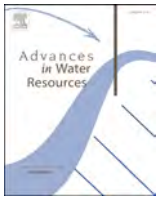
Residual dynamics in hydrological models: insights from a large sample of catchments and models

Advances in Water Resources 206 (2025) 105165

Contents lists available at ScienceDirect

Advances in Water Resources

journal homepage: www.elsevier.com/locate/advwatres



Residual dynamics in hydrological models: insights from a large sample of catchments and models

Luca Lombardo [a,*], Simon Michael Papalexiou [b,c,d], Cyril Thébault [c], Martyn P. Clark [c], Richard M. Vogel [e], Alberto Viglione [a]

[a] Dipartimento di Ingegneria dell'Ambiente, del Territorio e delle Infrastrutture, Politecnico di Torino, Turin, Italy
[b] Institute of Global Water Security, Hamburg University of Technology (TUHH), Hamburg, Germany
[c] Department of Civil Engineering, University of Calgary, Calgary, Alberta, Canada
[d] Department of Civil, Geological and Environmental Engineering, University of Saskatchewan, Saskatoon, Saskatchewan, Canada
[e] Department of Civil & Environmental Engineering, Tufts University, Medford, MA, USA

ARTICLE INFO

Keywords:
Hydrological modeling
Model residuals
Post-Processing

ABSTRACT

Residual analysis is a cornerstone of hydrological modelling, providing the basis for assessing model performance, diagnosing structural deficiencies, and building stochastic error models for uncertainty quantification. This study investigates the properties of residuals from 78 hydrological models applied to 422 catchments spanning the contiguous United States offering an unprecedented multi-catchment, multi-model perspective. Key aspects examined include shape properties (L-skewness and L-kurtosis) assessed through both conventional L-moment diagrams and diagrams adapted for symmetric distributions. Residual heteroscedasticity, and temporal correlation are also analyzed, together with their variability due to model selection, hydrological regimes, and under different discharge transformations (Box-Cox and logarithmic). In addition, we evaluated the impact of seasonality removal, showing that while it substantially stabilizes higher-order moments and reduces heavy tails, it is less effective in mitigating heteroskedasticity, for which transformations play a crucial role. Finally, upper and lower temporal tails correlations are explored, revealing distinct behaviors that differ from general temporal correlation patterns. Collectively, these results provide a robust empirical foundation for the design of generalizable stochastic error models, with direct implications for predictive accuracy and uncertainty quantification in hydrology.

1. Introduction

Hydrological systems are inherently complex, with nonlinear interactions across multiple spatial and temporal scales that are difficult to observe and represent explicitly in mechanistic models (Beven, 2012; Clark et al., 2015). To make this complexity tractable, hydrologists are often required to introduce simplifications and compromises, both to reduce data requirements and to ensure that models can be implemented within reasonable computational constraints. Conceptual ("bucket-style") hydrological models respond to this need by providing a parsimonious representation of catchment function. The balance of conceptual models between realism, flexibility, and computational efficiency has made them highly successful in both research and operational contexts (Jakeman & Hornberger, 1993; Perrin et al., 2003). Nevertheless, such models remain imperfect descriptions of reality, as their simplified structure inevitably leaves some hydrological processes only approximately or implicitly represented (Beven, 2012; Clark et al., 2015; Hrachowitz & Clark, 2017). These structural limitations highlight the importance of explicitly accounting for model uncertainty, both in terms of parameter estimation and in the ability of the model to reproduce observed hydrological variability.

Model residuals incorporate the discrepancies between the observed hydrological variable and the model prediction. It is important to distinguish between the terms “residual” and “error”: while the first indicates a single realization (hence a time series), the second refers instead to the underlying stochastic process (Kavetski et al., 2006). Model residuals are fundamental to evaluate hydrological model performance and improve model structure. By analyzing residual properties, hydrologists can identify model deficiencies, such as oversimplified representations of hydrological processes or missing nonlinear

* Corresponding author.
E-mail address: luca_lombardo@polito.it (L. Lombardo).

https://doi.org/10.1016/j.advwatres.2025.105165
Received 23 April 2025; Received in revised form 4 October 2025; Accepted 2 November 2025
Available online 3 November 2025

dynamics. These insights are not merely diagnostic but provide a foundation for uncertainty frameworks (Evin et al., 2013, 2014; Schoups & Vrugt, 2010; Shabestanipour et al., 2023).

