-\tau)w(\tau)d\tau \approx \Delta t \sum_{k=0}^t Perc_{t-k}P_{n_{t-k}}w_k.$$

For daily simulation, where $\Delta t = 1$ day as in the GRHyMoLAP model, substituting this approximation into the model discrete ODE formulation, Eq. (2), yields

$$Q_t = Q_{t-1} - \frac{\mu}{\lambda} Q_{t-1}^{2\mu-1} + \gamma \sum_{k=0}^t w_k Perc_{t-k} P_{n_{t-k}}. \quad (5)$$

In practice, the memory is truncated to a finite length M , leading to

$$Q_t = Q_{t-1} - \frac{\mu}{\lambda} Q_{t-1}^{2\mu-1} + \gamma \sum_{k=0}^{M-1} w_k Perc_{t-k} P_{n_{t-k}}. \quad (6)$$

For all discrete kernel formulations, the weights are non-negative and normalized over the finite memory window,

$$w_k \geq 0, \quad \sum_{k=0}^{M-1} w_k = 1. \quad (7)$$

This normalization separates the overall magnitude of the runoff-generation contribution, represented by γ , from its relative temporal allocation across lags, represented by the kernel weights. The original GRHyMoLAP formulation is recovered as the special case $w_0 = 1$ and $w_k = 0$ for all $k > 0$.

Figure 1 illustrates the architectures of GR4J, GRHyMoLAP, and mGRHyMoLAP. In panel (1b), the forcing term G_t distinguishes the GRHyMoLAP and mGRHyMoLAP formulations: for mGRHyMoLAP, it represents the weighted history of the runoff-generation forcing, $G_t = \sum_{k=0}^{M-1} w_k Perc_{t-k} P_{n_{t-k}}$, whereas for GRHyMoLAP, it corresponds to the instantaneous forcing, $G_t = Perc_t P_{n_t}$.

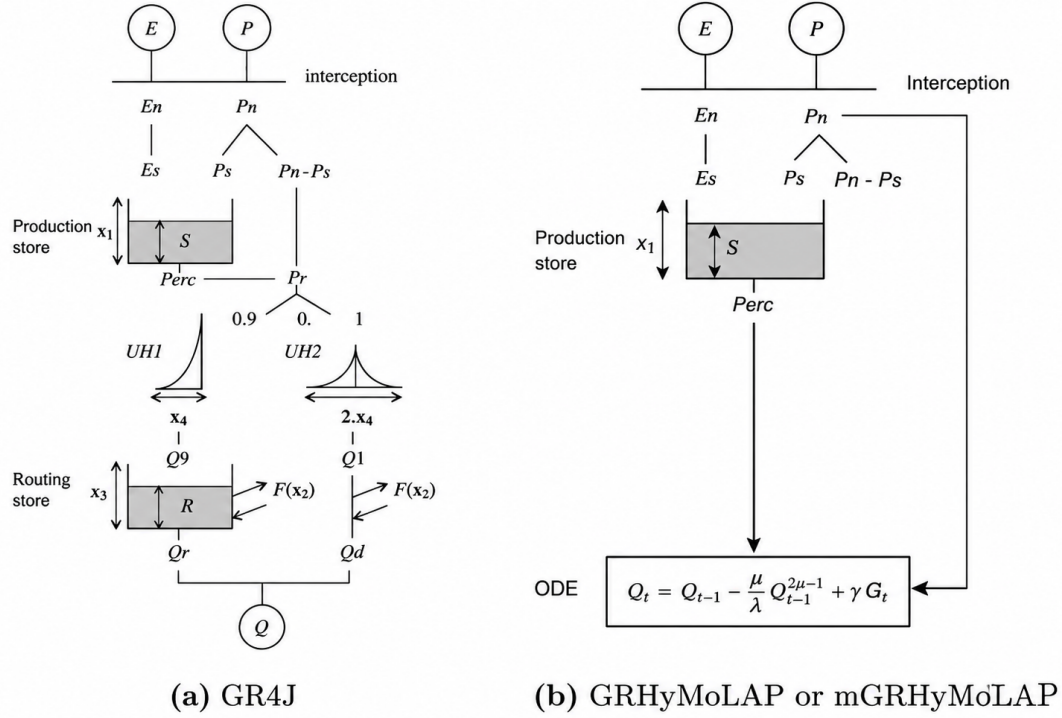


Figure 1. GR4J, GRHyMoLAP and mGRHyMoLAP model architectures.

Without imposing a specific temporal structure on the kernel, we start with a free-form representation by assigning a fixed weight to each lag.

2.2.3 Explicit discrete kernel

A first and simple assumption is to represent the kernel through one independently calibrated non-negative coefficient for each lag. Let $a_k \geq 0$ denote these lag-specific shape parameters and let $A = \sum_{j=0}^{M-1} a_j > 0$. The normalized explicit kernel is then defined as

$$w_k = \frac{a_k}{A} = \frac{a_k}{\sum_{j=0}^{M-1} a_j}, \quad k = 0, 1, \dots, M-1. \quad (8)$$

The coefficients determine the flexible shape of the kernel, whereas normalization ensures that their common scale is not confounded with the runoff-generation scaling parameter γ .

This formulation does not impose any functional structure on the memory decay, and the kernel is directly represented as a finite-dimensional parameter vector. It therefore provides a flexible baseline representation of forcing memory without assuming a specific parametric distribution. As illustration, Figure 2 shows the estimated forcing-memory kernel trajectories for the first two and last two catchments in the database.

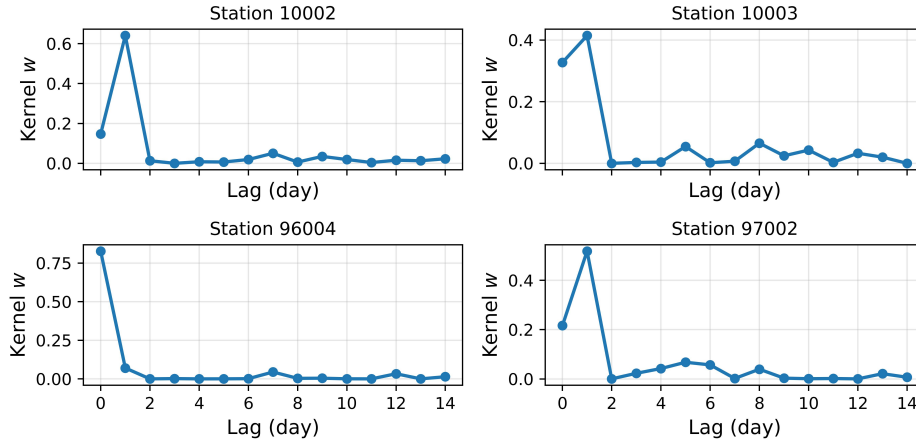


Figure 2. Estimated explicit discrete kernel for the first two and last two gauging stations in the CAMELS-GB dataset. The kernel describes how the influence of past runoff-generation forcing is distributed across different time lags.

However, although this approach is straightforward to implement, it presents some limitations. In particular, we highlight two main drawbacks:

- High computational cost: the number of parameters increases linearly with the selected memory length, which can increase calibration complexity and the risk of over-parameterization, especially when applied across a large number of catchments.
- Limited interpretability: since each lag coefficient is independently estimated, the resulting kernel does not explicitly describe an underlying forcing memory process or provide a direct interpretation of how past and current forcing inputs influence current flow.

Nevertheless, Figure 2 shows that the estimated explicit kernels exhibit recognizable patterns resembling classical probability distributions, including peaks at non-zero lags and gradual decay structures. This observation, together with the limitations highlighted above, motivated the exploration of parametric probability-based kernels. The choice of probability distributions was therefore guided by the observed shapes of the explicit discrete kernels, allowing forcing memory to be represented using fewer and more interpretable parameters.

2.2.4 Probability-based kernel

A more interpretable approach consists in modeling $w(\tau)$ as a parametric probability distribution defined over time. In continuous form, the kernel is assumed to satisfy:

$$w(\tau) \geq 0, \quad \int_0^{\infty} w(\tau) d\tau = 1, \quad (9)$$



so that it can be interpreted as a probability distribution of the temporal contributions of effective forcing to current runoff generation. Guided by the patterns observed in the explicit discrete kernels shown in Figure 2, two probability distributions were selected in this study: the exponential and log-normal distributions. These kernels provide complementary representations of forcing-memory patterns, ranging from monotonic decay to asymmetric distributions with peaks at non-zero lags and contributions distributed over longer time scales.

Since the mGRHyMoLAP model is implemented in discrete time, the kernel is discretized over a finite memory window M . The continuous density is therefore converted into normalized discrete weights as:

$$w_k = \frac{g(k; \theta)}{\sum_{j=0}^{M-1} g(j; \theta)}, \quad k = 0, 1, \dots, M-1, \quad (10)$$

where $g(\cdot; \theta)$ denotes a kernel generator derived from a classical probability density function (Table 1). Figure 3 illustrates the behavior of the exponential and log-normal kernels.

Table 1. Kernel generators derived from classical probability density functions. Multiplicative normalization constants are omitted because the resulting kernels are subsequently normalized in discrete time to obtain the weights w_k .

Kernel type	Kernel generator	Parameters	Parameter meaning
Exponential	$g(\tau; \theta) = \exp(-\tau/\tau_0)$	$\tau_0 > 0$	Characteristic memory time scale controlling exponential decay
Log-normal	$g(\tau; \theta) = \frac{1}{(1+\tau)\sigma_{\log N}} \exp\left(-\frac{(\ln(1+\tau) - \mu_{\log N})^2}{2\sigma_{\log N}^2}\right)$	$\mu_{\log N} \in \mathbb{R}, \sigma_{\log N} > 0$	$\mu_{\log N}$: log-location parameter controlling peak timing; $\sigma_{\log N}$: log-scale parameter controlling memory spread

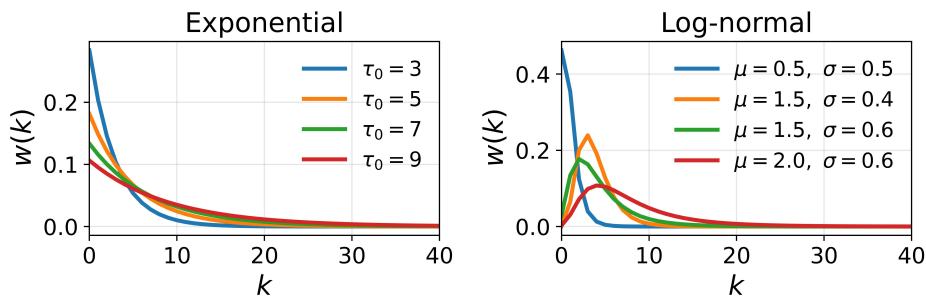


Figure 3. Exponential and log-normal kernels.

To ensure numerical stability and consistency across different kernel formulations, distinct treatments are adopted depending on the functional form. The exponential kernel is naturally well-defined at zero lag and is therefore directly evaluated on the discrete support k , as it does not exhibit any singularity or degeneracy at $k = 0$. In contrast, the log-normal kernel requires special attention at zero lag due to their functional structure in discrete time. The standard continuous formulations become ill-defined at $\tau = 0$, which motivates the use of a shifted discrete support $(k + 1)$ to ensure well-posedness.



Both kernels are normalized in discrete time to obtain weighting functions over the memory window. This guarantees a coherent redistribution of current and past inputs in time, where the relative contribution of recent versus older forcings is governed by the shape of the selected kernel.

2.3 Hydrological interpretation of probability-based kernels

Each kernel provides a representation of forcing-memory structure by assigning temporal weights to current and antecedent runoff-generation forcing.

The exponential kernel exhibits a monotonic decay, representing a forcing-memory structure in which the temporal weight assigned to runoff-generation forcing decreases continuously with increasing lag. This formulation therefore emphasizes recent forcing, with progressively smaller weights assigned to antecedent forcing.

In contrast, the log-normal kernel can allow for asymmetric memory structures with a peak at a non-zero lag. In this case, the maximum temporal weight is not necessarily assigned to the current forcing, as the memory kernel can increase across successive lags before reaching a maximum and subsequently decreasing. This formulation therefore provides more flexibility for representing forcing-memory structures in which the temporal organization of the runoff-generation response can be shifted toward antecedent forcing and distributed over multiple lags. The log-normal kernel can consequently represent both skewed memory structures and non-zero-lag peaks that cannot be represented by a monotonic exponential decay.

Overall, these probability-based kernels provide complementary representations of the temporal structure of forcing memory, ranging from monotonic decay dominated by recent forcing to asymmetric memory structures with greater temporal emphasis on antecedent forcing. This allows the memory formulation to accommodate different temporal organizations of the runoff-generation response across catchments.

2.4 Quantifying forcing-memory characteristics from kernel distributions

To characterize the temporal structure of forcing memory across catchments, two summary metrics were derived from the calibrated log-normal memory kernels: the peak memory lag and the mean memory lag. The peak memory lag is defined as the lag at which the kernel reaches its maximum value:

$$\text{Peak memory lag} = \arg \max_{k \in \{0, \dots, M-1\}} w_k, \quad (11)$$

indicating the lag at which the model assigns the greatest temporal weight to the runoff-generation forcing. Hydrologically, a peak at lag zero indicates that the model assigns its greatest temporal weight to current runoff-generation forcing, whereas a peak at a non-zero lag indicates greater temporal weighting of antecedent runoff-generation forcing.