Residual analysis has two primary applications in hydrological modelling. First, residual analysis serves as the foundation for evaluating goodness of fit. For instance, the commonly-used Nash-Sutcliffe Efficiency (NSE - Nash & Sutcliffe, 1970) performance metric is based on the ratio of the sum-of-squared residuals and the estimated variance of the observations. Second, residual analysis forms the basis for proposing error models for stochastic parameter inference and uncertainty quantification, generating streamflow ensembles, which are both useful and even necessary for nearly all risk-based hydrological planning activities (Vogel, 2017).

When analyzing model residuals and developing methods based on their statistical characterization, a central challenge lies in distinguishing patterns that are consistent across diverse catchments from differences that arise due to their heterogeneous characteristics, including variations in climate, morphology, flow regimes, and hydrological processes. Such complexity is further amplified by the variability introduced when different hydrological models are applied to the same catchment, reflecting differences in process representation and parameterization (Clark et al., 2011; Gupta et al., 2009). In order to allow the development of error models that are robust and transferable across the widest possible range of conditions, rather than relying on assumptions that hold only in specific contexts, comprehensive and well-calibrated multi-model and multi-catchment datasets are essential.

To the authors' knowledge, previous studies in the literature that addressed such issues have always focused only on part of this variability: most of those studies (e.g., Schoups & Vrugt, 2010; Tajiki et al., 2020) generally focused on few catchments, often characterized by a limited climate and hydrological variability or just considering two catchments with radical opposite characteristics (i.e, very arid and very humid), and applying single or few (usually at most two, e.g. Evin et al., 2014) different rainfall-runoff model structures. Other examples (Bourgin et al., 2014; McInerney et al., 2017; Shabestanipour et al., 2023) involved larger catchment ensembles, but with the application of a single model scheme. Moreover, the same larger ensemble studies have mainly focused on uncertainty estimation or error models intercomparisons rather than systematically examining residual characteristics across catchments.

The focus of this work is targeted at a detailed pure statistical description of the properties of aggregated rainfall-runoff model residuals, systematically exploring a new large multi-model and multi-catchment dataset (described in detail in Section 2.1), substantially larger and geographically more diverse than that used by previous multi-catchment studies. This large-sample dataset ensures that residual analyses reflect robust general properties rather than being representative of few individual cases. Additionally to the characterization of "raw" model residuals, the effect of transformations (Section 2.2) is also investigated to address how residual properties can be stabilized and be reconducted to more uniform properties. The broader intent of this study is to enable hydrologists to gain a deeper understanding of the structure of conceptual rainfall–runoff model residuals. This, in turn, will lead to improvements in goodness-of-fit evaluation and overall model reliability, as well as support the development of more generalizable error-modeling schemes.

The remainder of the study is structured as follows: Section 2 introduces the multi-model dataset and the analysis of residuals; Section 3 examines seasonality, shape properties, and heteroscedasticity in residuals; Section 4 investigates temporal dependencies through correlation analyses; finally, Section 5 discusses and synthesizes previous findings to propose residual modelling strategies, emphasizing their implication for predictive accuracy and uncertainty quantification.

2. Data and methods

2.1. Models and simulations dataset – FUSE

This study utilizes a subset of the CAMELS dataset (Addor et al., 2017; Newman et al., 2015), containing data for 559 catchments across the contiguous United States, spanning a wide range of climatic and hydrological characteristics (such subset was obtained from the original 671 catchments dataset after processing for consistency in catchment size and water balance, see Knoben et al., 2020). Simulations (at a daily time scale) are conducted using 78 configurations of the Framework for Understanding Structural Errors (FUSE) software (M. P. Clark et al., 2008), enabling an exploration of differences in hydrological process representation. Each FUSE configuration is a different conceptual hydrological model where the vertical dimension is always discretized into two zones (unsaturated and saturated), but in which the different configurations vary in the definition of the architecture of the zones (the number and arrangement of "buckets" that represent the storage of water) and their parameterization (the equations that represent the model fluxes such as infiltration, vertical percolation, and baseflow, that affect the time evolution of water storage in each of the conceptual buckets).

The dataset leverages recent efforts to calibrate a large ensemble of FUSE models to a large sample of CAMELS catchments (Thébault, C., 2025). The dataset includes over 43,600 simulated discharge time series (78 FUSE models for each of the 559 catchments) that cover regions from humid eastern climates to arid southwestern zones. Model calibration is performed using the Shuffled Complex Evolution algorithm (SCE-UA; Duan et al., 1992), with 1,000 iterations, and the Kling–Gupta Efficiency (KGE; Gupta et al., 2009) as the objective function. Each of the FUSE models was calibrated over the time period 1989–1998, following a one-year warm-up period (1988), and validated over the period 1999–2009. To maintain robust ensemble analyses, the catchments with intermittent flow are removed: the reason behind this choice is that catchments with observations of zero flow present completely different characteristics from any other catchments (see Section 5 for a more detailed discussion). This filtering reduces the dataset from 559, to 422 catchments, with residual analysis restricted to the validation period only.

2.2. Residuals definition and catchments' clustering

In conceptual rainfall–runoff modeling, predictive uncertainty arises from multiple sources, including errors in forcing data (e.g., precipitation, evaporation, discharge measurements), parameter uncertainty, structural deficiencies of the model, and observational noise. Two major methodological families have emerged in this context: joint inference and post-processing, which differ in how model and error components are treated.

Joint inference refers to approaches in which model parameters, error model parameters (e.g., describing autocorrelation or heteroscedasticity), and sometimes even input or observation uncertainties are estimated simultaneously within a single statistical framework. This is typically done through specification of a likelihood function that accounts for hydrological model errors, leading either to maximum likelihood estimates or Bayesian posteriors (Evin et al., 2013, 2014; Kuczera et al., 2006; Renard et al., 2011; Schaefli et al., 2007; Schoups & Vrugt, 2010; Tajiki et al., 2020, and others). The advantage of joint inference lies in its statistical coherence and ability to capture interactions between model and error parameters, but it can suffer from identifiability issues and high computational cost. By contrast, post-processing approaches calibrate the hydrological model first (often assuming a simple error structure), and then apply a statistical correction or residual model a posteriori to account for predictive uncertainty. While less comprehensive, post-processing is computationally efficient and often more robust in practice, particularly when joint inference struggles with

identifiability (Evin et al., 2014). Within this framework, PP methods can be further distinguished into likelihood-based approaches (see Kuczera et al., 2017) and likelihood-free approaches (Montanari & Koutsoyiannis, 2012; Quilty & Adamowski, 2020).

Both joint inference and PP methods differ in whether they disaggregate uncertainty sources (explicitly modeling input, structural, parameter, and observation errors separately, for example see Renard et al., 2010) or aggregate them into a single residual term (e.g., (Schoups & Vrugt, 2010; Shabestanipour et al., 2023). Disaggregation provides valuable diagnostic insights but can suffer from equifinality, whereas aggregation simplifies the estimation at the cost of potentially misattributing uncertainty. While the former allows for a more detailed interpretation of uncertainty origins, the latter has often demonstrated superior performance in estimating prediction intervals (Valdez et al., 2022).

Regardless of whether joint estimation or PP is adopted, all methods ultimately rely on assumptions about the statistical properties of hydrological model residuals. Recent studies have sought to address the complexities of residual behavior by relaxing distributional assumptions, particularly within PP frameworks, through the use of bootstrap techniques (Koutsoyiannis & Montanari, 2022; Shabestanipour et al., 2023; Sikorska et al., 2015). However, even non-parametric resampling approaches still rely on assumptions, such as normality and homoscedasticity, that require verification, underscoring the continued need for careful statistical characterization of model residuals.