The mean memory lag was calculated as the first moment of the discrete kernel distribution:

$$\text{Mean memory lag} = \sum_{k=0}^{M-1} k w_k, \quad (12)$$



235 summarizing the average temporal position of the memory distribution within the considered memory window. A larger mean memory lag indicates that the memory structure is distributed further toward antecedent forcing, reflecting greater persistence in the temporal organization of the runoff-generation forcing.

These two complementary metrics capture different aspects of forcing memory: the peak memory lag identifies the lag receiving the greatest temporal weight from the model, whereas the mean memory lag characterizes the overall temporal position of the memory structure. Together, they provide complementary information on how the model distributes temporal weight between current and antecedent runoff-generation forcing.

2.5 Assessment of streamflow change dynamics

Beyond performance metrics, model behaviour was further evaluated under different flow change regimes. Individual time steps were classified according to the sign of observed flow variations, distinguishing non-increasing-flow steps ($\Delta Q_{\text{obs}} \leq 0$) and increasing-flow steps ($\Delta Q_{\text{obs}} > 0$). Model consistency was quantified using the agreement between simulated and observed flow change directions, referred to as sign consistency.

To further investigate model behaviour, a complementary analysis was conducted under no effective precipitation conditions ($P_n(t) = 0$). Under these conditions, time steps were separated into non-increasing-flow and increasing-flow cases to evaluate the ability of models to reproduce flow variations without current effective precipitation input.

2.6 Wilcoxon, Mann–Whitney U, and Cliff’s effect size tests

Statistical comparisons were conducted using non-parametric tests selected according to the structure of the compared samples. For paired comparisons of GRHyMoLAP and mGRHyMoLAP across the same catchments, the Wilcoxon signed-rank test (Wilcoxon, 1945) was used. This test evaluates whether the paired differences are systematically centered around zero by ranking their absolute magnitudes while retaining their signs. Its non-parametric nature makes it appropriate when the distribution of paired differences cannot be assumed to be normal. Statistical significance was evaluated at $\alpha = 0.05$.

For comparisons between independent groups of catchments, the two-sided Mann–Whitney U test was employed. This test assesses whether the distributions of the two groups differ based on the relative ranking of their observations, without requiring normally distributed data. It was used, in particular, to compare catchment characteristics between groups defined according to the timing of peak forcing-memory kernel.

To complement the hypothesis tests, Cliff’s delta (δ) was calculated as a distribution-free measure of effect size. The statistic indicates both the direction and magnitude of the difference between two groups or paired model performances. Following the commonly adopted thresholds (Meissel and Yao, 2024), effect sizes are interpreted as negligible ($|\delta| < 0.15$), small ($0.15 \leq |\delta| < 0.33$), medium ($0.33 \leq |\delta| < 0.47$), and large ($|\delta| \geq 0.47$).



2.7 Model calibration and evaluation

GRHyMoLAP was calibrated following the procedure described in Houénafa et al. (2026b), using the Nelder-Mead simplex algorithm with a multi-start strategy and $1 - \text{NSE}$ as the calibration objective, where NSE denotes the Nash–Sutcliffe efficiency (Nash and Sutcliffe, 1970). GR4J was calibrated independently for each catchment using the same objective function, the L-BFGS-B optimization algorithm, with its four parameters constrained to $X_1 \in [1, 3500]$, $X_2 \in [-15, 15]$, $X_3 \in [1, 1200]$, and $X_4 \in [0.5, 5]$. Twenty randomly generated initial parameter sets were tested for each catchment, and the parameter set yielding the highest NSE over the calibration period was retained. For both models, the first 15 days were excluded from the calculation of metrics to ensure consistency with mGRHyMoLAP.

The six parameters (μ , λ , X_1 , γ , $\mu_{\log N}$, and $\sigma_{\log N}$) of the log-normal mGRHyMoLAP formulation were estimated independently for each catchment using the *Nelder–Mead simplex* algorithm (Nelder and Mead, 1965). The calibration minimized $1 - \text{NSE}$. A fixed memory window M of 15 days was used for all catchments, and the log-normal kernel weights were normalized over this window. The parameter bounds were set to $\mu \in [0.5, 3.5]$, $\lambda \in [10^{-4}, 300]$, $X_1 \in [10^{-4}, 5000]$, and $\gamma \in [0, 20]$, following the same ranges as GRHyMoLAP, while the kernel parameters were constrained to $\mu_{\log N} \in [-20, 20]$ and $\sigma_{\log N} \in [10^{-3}, 10]$. To reduce sensitivity to the initial parameter values, the optimization was repeated from five distinct initial parameter sets, and the parameter set yielding the lowest objective function was retained. Each optimization run was allowed a maximum of 3000 iterations, and the first 15 days, corresponding to the memory window, were excluded from the calculation of performance metrics. The same calibration strategy was also adopted for the explicit-kernel and exponential-kernel formulations of mGRHyMoLAP.

After calibration, model performance was evaluated using the NSE and the Kling-Gupta efficiency (KGE) (Gupta et al., 2009). The NSE can take values from $-\infty$ to 1, where a value of 1 indicates perfect agreement between simulated and observed flow. KGE evaluates model performance by simultaneously considering correlation, variability, and bias. For both measures, values approaching 1 indicate closer correspondence between simulated and observed flows and thus better model performance. They are mathematically expressed as:

$$\text{NSE} = 1 - \frac{\sum_{n=1}^N (Q_n^{\text{sim}} - Q_n^{\text{obs}})^2}{\sum_{t=1}^N (Q_t^{\text{obs}} - \bar{Q}^{\text{obs}})^2}, \quad (13)$$

$$\text{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}, \quad (14)$$

where N is the time series length, r is the Pearson correlation coefficient between simulated and observed flow, $\alpha = \sigma_{\text{sim}}/\sigma_{\text{obs}}$ is the variability ratio, with σ_{sim} and σ_{obs} denoting the standard deviations of simulated and observed flow, respectively, and $\beta = \mu_{\text{sim}}/\mu_{\text{obs}}$ is the bias ratio, with μ_{sim} and μ_{obs} denoting the corresponding mean flow.

2.8 Sensitivity and identifiability analyses

To assess how much temporal support is needed for the predictive benefit of the log-normal kernel, we calibrated the model separately with memory-window lengths $M = 1, 2, 5, 15, 30, 50$, and 100 days. The original GRHyMoLAP model, which does not convolve runoff-generation forcing, was included as a separate comparator and is denoted by $M = 0$ in Figure 8. The



$M = 1$ kernel assigns all weight to current-day forcing and therefore reproduces the original formulation. For each calibrated window, we summarized validation NSE and KGE, the peak memory lag, the mean memory lag, and the kernel spread. Kernel spread was calculated as the weighted standard deviation of lag,

$$s_w = \left[\sum_{k=0}^{M-1} w_k (k - \bar{k}_w)^2 \right]^{1/2}, \quad \bar{k}_w = \sum_{k=0}^{M-1} k w_k. \quad (15)$$

300 We used the $M = 15$ calibration employed in the main analysis as the reference when comparing catchment-wise peak-lag classifications across window lengths.

To assess sensitivity to calibration choices, we repeated the $M = 15$ log-normal calibration with expanded bounds for all six parameters and 15 rather than five initial parameter sets. The objective function, maximum number of iterations, calibration period, and validation period were unchanged. The bounds and numbers of initial sets are reported in Table A1. In a separate
305 temporal sensitivity analysis, the $M = 15$ model was calibrated using each of three non-overlapping thirds of the original calibration period; all resulting parameter sets were evaluated over the same validation period.

Finally, we compared the calibrated explicit and log-normal formulations catchment by catchment. We defined tail mass as $\sum_{k=2}^{14} w_k$. To examine possible compensation between kernel tails and recession dynamics, we evaluated the recession-loss term

$$310 \quad R(Q) = \frac{\mu}{\lambda} Q^{2\mu-1} \quad (16)$$

for each formulation at the catchment-specific median observed validation discharge, Q_{50} . We related the explicit-minus-log-normal difference in tail mass to $\log[R_{\text{explicit}}(Q_{50})/R_{\text{logN}}(Q_{50})]$ using Spearman rank correlation.

3 Data used

The analysis relies on the CAMELS-GB dataset (Coxon et al., 2020), which contains hydrometeorological observations and
315 catchment descriptors for 671 gauged catchments distributed across Great Britain (Figure 4). The dataset combines daily meteorological forcing, observed streamflow, and a broad collection of catchment attributes, thereby providing a consistent basis for evaluating rainfall-runoff models across catchments with contrasting hydroclimatic and physical characteristics.

Daily precipitation P [mm/d] and potential evapotranspiration E [mm/d] are used as meteorological inputs, while daily runoff Q [mm/d] provides the reference data for model calibration and subsequent performance assessment.



Figure 4. Location of the 671 CAMELS-GB river flow gauging stations across Great Britain, adapted from Mathias et al. (2026).

320 Beyond the time series above, CAMELS-GB includes static descriptors characterizing a wide range of catchment properties. These variables cover climatic conditions, topography, hydrometry, hydrogeology, land cover, soil properties, hydrological behaviour, and human influence. A subset of 20 attributes was considered in this study to examine catchment characteristics associated with the mGRHyMoLAP model behaviour (Table 2).

The entire CAMELS-GB time series cover 1970-2015. However, this study considers a period restricted to 1990-2014, as in
 325 Gohouede et al. (2026). The resulting records were divided chronologically into calibration and validation subsets, with 70% of the available observations assigned to calibration and the remaining 30% used for independent validation. Catchments for which more than 10% of streamflow observations were missing were removed before model analysis. This filtering resulted in 623 catchments retained for the final assessment.



Table 2. Catchment attributes considered in this study.

Category	Attribute	Description
Climate	p_mean	Mean daily precipitation (mm d^{-1})
	pet_mean	Mean daily potential evapotranspiration (mm d^{-1})
	aridity	Aridity index (PET/P)
	p_seasonality	Precipitation seasonality
Topography	area	Catchment area
	dpsbar	Mean drainage path slope
	elev_mean	Mean catchment elevation
Hydrogeology	frac_high_perc	Highly productive aquifers
	frac_low_perc	Low-productivity aquifers
	inter_low_perc	Intergranular low-productivity aquifers
Land cover	crop_perc	Cropland cover
	dwood_perc	Deciduous woodland cover
Soil	clay_perc	Clay content
	tawc	Total available water capacity
	root_depth	Effective rooting depth
	conductivity_cosby	Saturated hydraulic conductivity
Hydrology	runoff_ratio	Runoff ratio
	baseflow_index	Baseflow index
	Q95	95th percentile flow
	high_q_dur	Duration of high-flow events

4 Results

4.1 Performance evaluation of memory-based model formulations

Figure 5 first compares the explicit discrete kernel with the calibrated exponential and log-normal kernels for the four representative catchments presented in Figure 2. The estimated explicit discrete kernels display monotonic decay, peaks at positive lags, and asymmetric memory structures. The exponential kernel represents patterns with weights decreasing from lag zero, whereas the log-normal kernel can also represent asymmetric patterns with a peak at a positive lag. Despite using only one and two kernel parameters, respectively, these formulations capture much of the structure seen in the explicit discrete kernels. The log-normal formulation generally resembles the explicit kernels more closely, particularly where their largest weight occurs at a positive lag.

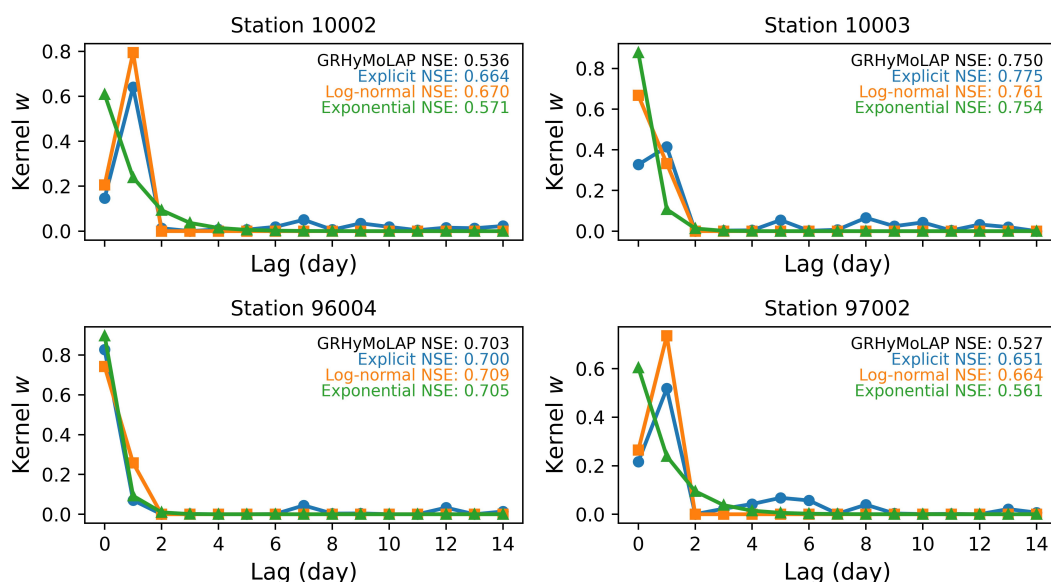


Figure 5. Estimated explicit, exponential, and log-normal forcing memory kernels of the mGRHyMoLAP model for the first two and last two gauging stations in the CAMELS-GB dataset. The NSE values reported in each subplot correspond to the validation performance of the corresponding mGRHyMoLAP formulation for each catchment.