In the context of a model, whether hydrological or not, residuals can typically be defined as:

$$\epsilon_{\mathscr{T}}(t) = \mathscr{T}(Q_s(t)) - \mathscr{T}(Q_o(t)) \quad (1)$$

Here, ε represents the model residuals, Q_s (the letter Q here is used to already refer to river discharge) denotes the simulated values from the model, Q_O refers to the observed values, $\mathscr{T}$ indicates a potential transformation, and t is the time index of the model simulation. Various transformations $\mathscr{T}$ for discharge have been explored in the literature, including logarithmic (Shabestanipour et al., 2023), log-sinh (Wang et al., 2012), and Box-Cox transformations (Box & Cox, 1964; McInerney et al., 2017). Such transformations are commonly employed to stabilize residual properties like heteroskedasticity, reducing skewness, or improving adherence to normality assumptions, thereby enabling more robust statistical analyses. For instance, the Box-Cox transformation has been extensively used to correct variance and distributional issues in residuals (McInerney et al., 2017). If no transformation is applied, $\mathscr{T}(x)$=x, and the residuals remain untransformed. In this study, residual analyses are conducted on untransformed residuals in their "natural" space, alongside two transformations applied to the simulated and observed discharge before computing their difference: (a) the logarithmic transformation $\mathscr{T}(x) = \ln(x)$ is adopted due to its widespread application; (b) a Box-Cox transformation with λ=0.5 as previous works (McInerney et al., 2017) suggested it as an improvement to the simpler log-transformation.

To complement the residual analysis, the dataset's catchments are stratified into five distinct hydrological categories based on previously proposed clustering (Brunner et al., 2020). The five clusters are: (1) Intermittent regime (75 catchments): mostly characterized by arid catchments; (2) Weak winter regime (171 catchments): with minimal seasonality but winter-dominated flows; (3) Strong winter regime (88 catchments): showcasing pronounced seasonality and significant winter peaks; (4) Melt regime (51 catchments): dominated by snowmelt, with spring-driven flows; and (5) New Year's regime (37 catchments): featuring strong seasonal flows peaking around the start of the year (Fig. 1). For clarity, the Intermittent Regime (that contained nearly all the intermittent catchment removed from the dataset), is still included due to the presence of arid, but not intermittent, catchments.

3. Statistical Characteristics of Residuals

3.1. Do residuals display seasonality?

River flows are well known to exhibit varying degrees of seasonality, driven by meteorological factors such as precipitation patterns, snow accumulation, and snowmelt. However, the periodic variations in model residuals have not been thoroughly analysed in the literature, despite being of interest for residual characterization, and having been implemented in uncertainty estimation method in the past (McInerney et al., 2020). Seasonality persistence in model residuals is directly related to how strong discharge seasonality is for the considered catchment. In Fig. 2, two examples are reported for the Gallatin River near Gallatin Gateway catchment (on the left, a catchment belonging to the Melt regime, displaying a strong seasonality), and for the Wolf River at Langlade (on the right, an arid catchment belonging to the Intermittent regime instead, displaying a weak seasonality) for Model no.1 for the sub period of 5 years (2005-2009) of the validation period.

It can be immediately noticed how in Fig. 2A, model residuals still show a very clear seasonal component in the signal, less evident instead

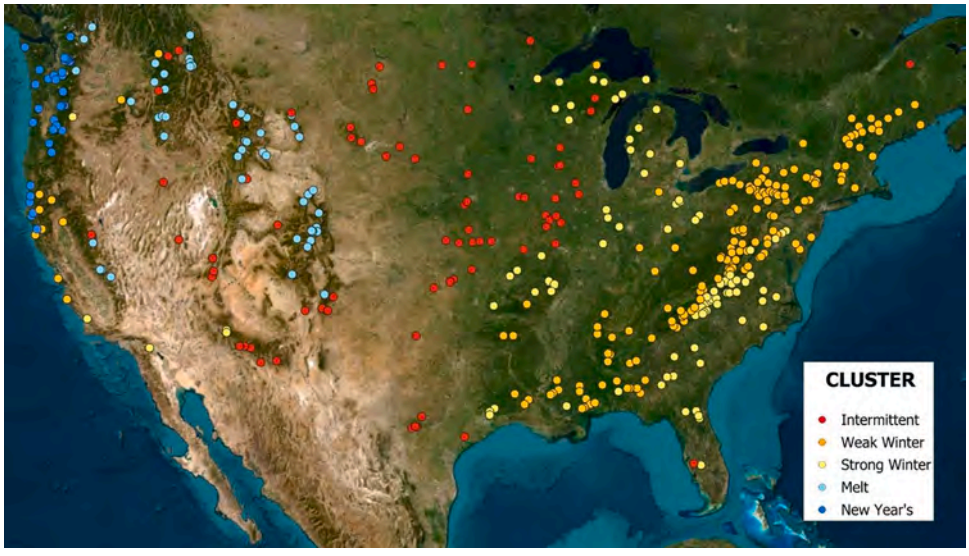


Fig. 1. Filtered CAMEL gauging stations over continental US used in the study. Each colour represents one of five discharge regimes following the classification by Brunner et al., 2020.