The influence of the proposed memory-driven formulations on predictive performance is summarized in Figure 6. All three memory-driven formulations shift the cumulative NSE and KGE distributions towards higher values than the original GRHyMoLAP model, with the largest shifts for the explicit discrete and log-normal kernels.

These differences are further reflected in the median performance values. For NSE, the median increases from 0.758 for GRHyMoLAP to 0.776, 0.800, and 0.801 for the exponential, explicit discrete, and log-normal mGRHyMoLAP formulations, respectively. A similar trend is observed for KGE, with the median increasing from 0.814 for GRHyMoLAP to 0.823, 0.834, and 0.836, respectively. All three memory-driven formulations also achieve higher median NSE and KGE values than the benchmark GR4J model.

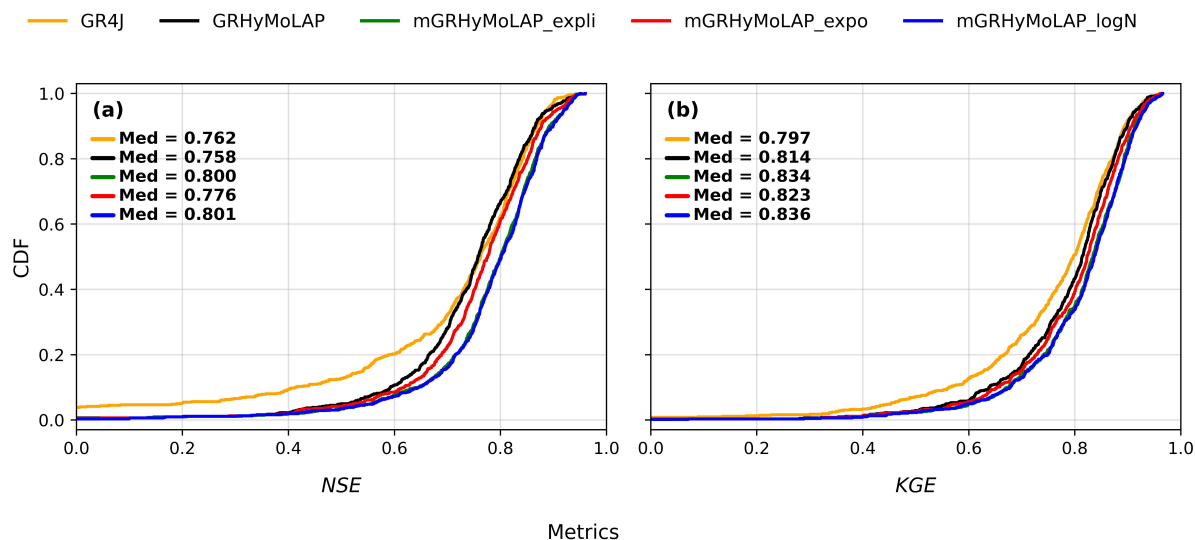


Figure 6. Cumulative distribution functions of validation NSE and KGE obtained with GR4J, GRHyMoLAP, and the three memory-driven mGRHyMoLAP formulations using explicit, exponential, and log-normal forcing memory kernels across the 623 CAMELS-GB catchments.

We assessed whether the performance differences between models were statistically significant using paired, one-sided Wilcoxon signed-rank tests (Table 3). For both NSE and KGE, log-normal mGRHyMoLAP outperformed GR4J, GRHyMoLAP, and the exponential formulation, with $p < 0.001$ in each comparison. It also outperformed the explicit discrete formulation ($p < 0.001$), although the median paired differences were small: +0.0012 for NSE and +0.0006 for KGE. The exponential

350 formulation significantly outperformed GR4J and GRHyMoLAP, but not the explicit discrete formulation.



Table 3. One-sided Wilcoxon signed-rank test comparing the proposed memory-driven mGRHyMoLAP over the validation period. The alternative hypothesis is that the first model listed in each comparison achieves higher performance than the second. Δ denotes the median paired difference (first model – second model). Statistical significance is indicated by *** (p -value < 0.001) and ** (p -value < 0.01).

Metric	Comparison	Δ (median)	p -value
NSE	mGRHyMoLAP_logN vs GR4J	+0.0205	2.51×10^{-39} ***
	mGRHyMoLAP_logN vs GRHyMoLAP	+0.0270	4.39×10^{-97} ***
	mGRHyMoLAP_logN vs mGRHyMoLAP_expli	+0.0012	2.06×10^{-10} ***
	mGRHyMoLAP_logN vs mGRHyMoLAP_expo	+0.0143	3.40×10^{-94} ***
	mGRHyMoLAP_expo vs GR4J	+0.0024	9.64×10^{-5} ***
	mGRHyMoLAP_expo vs GRHyMoLAP	+0.0111	1.78×10^{-95} ***
	mGRHyMoLAP_expo vs mGRHyMoLAP_expli	−0.0138	1.00
KGE	mGRHyMoLAP_logN vs GR4J	+0.0207	6.73×10^{-46} ***
	mGRHyMoLAP_logN vs GRHyMoLAP	+0.0155	3.08×10^{-92} ***
	mGRHyMoLAP_logN vs mGRHyMoLAP_expli	+0.0006	3.65×10^{-4} ***
	mGRHyMoLAP_logN vs mGRHyMoLAP_expo	+0.0087	9.61×10^{-86} ***
	mGRHyMoLAP_expo vs GR4J	+0.0101	8.99×10^{-14} ***
	mGRHyMoLAP_expo vs GRHyMoLAP	+0.0065	3.84×10^{-88} ***
	mGRHyMoLAP_expo vs mGRHyMoLAP_expli	−0.0090	1.00

Figure 7 shows how the validation NSE differences are distributed across catchments. The log-normal formulation has higher NSE than GRHyMoLAP in 93.1% of catchments and than GR4J in 68.1%; for the exponential formulation, the corresponding shares are 92.0% and 51.4%. Compared with the explicit discrete kernel, log-normal mGRHyMoLAP has higher NSE in 60.4% of catchments and exponential mGRHyMoLAP in 19.6%.

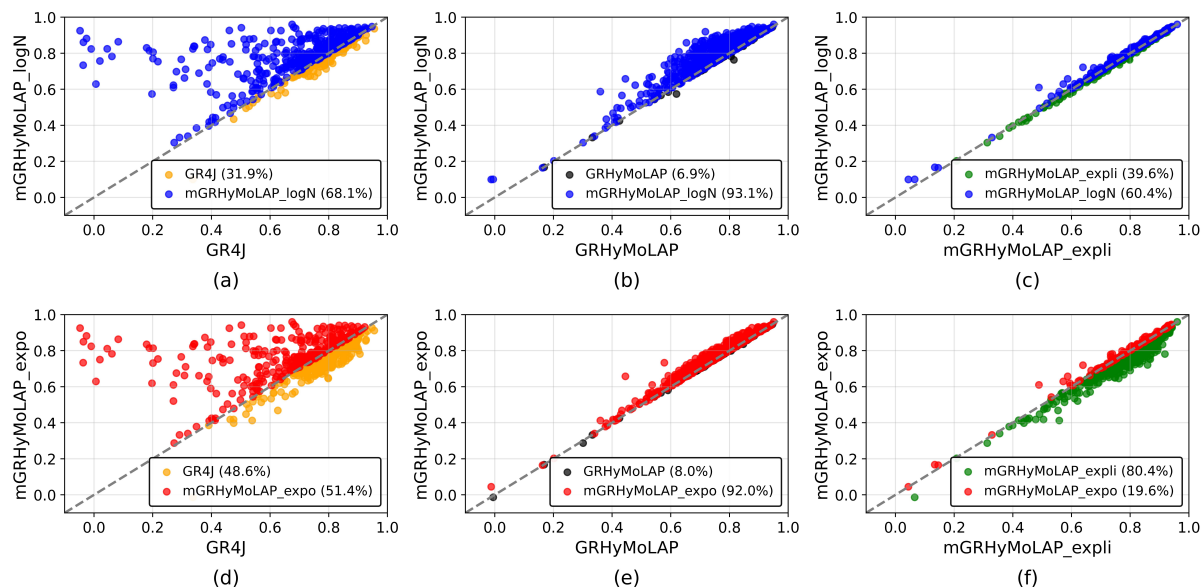


Figure 7. Validation NSE scatter plots for the 623 CAMELS-GB catchments. Points above the 1:1 line indicate catchments for which the model on the y-axis achieves higher NSE values than the model on the x-axis. A few extreme negative NSE values are truncated to improve figure readability.

355 The improvement is also reflected in the distribution of performance classes (Table 4). The proportion of catchments classified as having very good performance ($NSE > 0.80$) increases from 33.2% for GRHyMoLAP to 38.8% for the exponential mGRHyMoLAP, while both the explicit discrete and log-normal kernels raise this proportion to approximately one half of all catchments (50.2% and 50.7%, respectively). Simultaneously, the percentage of unsatisfactory simulations decreases from 5.0% for GRHyMoLAP to around 3.4–4.2% for the memory-driven mGRHyMoLAP formulations. Taken together, these class
 360 shifts show that the improvement extends beyond the median: more catchments achieve very good performance, while fewer remain in the unsatisfactory category.



Table 4. NSE performance classification across the evaluated models during the validation period according to Moriasi et al. (2015): Very good ($NSE > 0.80$), Good ($0.70 < NSE \leq 0.80$), Satisfactory ($0.50 < NSE \leq 0.70$), and Unsatisfactory ($NSE \leq 0.50$). Values indicate the number (and percentage) of catchments belonging to each performance class.

Model	Very good	Good	Satisfactory	Unsatisfactory
GR4J	235 (37.7%)	187 (30.0%)	122 (19.6%)	79 (12.7%)
GRHyMoLAP	207 (33.2%)	242 (38.8%)	143 (23.0%)	31 (5.0%)
mGRHyMoLAP _{expli}	313 (50.2%)	206 (33.1%)	83 (13.3%)	21 (3.4%)
mGRHyMoLAP _{expo}	242 (38.8%)	241 (38.7%)	114 (18.3%)	26 (4.2%)
mGRHyMoLAP _{logN}	316 (50.7%)	207 (33.2%)	78 (12.5%)	22 (3.5%)

How much of the performance gain requires a long memory window? We addressed this question by comparing separately calibrated log-normal formulations with $M = 1$ to 100 days (Figure 8). For visual reference, the figure also shows the original GRHyMoLAP model as $M = 0$; this denotes the model without a convolution rather than an additional kernel calibration. As expected, $M = 1$, which permits current-day forcing only, reproduced the original model's median validation performance. Increasing the window to $M = 2$ raised median NSE from 0.758 to 0.801 and median KGE from 0.814 to 0.836. For $M = 5$ to 100, median NSE remained between 0.801 and 0.803 and median KGE between 0.835 and 0.836. Thus, allowing the previous day's runoff-generation forcing captured most of the predictive benefit; longer admissible windows provided no material additional gain in median performance. We use $M = 15$ as the reference for the following kernel-characteristic analyses because these analyses also examine peak timing and longer tails, rather than median predictive performance alone.

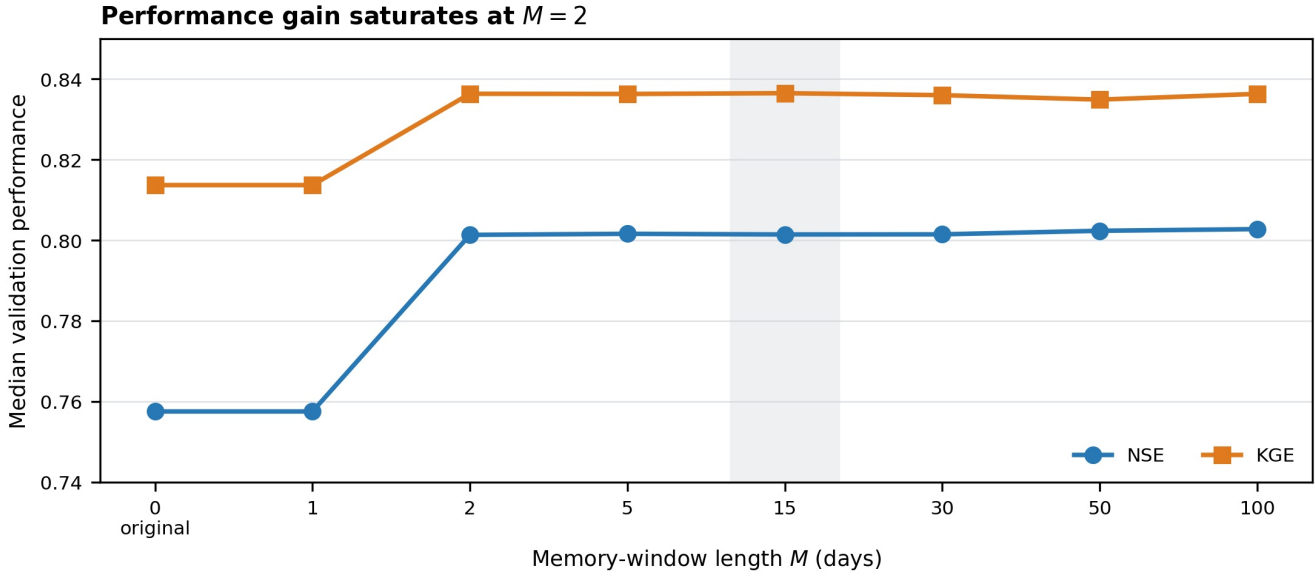


Figure 8. Validation performance across memory-window lengths for 623 CAMELS-GB catchments. Median validation NSE and KGE are shown for the original GRHyMoLAP model ($M = 0$, no convolution) and separately calibrated log-normal mGRHyMoLAP formulations with $M = 1$ to 100 days. Shading marks the $M = 15$ reference used in the main analysis.