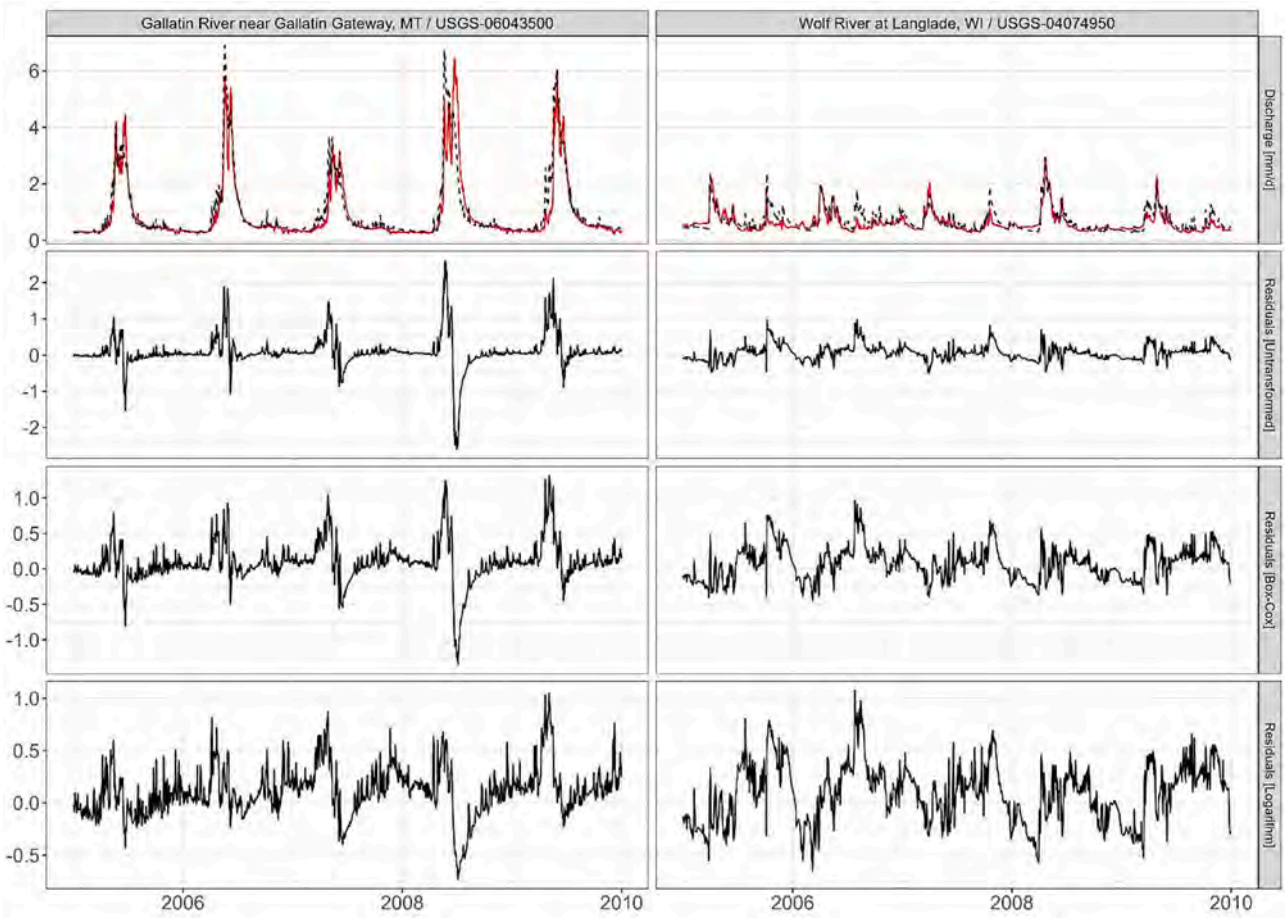


Fig. 2. Observed discharge and simulated discharge (in red), untransformed residuals, residuals after Box-Cox (λ=0.5) transformation, and residuals after logarithmic transformation for two catchments (left - gauge USGS gauge 06043500, Melt Regime; right - USGS gauge 04074950, Intermittent Regime), for Model no. 1.

in Fig. 2B. Moreover, such seasonality is not only relevant for untransformed residuals, but it is also persistent after the application of the Box-Cox and logarithmic transformations. Motivated by this evidence, in order to isolate non-seasonal components, a deseasonalization of the residual time series is performed. The seasonal component is estimated by calculating the mean and standard deviation of residuals for each day of the year across the validation period, then standardizing the residuals accordingly (below for the case of residuals in "natural" space):

$$\widehat{\epsilon}_{d,j} = \frac{\epsilon_{d,j} - \mu_d}{\sigma_d} \tag{2}$$

Where d indicates the day of the year, j the specific year, μ_d the mean and σ_d the standard deviation. The same transformation is also applied analogously to the series resulting after transformation of the observed and simulated discharge variables. This operation is basically a normalization of the residuals on a daily basis. While more sophisticated methods, such as Fourier-based deseasonalization or harmonic regression, could be applied (Sang, 2013; Sang et al., 2016) the simplicity and effectiveness of the chosen method make it a practical choice for this study. From this point forward, the deseasonalized residuals will be referred to as the DS time series.

3.2. What is the probability distribution of model residuals?

Key features of model residuals are their shape properties, which frequently exhibit heavy tails and asymmetry (Evin et al., 2014; Schoups & Vrugt, 2010; Shabestanipour et al., 2023). These characteristics pose significant challenges for statistical modelling, particularly when residuals deviate substantially from normality. A simple and intuitive example is shown in Fig. 3, where, for each catchment, the percentage of days contributing to 50% of the sum-of-squared error (SSE) is reported (see also Clark et al., 2021). The SSE is defined as:

$$\sum_{t=n-k+1}^{n} \widehat{\epsilon}^2(t) = 0.5 \text{ SSE} \tag{3}$$

$$\text{SSE} = \sum_{t=1}^{n} (Q_s(t) - Q_o(t))^2 = \sum_{t=1}^{n} \epsilon^2(t) \tag{4}$$

With n the length of the residual timeseries, $\widehat{\epsilon}^2$ the sorted (in ascending order) squared untransformed residuals and k determining a subset of the largest residuals. Once solved Eq. 3 for k, the corresponding percentage is then easily calculated as k/n. It is immediately clear how just few data points, especially for the westside catchments, can heavily influence the estimate of the MSE, hence the distribution is heavy tailed (see Fig. 3).