4.2 Evaluation of the log-normal mGRHyMoLAP formulation

Given its similar predictive performance compared to the explicit discrete kernel formulation of mGRHyMoLAP, together with its parsimonious parameterization, the log-normal mGRHyMoLAP model was selected for further evaluation against GR4J, the original GRHyMoLAP model, and the exponential kernel formulation.

375 The calibrated parameters of log-normal mGRHyMoLAP remain generally close to those of the original GRHyMoLAP, as shown by their alignment with the 1:1 relationship in Figure 9. This close correspondence is observed for μ , λ , X_1 , and γ , indicating that introducing the probability-based forcing-memory kernel does not substantially alter the parameter structure of the original GRHyMoLAP formulation. Although some catchments exhibit larger deviations, particularly for λ and X_1 , most calibrated parameter pairs remain close to the 1:1 line. Thus, the improved formulation largely preserves the original model
 380 parameterization while introducing an additional representation of antecedent forcing memory.

Figure 10 shows the distributions of the calibrated log-normal mGRHyMoLAP kernel parameters $\mu_{\log N}$ and $\sigma_{\log N}$ across the 623 CAMELS-GB catchments. Most catchments are characterized by relatively small values of both $\mu_{\log N}$ and $\sigma_{\log N}$, although a limited number exhibit substantially larger parameter values. Overall, these results indicate that the forcing-memory kernels are generally concentrated over a relatively narrow range of lags for most catchments, while a smaller number exhibit more
 385 broadly distributed memory structures.

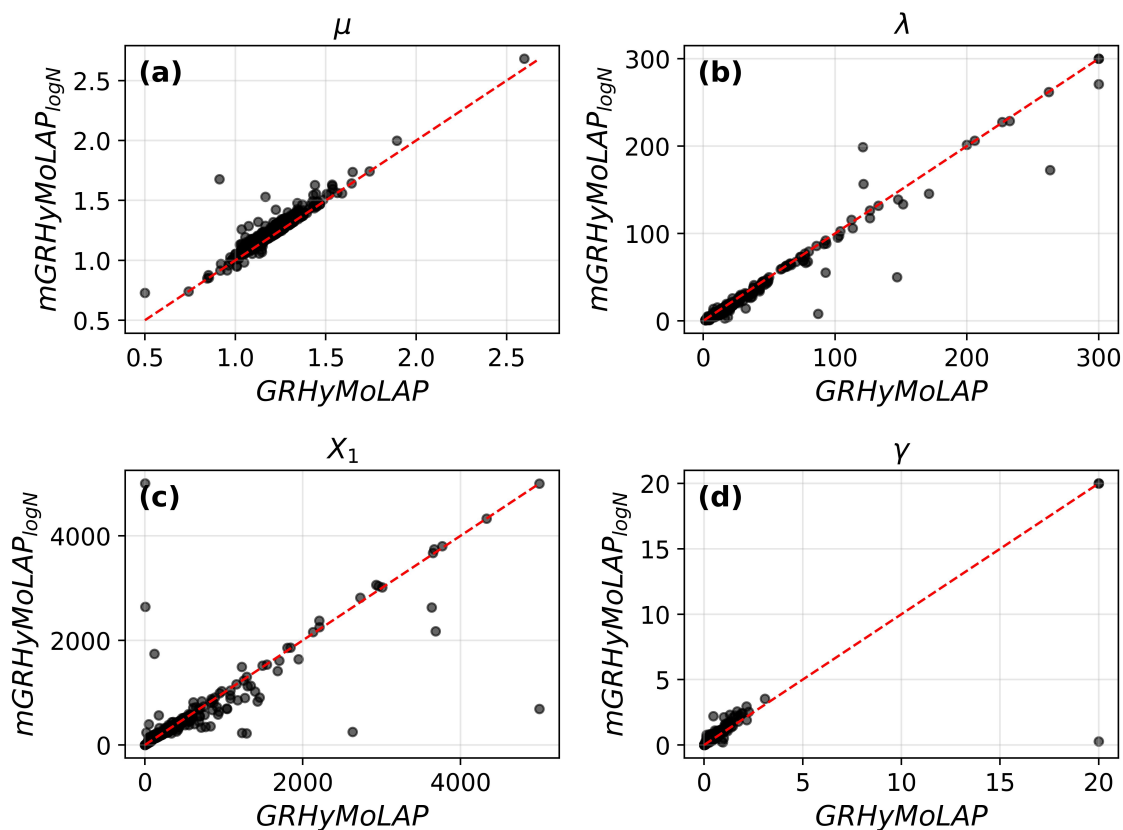


Figure 9. Comparison of the calibrated parameters μ , λ , X_1 , and γ between GRHyMoLAP and the log-normal mGRHyMoLAP across the 623 CAMELS-GB catchments. The dashed red line denotes the 1:1 relationship.

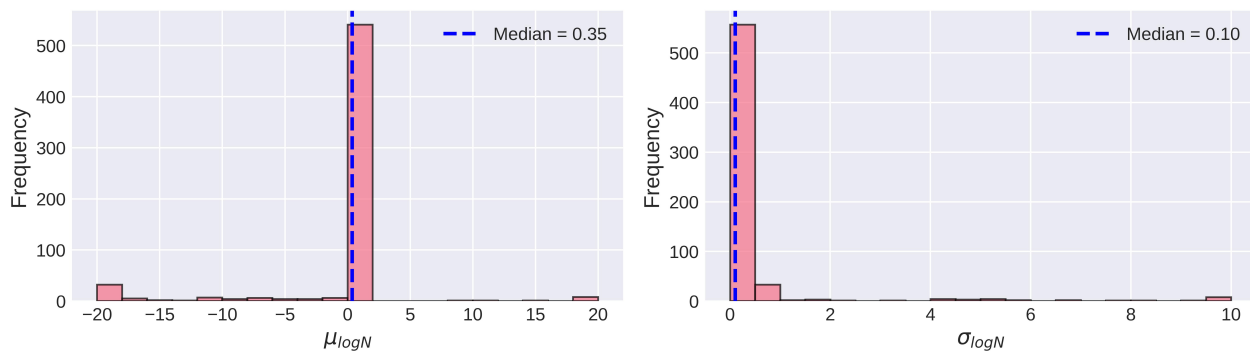


Figure 10. Distributions of the calibrated log-normal kernel parameters $\mu_{\log N}$ and $\sigma_{\log N}$ within mGRHyMoLAP.

Figure 11 illustrates the spatial distribution of the validation NSE differences across the 623 CAMELS-GB catchments. Compared with GR4J (Figure 11a), the original GRHyMoLAP exhibits both performance improvements and deteriorations

throughout Great Britain. In contrast, the log-normal mGRHyMoLAP (Figure 11b) substantially increases both the proportion and spatial extent of positive Δ NSE values, with improvements becoming widespread across England, Wales, and Scotland. Relative to the original GRHyMoLAP (Figure 11c), the log-normal mGRHyMoLAP yields predominantly positive Δ NSE values across the entire study area, while only a limited number of isolated catchments exhibit lower NSE values. Compared with the exponential kernel formulation (Figure 11d), the differences are generally smaller and concentrated around zero, although positive Δ NSE values remain more frequent than negative ones across most regions. The Wilcoxon signed-rank test results reported in Table 3 further confirm that these performance differences are statistically significant in favour of the log-normal mGRHyMoLAP.

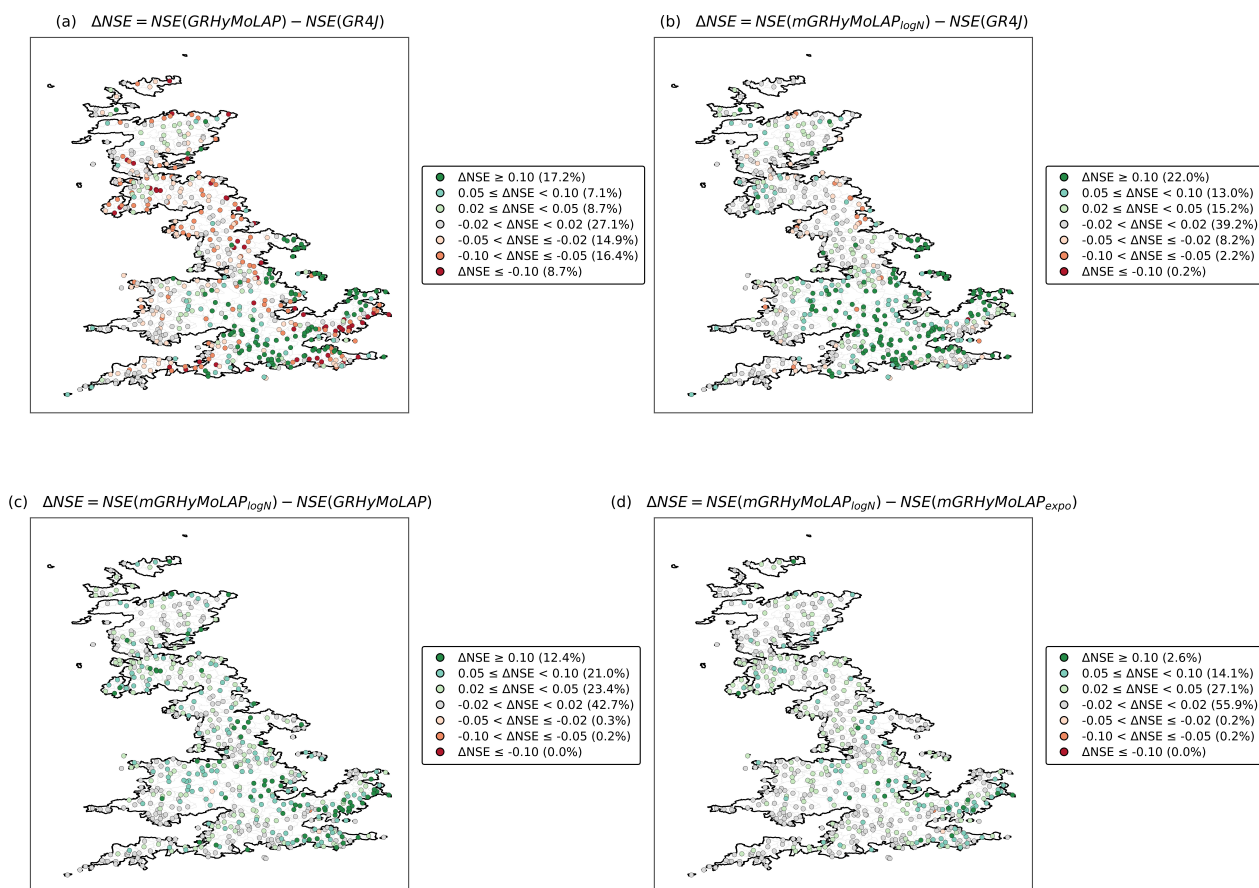


Figure 11. Spatial distribution of validation NSE differences (Δ NSE) across the 623 CAMELS-GB catchments for four model comparisons: (a) GRHyMoLAP versus GR4J, (b) log-normal mGRHyMoLAP versus GR4J, (c) log-normal mGRHyMoLAP versus GRHyMoLAP, and (d) log-normal mGRHyMoLAP versus exponential mGRHyMoLAP.

4.3 Consistency in simulating the direction of flow changes

The sign consistency between observed and simulated daily flow changes across the 623 CAMELS-GB catchments is compared in Figure 12. When all time steps are considered (Figures 12a–b), the three models exhibit similarly high median and mean sign consistency under non-increasing-flow conditions. Although their central tendencies remain comparable, the distributions differ slightly. In particular, the original GRHyMoLAP exhibits a shorter lower tail than both GR4J and the log-normal mGRHyMoLAP, indicating fewer catchments with very low sign consistency. By contrast, more substantial differences are observed under increasing-flow conditions ($\Delta Q_{\text{obs}} > 0$). Here, the mGRHyMoLAP achieves the highest sign consistency agreement. The median sign consistency increases from 0.555 for GRHyMoLAP to 0.722 for mGRHyMoLAP_logN. Moreover, the distribution obtained with the log-normal mGRHyMoLAP is shifted towards higher values, indicating improved agreement with the observed direction of flow increases across a large proportion of catchments.

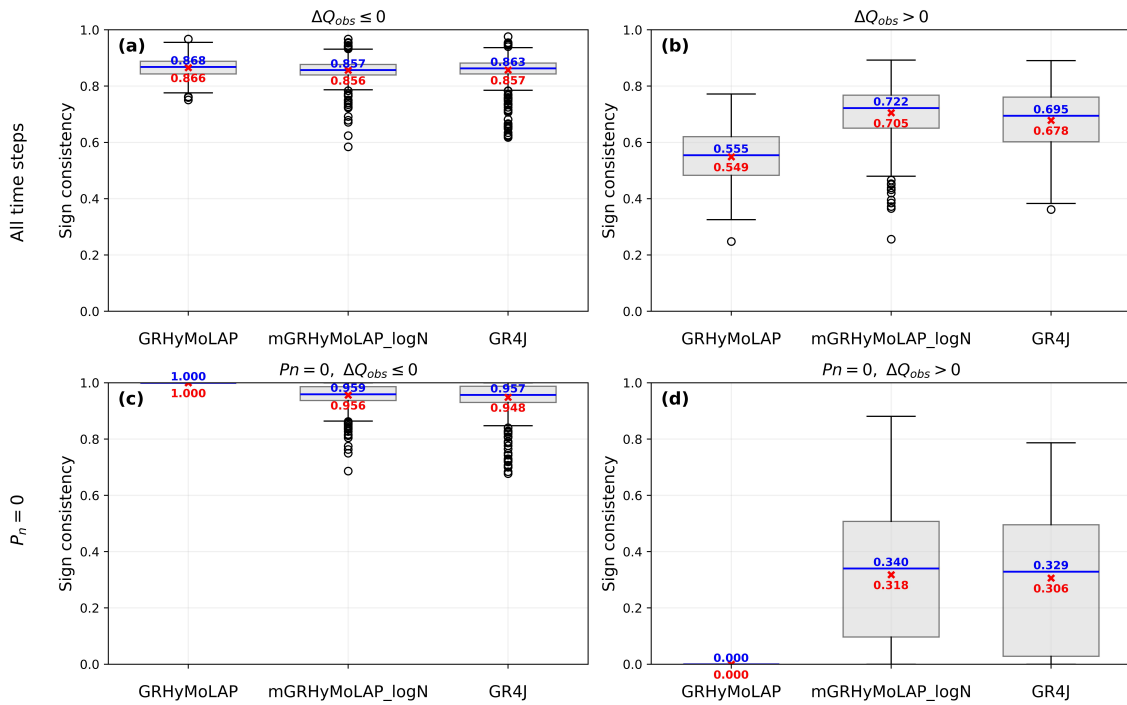


Figure 12. Sign consistency between observed and simulated daily flow changes across the 623 CAMELS-GB catchments. Panels (a) and (b) consider all time steps under non-increasing-flow ($\Delta Q_{\text{obs}} \leq 0$) and increasing-flow ($\Delta Q_{\text{obs}} > 0$) conditions, respectively. Panels (c) and (d) show the corresponding results for time steps without effective precipitation ($P_n(t) = 0$). Blue and red values indicate the median and mean sign consistency across catchments, respectively.

Restricting the analysis to days with no effective precipitation ($P_n(t) = 0$) (Figures 12c–d) further accentuates the differences between the models. Under non-increasing-flow conditions ($\Delta Q_{\text{obs}} \leq 0$), all three models exhibit very high sign consistency, with median values close to one. In particular, the original GRHyMoLAP reaches a sign consistency of one for all catchments,



whereas GR4J and the log-normal mGRHyMoLAP exhibit slightly lower median values together with a broader distribution.

410 Under increasing-flow conditions ($\Delta Q_{\text{obs}} > 0$), however, the behaviour changes markedly. The original GRHyMoLAP exhibits zero sign consistency across all catchments, whereas GR4J and the log-normal mGRHyMoLAP achieve median values of 0.329 and 0.340, respectively. Although these values remain substantially lower than those obtained when all time steps are considered, they indicate that both models reproduce a substantial fraction of observed streamflow increases occurring in the absence of effective precipitation, with the log-normal mGRHyMoLAP exhibiting the highest overall consistency.

415 Figure 13 further illustrates the differences in simulated flow dynamics for catchments spanning a wide range of drainage areas. For the smaller catchment (Figure 13a), the three models produce broadly similar flow-change behaviour, although GR4J achieves a higher NSE. For the intermediate-sized catchment (Figure 13b), log-normal mGRHyMoLAP provides a closer representation of the observed flow variations and achieves the highest NSE among the three models. The differences become more apparent for the largest catchment (Figure 13c), where mGRHyMoLAP more closely follows the timing of

420 observed flow changes across several events, particularly during the higher-flow period toward the end of the validation period. In contrast, GRHyMoLAP tends to provide a smoother representation of these flow variations, with less accurate timing of flow increases. These examples illustrate that the benefit of the memory formulation is not limited to overall flow magnitude but can also extend to the temporal representation of flow changes.

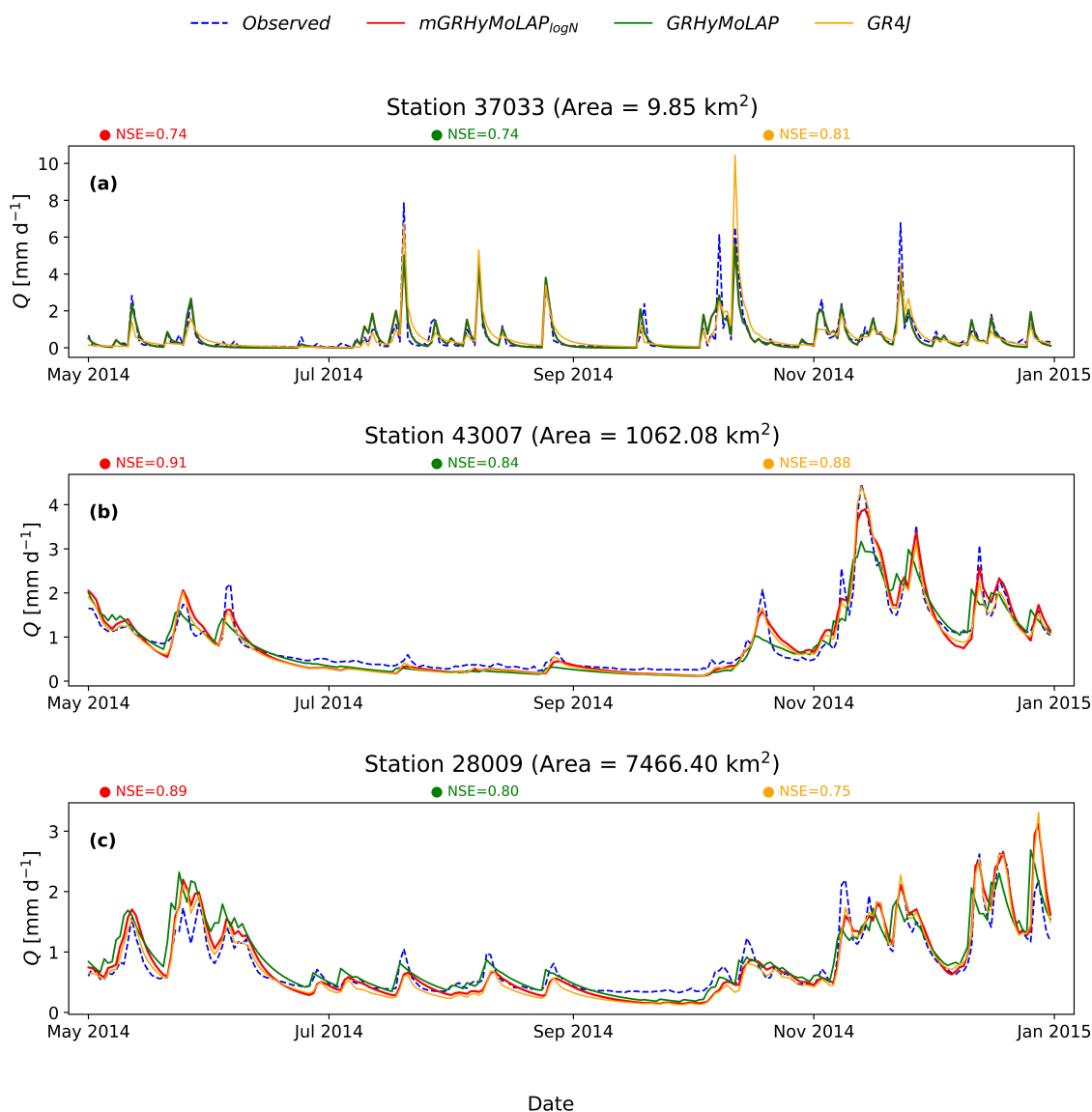


Figure 13. Comparison of observed and simulated daily flow for three catchments with contrasting drainage areas. Panels (a)–(c) show catchments selected across the drainage-area range. Flow simulations from log-normal mGRHyMoLAP, GRHyMoLAP, and GR4J are compared with observations over the last eight months of 2014. The reported NSE values are calculated over the entire validation period.

4.4 Forcing-memory characteristics: peak memory lag and mean memory lag

425 To facilitate the interpretation of the calibrated log-normal kernels, Figure 14 summarizes two derived forcing-memory characteristics: the peak memory lag and the mean memory lag.



The peak memory lag is strongly concentrated at short lags. A substantial majority of catchments exhibit their maximum kernel value at the current day (lag 0), while most of the remaining catchments reach their peak at lag 1 (previous day). Peak memory lags beyond lag 1 are uncommon. Thus, although the calibrated kernels allow for extended temporal dependence, their
 430 dominant memory structure is generally centred on lag 0 or lag 1.

A similar concentration toward short lags is observed for the mean memory lag. The distribution is strongly right-skewed, with most catchments exhibiting mean memory lags below three days, corresponding to mean lags within the first three days of the memory window. The frequency of catchments decreases rapidly as the mean memory lag increases, although a small number of catchments display considerably larger values. These larger values indicate that the memory structure of these
 435 catchments extends further toward antecedent runoff-generation forcing within the considered memory window.

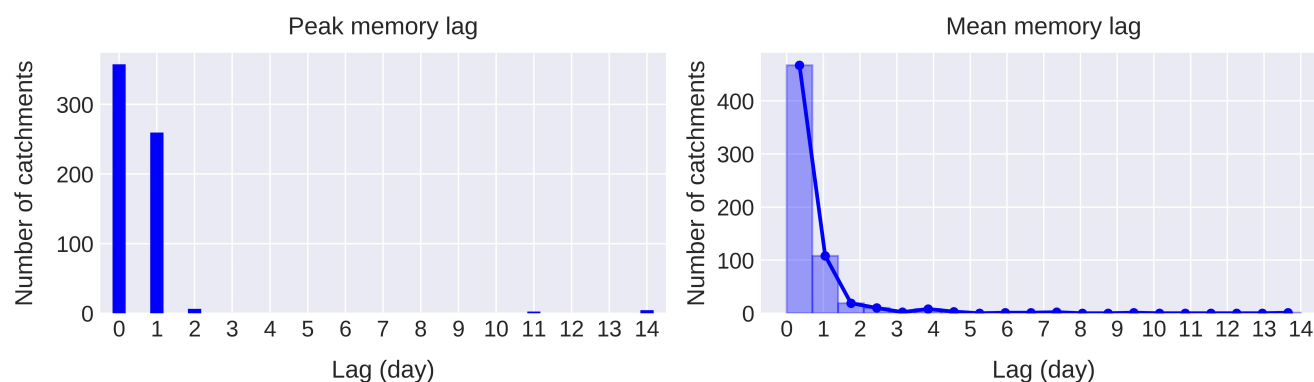


Figure 14. Distributions of two derived forcing-memory characteristics computed from the calibrated kernels: the peak memory lag and the mean memory lag. Lag 0 corresponds to the current day, while positive lags refer to antecedent days.

Although both characteristics are dominated by short lags, they describe complementary aspects of the calibrated memory kernels. The peak memory lag identifies the lag at which the kernel assigns its greatest temporal weight to runoff-generation forcing, whereas the mean memory lag summarizes the overall temporal position of the memory distribution. Together, they show that the forcing-memory structure is predominantly centred on recent forcing across the CAMELS-GB catchments, while
 440 a relatively small subset exhibits memory structures that extend further toward antecedent forcing.

4.5 Robustness and identifiability of forcing-memory characteristics

We assessed whether the inferred timing of the dominant kernel contribution depends on the chosen window. Figure 15a shows, for each M , the proportion of catchments whose largest kernel weight occurs at lag 0, lag 1, or a later lag. Taking the $M = 15$ calibration used throughout the main analysis as the reference, we classified a catchment as either current-day dominated (peak at lag 0) or antecedent-forcing dominated (peak at any positive lag). Two window lengths were counted as agreeing if they
 445 assigned the same catchment to the same one of these two classes. Agreement was necessarily lower for $M = 1$, which cannot represent an antecedent peak. Among formulations permitting antecedent forcing ($M \geq 2$), agreement with $M = 15$ ranged



from 96.6% to 98.1%. Across these windows, approximately 56–58% of catchments peaked at lag 0, 40–44% at lag 1, and no more than 2.1% at longer lags. The dominant peak-timing distinction was therefore largely independent of the admissible window.

This stability does not extend equally to every characteristic of the kernel. Figure 15b contrasts the medians and 90th percentiles of mean lag and kernel spread across window lengths. Their medians changed little as M increased, whereas their upper tails grew markedly for the longest windows. For example, between $M = 15$ and $M = 100$, the median catchment-wise change in mean lag was only 0.003 days, although some catchments acquired substantially longer mean lags and wider kernels. Longer admissible windows thus mainly affected a minority of long-tailed solutions, without changing the dominant peak-lag classification or the median predictive performance.

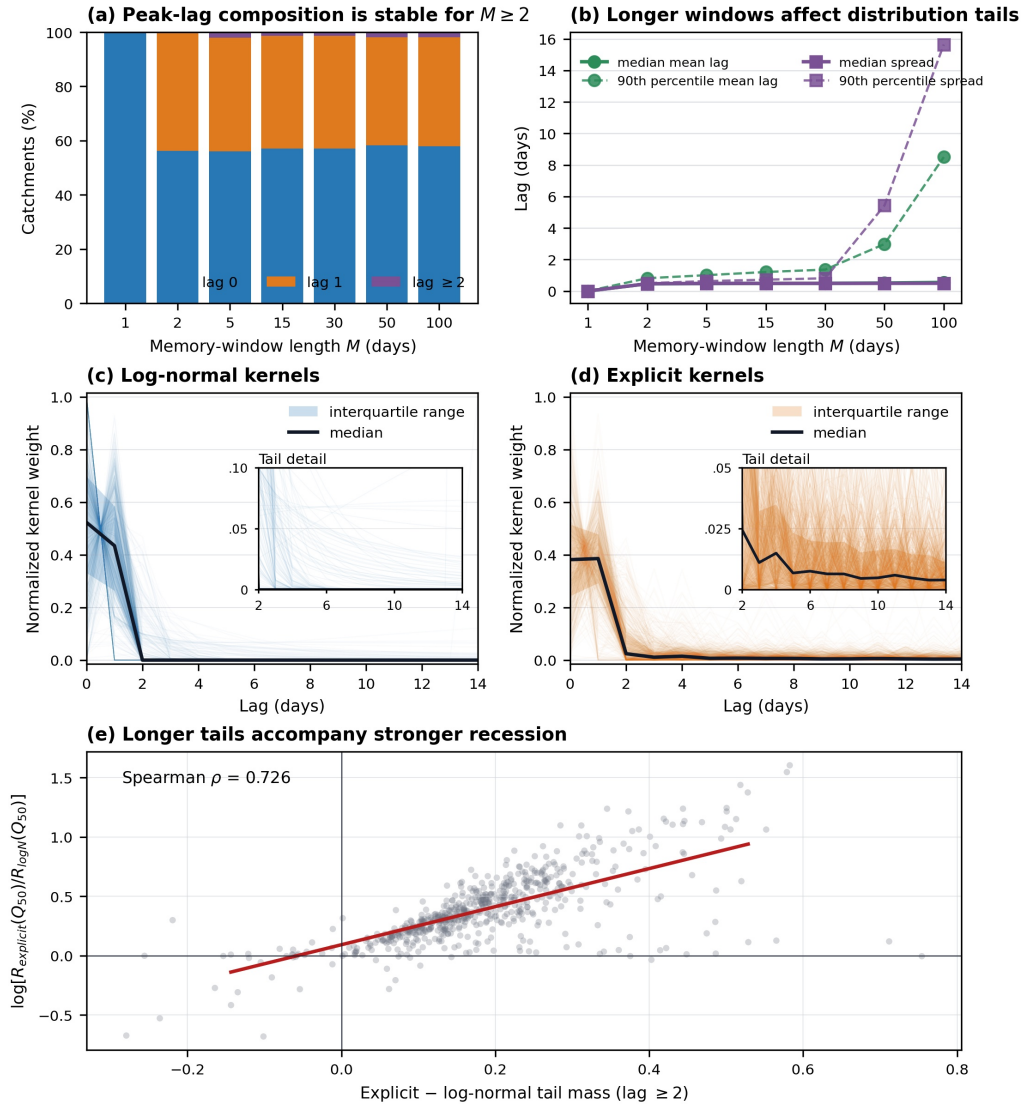


Figure 15. Robustness and identifiability of forcing-memory characteristics across 623 CAMELS-GB catchments. (a) Proportions of catchments whose maximum kernel weight occurs at lag 0, lag 1, or lag 2 or later. (b) Medians (solid) and 90th percentiles (dashed) of mean memory lag and kernel spread across window lengths. (c,d) Normalized log-normal and explicit kernels for all catchments (thin lines), with their pointwise medians (dark lines) and interquartile ranges (shading). Both kernel panels use linear scales; insets enlarge lags 2–14, with vertical limits of 0.10 in (c) and 0.05 in (d). (e) Explicit-minus-log-normal kernel mass at lags 2–14 versus the logarithm of the explicit-to-log-normal recession-loss ratio evaluated at each catchment’s median observed validation discharge, Q_{50} . The reported ρ is Spearman’s rank correlation; the red line is a linear visual guide.

We then tested whether the main results depend on the optimization setup. Expanding all six parameter bounds and increasing the number of initial parameter sets from five to fifteen (Table A1) produced nearly identical validation performance



to the reference calibration: median NSE was 0.801 under both setups, and median KGE was 0.835 and 0.836, respectively. Catchment-wise correlations between the two setups were 0.998 for NSE and 0.996 for KGE, while the current-day versus antecedent-forcing peak classification agreed for 98.1% of catchments. In a separate test of temporal sensitivity, calibrating on each of three non-overlapping thirds of the original calibration period retained peak-class agreement of 90.2–91.3% with the full-period calibration. These tests support the robustness of predictive performance and the principal peak-timing distinction, although they do not imply that every individual parameter or detailed kernel moment is uniquely identified.

The contrast between the two calibrated kernel families further illustrates this distinction (Figure 15c–d). Both families place most of their weight on recent forcing, but the explicit kernels admit more heterogeneous and persistent tails. Compared with the log-normal kernels, their median additional mass at lags of two days or more was 0.173; median mean lag and spread were larger by 1.10 and 2.30 days, respectively. Yet the median explicit-minus-log-normal NSE difference was only -0.001 . The tail contrast is therefore much larger than the difference in predictive performance.

To investigate a possible reason, Figure 15e relates the additional explicit-kernel tail mass to the corresponding change in effective recession loss at Q_{50} . Catchments with more additional tail mass tended also to have stronger recession loss in the explicit formulation (Spearman $\rho = 0.726$). This association is consistent with partial compensation between delayed runoff-generation contributions and recession dynamics: different combinations can yield similar predictions. Because it compares independently calibrated formulations rather than experimentally fixing one component while varying the other, the analysis supports this interpretation but does not establish a unique causal compensation mechanism. Taken together, the sensitivity analyses support peak lag as the more robust diagnostic of dominant forcing-memory timing, while detailed long-tail moments warrant more cautious interpretation.

4.6 Forcing-memory persistence and performance benefit

The relationship between forcing-memory persistence and the performance benefit of the memory formulation is examined from two complementary perspectives. First, catchments are separated according to whether the peak memory lag occurs at lag 0 or at a positive lag. Second, the mean memory lag is considered as a continuous measure of the temporal persistence of forcing memory, allowing its association with model improvement to be assessed across the full range of kernel distributions.

The comparison based on peak memory lag shows a clear separation in performance gains between the two groups (Figure 16a). Catchments with a positive peak memory lag exhibit substantially larger ΔNSE than those with a peak at lag 0. Mean ΔNSE is 0.017 for catchments with a lag-0 peak and 0.074 for those with a positive lag peak (Table 5). The proportion of catchments benefiting from the memory formulation is also higher for positive peak lags, with positive ΔNSE for 100.0% of these catchments compared with 87.9% for lag-0 peaks.

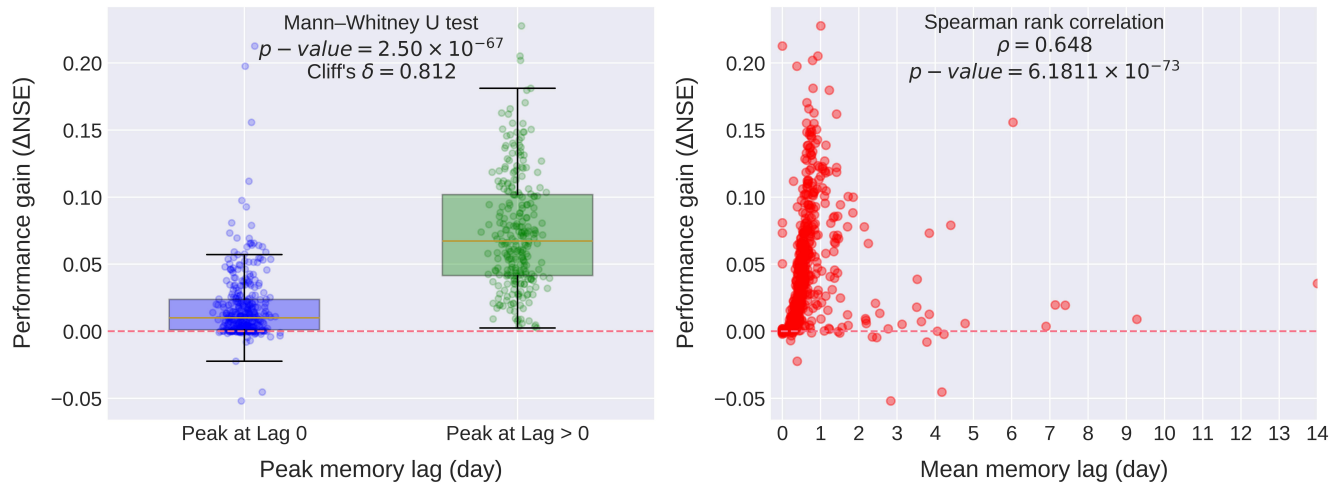


Figure 16. Relationship between forcing-memory persistence and performance benefit. ΔNSE is calculated as the mGRHyMoLAP NSE minus the GRHyMoLAP NSE. (a) Distribution of ΔNSE for catchments with peak memory lags at lag 0 and at positive lags. The Mann–Whitney U test and Cliff’s δ quantify the difference between the two distributions. (b) Relationship between mean memory lag and ΔNSE . The Spearman rank correlation quantifies the monotonic association between the two variables.

Table 5. Memory benefit according to peak memory lag. n denotes the number of catchments.

Peak memory lag	n	Mean ΔNSE	Median ΔNSE	Benefiting	Benefiting (%)
Lag 0	356	0.017	0.010	313	87.9
Positive lag	267	0.074	0.067	267	100.0

The statistical comparison confirms that the two ΔNSE distributions differ strongly. The Mann–Whitney U test gives $p\text{-value} = 2.50 \times 10^{-67}$, while Cliff’s $\delta = 0.811$ indicates a large effect size. Thus, the separation between lag-0 and positive-lag catchments reflects a substantial shift in the distribution of performance gains rather than only a statistically detectable difference.

The continuous analysis based on mean memory lag provides a complementary view of the same pattern. Across the 623 catchments, ΔNSE generally increases with mean memory lag (Figure 16b). Catchments with shorter mean memory lags are predominantly associated with smaller performance gains, whereas larger gains become more prevalent as the mean memory lag increases. Spearman’s rank correlation gives $\rho = 0.639$ with $p\text{-value} = 6.18 \times 10^{-73}$, indicating a positive monotonic association between mean memory lag and ΔNSE .

Together, these results show a consistent association between the temporal persistence of forcing memory and the benefit of representing it explicitly. Catchments with positive peak memory lags exhibit substantially larger performance gains than those with peaks at lag 0, while the mean memory lag indicates that this association becomes stronger as the memory structure



500 extends further toward antecedent forcing. Thus, greater temporal persistence of forcing memory is systematically associated with a larger performance benefit of mGRHyMoLAP across the CAMELS-GB catchments.

4.7 Catchment characteristics associated with the peak memory lag

505 Following the comparison of performance benefit according to forcing-memory persistence, we examine the catchment characteristics associated with different peak memory lags. Catchments are grouped according to whether the forcing-memory kernel reaches its maximum at lag 0 or at a positive lag.

The comparison shows systematic differences in several catchment characteristics between the lag-0 and positive-lag groups (Table 6). The largest differences in median values occur for catchment area, cropland fraction, mean elevation, drainage-path slope, and mean precipitation. The positive-lag group exhibits a higher median catchment area (278.050 versus 106.975) and cropland fraction (27.950 versus 5.145), and lower mean elevation (125.000 versus 195.500) and drainage-path slope (60.400 versus 93.300). Mean precipitation is also lower in the positive-lag group (2.280 versus 3.105), whereas aridity is higher (0.620 versus 0.415).

Table 6. Comparison of catchment attributes between lag-0 and positive-lag peak memory groups. The difference is calculated as the median value in the positive-lag group minus the median value in the lag-0 group. Statistical differences were assessed using two-sided Mann–Whitney U tests. Cliff’s δ quantifies the corresponding effect size. *Note:* $L0$ = lag 0; PL = positive lag; n = number of catchments; Med = median.

Group	Attribute	n_{L0}	n_{PL}	Med $L0$	Med PL	Difference	p -value	Cliff’s δ
Topography	area	356	267	106.975	278.050	171.075	1.82×10^{-26}	0.498
Hydrology	high_q_dur	356	267	1.390	1.810	0.420	2.91×10^{-25}	0.486
Climate	p_mean	356	267	3.105	2.280	-0.825	9.24×10^{-17}	-0.389
Climate	aridity	356	267	0.415	0.620	0.205	4.99×10^{-16}	0.379
Land cover	crop_perc	356	267	5.145	27.950	22.805	7.08×10^{-16}	0.377
Hydrology	runoff_ratio	356	267	0.620	0.440	-0.180	3.29×10^{-13}	-0.341
Topography	dpsbar	356	265	93.300	60.400	-32.900	1.11×10^{-12}	-0.334
Hydrology	Q95	356	267	6.550	3.280	-3.270	2.62×10^{-12}	-0.327
Climate	p_seasonality	356	267	-0.170	-0.090	0.080	4.26×10^{-11}	0.308
Topography	elev_mean	356	265	195.500	125.000	-70.500	1.51×10^{-10}	-0.300
Soil	clay_perc	356	267	23.100	24.470	1.370	2.25×10^{-8}	0.262
Soil	conductivity_cosby	356	267	1.390	1.300	-0.090	3.34×10^{-8}	-0.258
Soil	root_depth	356	267	1.050	1.150	0.100	1.38×10^{-7}	0.246
Climate	pet_mean	356	267	1.350	1.400	0.050	4.61×10^{-6}	0.214
Hydrogeology	frac_high_perc	356	267	0.000	0.000	0.000	6.58×10^{-5}	0.152
Hydrology	baseflow_index	356	267	0.530	0.580	0.050	1.47×10^{-4}	0.178
Hydrogeology	frac_low_perc	356	267	1.860	0.000	-1.860	9.40×10^{-3}	-0.116
Land cover	dwood_perc	356	267	5.700	5.970	0.270	1.14×10^{-1}	0.074
Hydrogeology	inter_low_perc	356	267	0.000	0.000	0.000	2.50×10^{-1}	0.030
Soil	tawc	356	267	90.580	91.370	0.790	3.73×10^{-1}	0.042



Several additional differences are observed across hydrological, climatic, soil, and hydrogeological attributes. The positive-lag group exhibits higher median high-flow duration, precipitation seasonality, clay fraction, root depth, potential evapotranspiration, and baseflow index, together with lower runoff ratio and 95th percentile flow. The corresponding effect sizes indicate moderate to large differences for the attributes showing the strongest separation. In particular, catchment area ($\delta = 0.498$) and high-flow duration ($\delta = 0.486$) show the largest effects, followed by mean precipitation ($\delta = -0.389$), aridity ($\delta = 0.379$), and cropland fraction ($\delta = 0.377$).

Differences are also detected for several hydrogeological attributes, although their effect sizes are generally smaller. The positive-lag group has a lower median fraction of low-productivity aquifers ($\delta = -0.116$), corresponding to a negligible effect, whereas the median fraction of highly productive aquifers is zero in both groups. Although the latter attribute shows a statistically significant difference in its distributions, the identical medians indicate that the difference is not reflected in their central values. Other attributes, including deciduous woodland ($\delta = 0.074$), intergranular low-productivity aquifers ($\delta = 0.030$), and total available water capacity ($\delta = 0.042$), also show negligible effects, indicating limited practical separation between the two groups.

Overall, catchments in which the forcing-memory kernel reaches its maximum at positive lags tend to be larger and to exhibit longer high-flow durations, lower precipitation, higher aridity, greater cropland fractions, and distinct runoff-regime characteristics. These differences suggest that the temporal organization of runoff-generation forcing may vary with catchment size, climatic conditions, land cover, and flow-regime characteristics. However, these attributes should be interpreted as characteristics associated with the calibrated peak memory lags rather than as individual controls of their occurrence.

5 Discussion

5.1 mGRHyMoLAP enhances rainfall–runoff prediction through log-normal kernel representation of forcing memory

In this study, the results show that representing forcing memory through probability-based kernels provides a consistent improvement over the original GRHyMoLAP formulation, supporting the broader evidence that rainfall–runoff response is influenced by the legacy of past hydroclimatic conditions (Nippgen et al., 2016). This pattern suggests that GRHyMoLAP’s instantaneous allocation of runoff-generation forcing is restrictive in many CAMELS-GB catchments. This should not, however, be interpreted as suggesting that existing rainfall–runoff models do not account for antecedent forcing. Conceptual models such as GR4J and GR6J commonly represent such effects implicitly through model states, reservoirs, routing, or transfer functions such as unit hydrographs (Perrin et al., 2003). The contribution of mGRHyMoLAP is instead to explicitly represent the temporal allocation of runoff-generation forcing in GRHyMoLAP through probability-based kernels, allowing its temporal structure within the model to be characterized.

Temporal dependencies have also been incorporated in data-driven and machine-learning approaches through sequences of lagged meteorological inputs and/or hydrological variables. For example, Ashfaq et al. (2026); Houénafa and Sylla (2026); Ao et al. (2025); Sezen and Šraj (2024) incorporated lagged or antecedent information to improve rainfall–runoff predictions.



545 In contrast, mGRHyMoLAP represents the temporal dependence of past runoff-generation forcing, defined by the interaction between percolation and effective precipitation, through a probability-based kernel. This formulation provides an interpretable representation of how temporal weights are assigned to runoff-generation forcing across different lags.

The explicit discrete kernel can represent a broad range of lag-specific shapes, including peaks at non-zero lags and substantial long tails. The log-normal formulation retains the ability to represent asymmetric short-lag weighting and non-zero peaks
 550 with substantially fewer parameters, while achieving similar validation performance. The sensitivity analysis helps explain this result: allowing one antecedent day captures most of the median predictive benefit, whereas the much longer tails fitted by some explicit kernels provide little additional predictive gain. Moreover, these tails are associated with changes in recession loss, consistent with partial compensation between model components. Thus, the evidence supports temporal redistribution of runoff-generation forcing as the important extension of GRHyMoLAP, but does not establish a uniquely identified long
 555 memory distribution. The log-normal kernel offers a parsimonious representation of the dominant model-implied timing.

The improved consistency in the direction of simulated flow changes provides an additional perspective on the improvement associated with mGRHyMoLAP. The higher agreement during increasing-flow periods suggests that representing forcing memory improves agreement with the observed direction of daily flow changes, rather than only the magnitude of simulated flows. This is particularly relevant during periods without effective precipitation, when observed increases in streamflow cannot
 560 be directly associated with contemporaneous runoff-generation forcing. This finding addresses a main limitation of GRHyMoLAP, whose runoff-generation term vanishes when no same-day effective input is present (Houénafa et al., 2026b). The sign consistency achieved by mGRHyMoLAP, which is even slightly higher than that of GR4J, further supports its ability to represent the temporal behaviour of hydrological systems. Overall, these results indicate that probability-based representations of current and past runoff-generation forcing can improve both the magnitude and temporal behaviour of simulated flow.

565 **5.2 Forcing-memory persistence and its dependence on catchment characteristics**

The calibrated kernels indicate that forcing memory is generally concentrated over short lags, with the kernel reaching its maximum most frequently at lag 0 (57% of catchments) or at lag 1. Positive peak memory lags in a subset of catchments show that the dominant timing of runoff-generation forcing varies across the CAMELS-GB dataset, consistent with evidence of variability in the temporal characteristics of rainfall–runoff interactions across catchments (Moges et al., 2022). This peak-
 570 lag distinction remains largely stable across memory-window lengths and calibration choices. Previous work has found that antecedent rainfall can matter over different time scales (Pathiraja et al., 2012). Here, forcing-memory persistence is characterized using the peak memory lag and the mean memory lag. The former identifies the lag at which the model assigns the greatest temporal weight to runoff-generation forcing, whereas the latter summarizes the overall temporal position of the memory distribution. The fitted mean lags and long-tail shapes are less stable across the sensitivity analyses and may partly trade
 575 off against recession dynamics.

This temporal structure is closely associated with the benefit obtained from representing forcing memory. Catchments with positive peak memory lags show substantially larger performance gains than those with peaks at lag 0. Notably, mGRHyMoLAP improves upon GRHyMoLAP in 100% of catchments with positive peak memory lags. This indicates that the probability-



based memory formulation is particularly beneficial for catchments in which the calibrated memory structure places greater
580 temporal weight on antecedent runoff-generation forcing, consistent with evidence that antecedent wetness can substantially
strengthen the runoff response to precipitation (Knapp et al., 2025).

The differences in catchment characteristics between lag-0 and positive-lag groups provide further hydrological context for
this behaviour. Catchments with positive peak memory lags tend to have larger areas, longer high-flow durations, lower mean
precipitation, higher aridity, greater cropland fractions, and distinct flow-regime characteristics. The strongest group separation
585 is observed for catchment area and high-flow duration, while several climatic and land-cover characteristics also show moderate
differences. These associations are relevant because previous studies have shown that catchment response times and the
temporal distribution of runoff responses can be systematically related to catchment attributes. For example, using a large set of
CAMELS-GB catchments, machine-learning approaches identified catchment area, aridity, cropland fraction, runoff ratio, and
high- and low-flow durations among the influential controls on catchment response time (Ntigliakis, 2020; Zheng et al., 2023).
590 Several of these attributes—catchment area, aridity, cropland fraction, and high-flow duration—also distinguish catchments
with positive peak memory lags in our analysis. More broadly, catchment response time has been related to catchment area,
geomorphology, climate, and land-cover characteristics, with spatial and temporal heterogeneity influencing runoff response
across scales (Zheng et al., 2023; Gericke, 2019; Eslami et al., 2025). Together, these studies suggest that catchment attributes
can provide useful information about the temporal characteristics of rainfall–runoff response. In our analysis, the association of
595 larger area, higher aridity, greater cropland fraction, and longer high-flow duration with non-zero peak memory lags indicates
that these characteristics are also associated with the temporal structure of the runoff-generation forcing represented by the
model. However, these results should be interpreted as catchment characteristics associated with the peak memory lag rather
than as evidence that individual attributes determine this lag.

Taken together, these findings indicate that the temporal structure of forcing memory varies across catchments and that this
600 variation is relevant to the benefit obtained from its representation. Beyond its use for rainfall–runoff prediction, mGRHy-
MoLAP provides a framework for characterizing how the model temporally weights current and antecedent runoff-generation
forcing in the representation of current runoff generation. The combined analysis of kernel behaviour, memory persistence,
performance benefit, and catchment characteristics provides a basis for identifying catchments where representing forcing
memory is most likely to improve rainfall–runoff prediction.

605 5.3 Future avenues

The sensitivity analysis puts the 15-day reference window into perspective: extending the admissible window allowed some
catchments to develop substantially longer kernel tails without corresponding predictive gains. A long fitted tail should there-
fore not automatically be interpreted as evidence of a comparably long physical runoff-generation memory. Future work could
examine whether independent information about recession behaviour or event-scale response helps distinguish delayed forc-
610 ing contributions from compensating model dynamics. Catchment-specific window lengths may still be useful, but their value
should be judged by demonstrable predictive or diagnostic gains rather than by the flexibility to fit longer tails alone.



Furthermore, future work could investigate more flexible kernel distributions, including multimodal formulations. The exponential and log-normal kernels considered in this study impose relatively simple temporal structures, with either monotonic decay or a single dominant region of temporal weighting. A multimodal kernel could instead represent forcing-memory structures characterized by multiple distinct regions of higher temporal weighting across lags, potentially providing a more flexible representation of catchments with complex temporal organization of runoff-generation forcing. Comparing unimodal and multimodal formulations could therefore help determine whether such additional flexibility provides meaningful predictive or hydrological benefits.

Although the proposed formulation is implemented within GRHyMoLAP, the underlying concept is not restricted to this model. More generally, the proposed probability-based memory operator can be applied to other process-based rainfall-runoff models in which runoff generation is driven by an effective hydrological input. Future work could therefore investigate its integration into other modelling frameworks to assess whether the benefits of representing forcing memory extend across different model structures. Such applications would also help determine whether the identified relationships between forcing-memory characteristics and catchment attributes are robust across modelling frameworks.

Furthermore, the proposed framework could be extended to investigate whether contributions from antecedent runoff-generation forcing have a greater influence on predictive flow uncertainty in catchments where these contributions are stronger. This would extend the analysis beyond the relationship between forcing memory and predictive performance to examine whether the temporal structure of runoff-generation forcing also affects predictive uncertainty. Such an analysis could provide further insight into how forcing memory influences probabilistic streamflow prediction.

6 Conclusions

The performance of mGRHyMoLAP shows that distributing runoff-generation forcing across current and antecedent days can improve daily streamflow prediction within the GRHyMoLAP model structure. The mGRHyMoLAP framework proposed here extends GRHyMoLAP by introducing probability-based forcing-memory kernels while preserving its process-driven ODE structure. Its overall performance across the 623 CAMELS-GB catchments, together with the comparison with the widely used GR4J model, demonstrates the potential of mGRHyMoLAP as an improved ODE-based rainfall-runoff model for representing flow dynamics.

mGRHyMoLAP addresses an important limitation of GRHyMoLAP, in which runoff-generation forcing vanishes when current effective precipitation is absent. This limitation limits the ability of GRHyMoLAP to represent continued runoff generation under antecedent forcing, even though hydrological response can persist beyond the current precipitation input. The proposed mGRHyMoLAP provides a simple way to overcome this limitation by allowing current and antecedent runoff-generation forcing to contribute to flow generation over time. This improvement is reflected not only in overall performance but also in the daily flow-change analysis: mGRHyMoLAP reproduces the observed direction of streamflow changes more consistently than GRHyMoLAP and achieves slightly higher sign consistency than GR4J.



Beyond prediction, the calibrated kernels describe how the fitted model allocates runoff-generation forcing across current and antecedent days. Most catchments place their largest weight at lag 0 or lag 1, while 42.9% have a non-zero peak. This peak-timing distinction remains largely robust across the sensitivity analyses, and all catchments with a non-zero peak benefit from the memory formulation relative to GRHyMoLAP. Larger mean lags are also associated with larger performance gains, but detailed kernel moments and long tails are less robust and may partly compensate for recession dynamics. Peak lag is therefore the more defensible model-based diagnostic of dominant forcing timing; neither it nor the full kernel should be read as a direct measurement of physical catchment memory.

The analysis of catchment characteristics further shows that positive peak memory lags are associated with larger catchment area, longer high-flow duration, lower mean precipitation, higher aridity, and greater cropland fraction, providing evidence that the temporal structure of runoff-generation forcing memory varies systematically across hydroclimatic, hydrological, topographic, and land-cover conditions. These findings highlight the value of calibrated memory kernels not only for improving rainfall–runoff prediction but also for investigating how catchment characteristics are related to the persistence and lag structure of runoff-generation forcing memory. Future work can test whether these model-based peak-lag relationships transfer to other rainfall–runoff models and whether independent hydrological information can distinguish long forcing-memory tails from compensating recession dynamics. More flexible or catchment-specific kernels should be evaluated for demonstrable predictive gains and stable diagnostic features.

Code and data availability. The data used in this study were obtained from the CAMELS-GB dataset (Coxon et al., 2020), available at <https://doi.org/10.5285/8344e4f3-d2ea-44f5-8afa-86d2987543a9>. The open-source code used for all simulations is available in the corresponding GitHub repository: <https://github.com/Ezesia-lab/mGRHyMoLAP>.

Appendix A: Calibration settings for the expanded-search sensitivity analysis

Table A1 documents the parameter bounds used in the reference $M = 15$ calibration and in the expanded-search sensitivity analysis. Both calibrations minimized $1 - \text{NSE}$ with Nelder–Mead, allowed up to 3000 iterations per initial parameter set, and used the same calibration and validation periods. The sensitivity analysis changed the bounds and the number of initial parameter sets, but not the model formulation or evaluation data.



Table A1. Parameter bounds and number of initial parameter sets in the reference and expanded-search calibrations of log-normal mGRHy-MoLAP.

Setting	Reference	Expanded search
μ	[0.5, 3.5]	[0.5, 10]
λ	$[10^{-4}, 300]$	$[10^{-4}, 1000]$
X_1	$[10^{-4}, 5000]$	$[10^{-4}, 10000]$
γ	[0, 20]	[0, 100]
$\mu_{\log N}$	[-20, 20]	[-100, 100]
$\sigma_{\log N}$	$[10^{-3}, 10]$	$[10^{-3}, 50]$
Initial parameter sets	5	15

Author contributions. Sianou Ezékiel Houénafa: Conceptualization, Methodology, Software, Formal analysis, Validation, Visualization, and Writing—original draft. Wolfgang Nowak: Methodology, Validation, Visualization, and Writing—review and editing.

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

Acknowledgements. The authors acknowledge the support by the Stuttgart Center for Simulation Science (SimTech), which is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC 2075, project number 390740016). OpenAI ChatGPT and Codex were used as assistive tools for language editing, reference consistency checks, LaTeX formatting, and preparation of submission materials. All scientific content and final wording were reviewed and approved by the authors, who take full
 675 responsibility for the manuscript.



References

- Afouda, A. and Alamous, E.: Modèle hydrologique basé sur le principe de moindre action (MODHYPMA), *Ann. Sci. Agron. Bénin*, 13, 23–45, <https://doi.org/10.4314/asab.v13i1.53862>, 2010.
- Ao, Y., Jiang, P., Zhang, R., Yang, B., Qu, S., Dunzhu, J., Wu, R., Fang, Z., She, L., and Yu, Z.: Incorporating accumulated temperature and lagged precipitation effects in LSTM modeling improves the streamflow simulation in glacier dominated basins, *J. Hydrol.-Reg. Stud.*, 62, 102872, <https://doi.org/10.1016/j.ejrh.2025.102872>, 2025.
- Ashfaq, S., Waseem, S., Tufail, M., and Tariq, A.: Evaluating the accuracy of machine learning models with short-term lag inputs for univariate daily streamflow time series forecasting, *Water Resour. Manag.*, 40, 177, <https://doi.org/10.1007/s11269-025-04409-x>, 2026.
- Bergström, S. and Forsman, A.: Development of a conceptual deterministic rainfall–runoff model, *Hydrol. Res.*, 4, 147–170, <https://doi.org/10.2166/nh.1973.0012>, 1973.
- 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 Syst. Sci. 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 Syst. Sci. Data*, 17, 1461–1479, <https://doi.org/10.5194/essd-17-1461-2025>, 2025.
- Eslami, Z., Seybold, H., and Kirchner, J. W.: Climatic, topographic, and groundwater controls on runoff response to precipitation: evidence from a large-sample data set, *Hydrol. Earth Syst. Sci.*, 29, 5121–5130, <https://doi.org/10.5194/hess-29-5121-2025>, 2025.
- Farrick, K. K. and Branfireun, B. A.: Soil water storage, rainfall and runoff relationships in a tropical dry forest catchment, *Water Resour. Res.*, 50, 9236–9250, <https://doi.org/10.1002/2014WR016045>, 2014.
- Fowler, K. J. A., Acharya, S. C., Addor, N., Chou, C., and Peel, M. C.: CAMELS-AUS: hydrometeorological time series and landscape attributes for 222 catchments in Australia, *Earth Syst. Sci. Data*, 13, 3847–3867, <https://doi.org/10.5194/essd-13-3847-2021>, 2021.
- Gericke, O. J.: GIS applications to investigate the linkage between geomorphological catchment characteristics and response time: a case study in four climatological regions, South Africa, *Water*, 11, 1072, <https://doi.org/10.3390/w11051072>, 2019.
- Gohouede, L. C., Houénafa, S. E., Biao, E. I., Olatunji, J., and Sezen, C.: A dimensionally consistent reformulation of the HyMoLAP rainfall–runoff model: interpretability and robustness across CAMELS datasets, SSRN [preprint], <https://doi.org/10.2139/ssrn.6333826>, 3 March 2026.
- 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.
- Hashemi, R., Brigode, P., Garambois, P. A., and Javelle, P.: How can we benefit from regime information to make more effective use of long short-term memory (LSTM) runoff models?, *Hydrol. Earth Syst. Sci.*, 26, 5793–5816, <https://doi.org/10.5194/hess-26-5793-2022>, 2022.
- Houénafa, S. E., Johnson, O., Ronoh, E. K., and Moore, S. E.: Hybridization of stochastic hydrological models and machine learning methods for improving rainfall–runoff modeling, *Results Eng.*, 25, 104079, <https://doi.org/10.1016/j.rineng.2025.104079>, 2025.
- Houénafa, S. E. and Sylla, M. B.: Enhancing conceptual rainfall–runoff modeling in data-scarce catchments using machine learning: Kolmogorov–Arnold networks compared to LSTMs, *Earth Syst. Environ.*, 10, 7789–7810, <https://doi.org/10.1007/s41748-025-01010-5>, 2026.



- Houénafa, S. E., Gohouede, L. C., Boy-ngbogbele, R. D., Latella, M., Johnson, O., and Sezen, C.: S-GRHyMoLAP: a stochastic extension of the GRHyMoLAP model with different diffusion mechanisms for probabilistic rainfall–runoff modeling, *J. Hydrol.*, 680, 136449, <https://doi.org/10.1016/j.jhydrol.2026.136449>, 2026a.
- 715 Houénafa, S. E., Latella, M., Gohouede, L. C., and Sezen, C.: GRHyMoLAP: a process-driven ODE catchment hydrology model inspired by GR4J and HyMoLAP approaches, *J. Hydrol.*, 675, 135597, <https://doi.org/10.1016/j.jhydrol.2026.135597>, 2026b.
- Knapp, J. L., Berghuijs, W. R., Floriancic, M. G., and Kirchner, J. W.: Catchment hydrological response and transport are affected differently by precipitation intensity and antecedent wetness, *Hydrol. Earth Syst. Sci.*, 29, 3673–3685, <https://doi.org/10.5194/hess-29-3673-2025>, 2025.
- 720 Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall–runoff modelling using long short-term memory (LSTM) networks, *Hydrol. Earth Syst. Sci.*, 22, 6005–6022, <https://doi.org/10.5194/hess-22-6005-2018>, 2018.
- Kunnath-Poovakka, A. and Eldho, T. I.: A comparative study of conceptual rainfall–runoff models GR4J, AWBM and Sacramento at catchments in the upper Godavari river basin, India, *J. Earth Syst. Sci.*, 128, 33, <https://doi.org/10.1007/s12040-018-1055-8>, 2019.
- Mathias, S. A., Thébault, C., and Ireson, A.: A theoretical appraisal of the GR4J rainfall–runoff modelling framework, *J. Hydrol.*, 664, 134393, <https://doi.org/10.1016/j.jhydrol.2025.134393>, 2026.
- 725 Meissel, K. and Yao, E. S.: Using Cliff’s delta as a non-parametric effect size measure: an accessible web app and R tutorial, *Pract. Assess. Res. Eval.*, 29, 2, 1–12, <https://doi.org/10.7275/pare.1977>, 2024.
- Moges, E., Ruddell, B. L., Zhang, L., Driscoll, J. M., and Larsen, L. G.: Strength and memory of precipitation’s control over streamflow across the conterminous United States, *Water Resour. Res.*, 58, e2021WR030186, <https://doi.org/10.1029/2021WR030186>, 2022.
- 730 Moriasi, D. N., Gitau, M. W., Pai, N., and Daggupati, P.: Hydrologic and water quality models: performance measures and evaluation criteria, *Trans. ASABE*, 58, 1763–1785, <https://doi.org/10.13031/trans.58.10715>, 2015.
- Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I—a discussion of principles, *J. Hydrol.*, 10, 282–290, [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6), 1970.
- Nelder, J. A. and Mead, R.: A simplex method for function minimization, *Comput. J.*, 7, 308–313, <https://doi.org/10.1093/comjnl/7.4.308>, 1965.
- 735 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, *Hydrol. Earth Syst. Sci.*, 19, 209–223, <https://doi.org/10.5194/hess-19-209-2015>, 2015.
- 740 Nguyen, T., Bennett, B., and Leonard, M.: Tracking rainfall contribution in a conceptual runoff model, *J. Hydrol.*, 672, 135366, <https://doi.org/10.1016/j.jhydrol.2026.135366>, 2026.
- Nippgen, F., McGlynn, B. L., Emanuel, R. E., and Vose, J. M.: Watershed memory at the Coweeta Hydrologic Laboratory: the effect of past precipitation and storage on hydrologic response, *Water Resour. Res.*, 52, 1673–1695, <https://doi.org/10.1002/2015WR018196>, 2016.
- Ntigkakakis, C.: Investigation of the most influential controls for catchment response time using machine learning, M.S. thesis, University of Bristol, Bristol, UK, <https://doi.org/10.13140/RG.2.2.14263.75684>, 2020.
- 745 Pathiraja, S., Westra, S., and Sharma, A.: Why continuous simulation? The role of antecedent moisture in design flood estimation, *Water Resour. Res.*, 48, W06534, <https://doi.org/10.1029/2011WR010997>, 2012.
- Payan, J.-L., Perrin, C., Andréassian, V., and Michel, C.: How can man-made water reservoirs be accounted for in a lumped rainfall–runoff model?, *Water Resour. Res.*, 44, W03420, <https://doi.org/10.1029/2007WR005971>, 2008.



- 750 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](https://doi.org/10.1016/S0022-1694(03)00225-7), 2003.
- Sezen, C. and Partal, T.: The utilization of a GR4J model and wavelet-based artificial neural network for rainfall–runoff modelling, *Water Supply*, 19, 1295–1304, <https://doi.org/10.2166/ws.2018.189>, 2019.
- Sezen, C. and Šraj, M.: Improving the simulations of the hydrological model in the karst catchment by integrating the conceptual model with
755 machine learning models, *Sci. Total Environ.*, 926, 171684, <https://doi.org/10.1016/j.scitotenv.2024.171684>, 2024.
- Wilcoxon, F.: Individual comparisons by ranking methods, *Biometrics Bull.*, 1, 80–83, <https://doi.org/10.2307/3001968>, 1945.
- Xu, Y., Hu, C., Wu, Q., Li, Z., Jian, S., and Chen, Y.: Application of temporal convolutional network for flood forecasting, *Hydrol. Res.*, 52, 1455–1468, <https://doi.org/10.2166/nh.2021.021>, 2021.
- Zheng, Y., Coxon, G., Woods, R., Li, J., and Feng, P.: Controls on the spatial and temporal patterns of rainfall–runoff event characteristics—
760 a large sample of catchments across Great Britain, *Water Resour. Res.*, 59, e2022WR033226, <https://doi.org/10.1029/2022WR033226>, 2023.