Traditionally, shape properties are quantified using product moment ratios (Hosking, 1990), such as skewness (C_s) and kurtosis (C_k). However, in many practical scenarios, the extreme nature of residual distributions can result in infinite or undefined values for C_s and C_k, thereby limiting their applicability (Vogel et al., 2024). This issue is particularly pronounced in hydrological contexts, where strong variability and outliers are common. To overcome these limitations, L-moments (Hosking, 1990) have been developed as a robust alternative for characterizing heavy-tailed distributions. L-moments are less sensitive to extreme values and provide a more stable framework for analysing asymmetry and heavy tails. They are defined as a linear combinations of probability-weighted moments (PWMs – Greenwood et al., 1979;

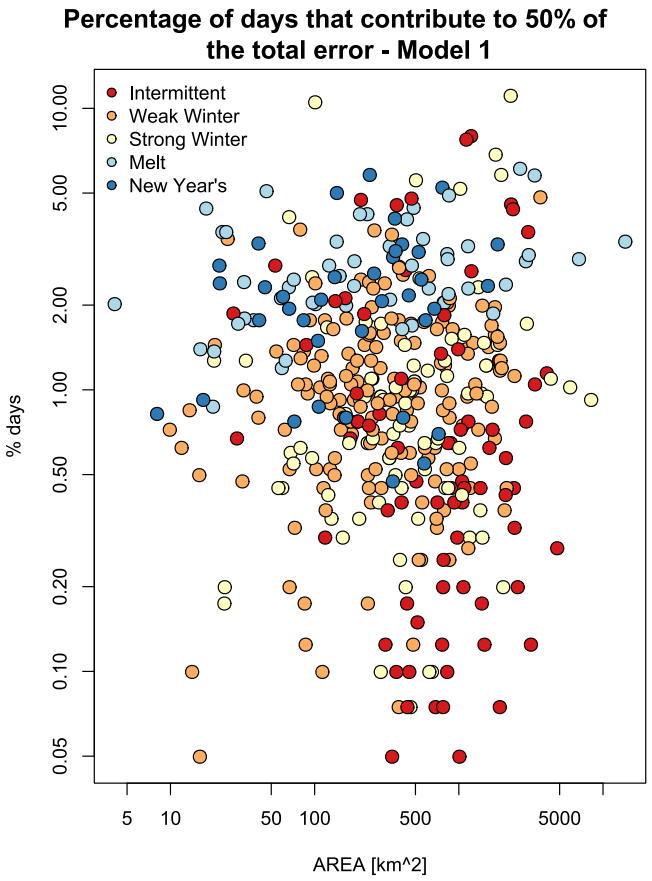


Fig. 3. Percentage of days contributing to 50% of the SSE for the 422 discharge gauge stations (model no.1). Regimes: Intermittent in red, Weak Winter in orange, Strong Winter in yellow, Melt in light-blue, and New Year's in blue.

Sillitto, 1951) and are denoted by the symbol λ (where λ_1 represents the mean, λ_2 is a measure of dispersion in a probability distribution and so on). Analogous to C_s and C_k, L-skewness and L-kurtosis are defined by the ratios $LC_s = \lambda_3/\lambda_2$ and $LC_k = \lambda_4/\lambda_2$, respectively, and retain the same interpretive meaning as their traditional counterparts. Moreover, L-moment diagrams offer a valuable tool for comparing empirical data samples to theoretical distributions.

Fig. 4 illustrates L-moment diagrams for untransformed residuals and residuals computed after the application of transformations. Each point represents median L-moment values across all models applied to a given catchment, ensuring robustness by avoiding reliance on single-point estimates. Being all the models being forced with the same climate forcings dataset, and assuming most of the residuals are explained by input uncertainty, looking at the median tendency per catchment is interesting because it explores a possible space of how the data can be modelled. For untransformed residuals (Fig. 4A), L-kurtosis values frequently exceed 0.4, a threshold suggested in the literature (Vogel et al., 2024) for classifying distributions as extremely heavy-tailed. Such extreme tails can disrupt most traditional statistical analyses. In some catchments, values reach as high as 0.8. These empirical points deviate substantially from the expected curves of theoretical distributions, underscoring the difficulty of modelling such data with conventional methods. By contrast, after the application of transformations (Fig. 4B and Fig. 4C) residuals exhibit substantially lower L-kurtosis, typically remaining below the 0.4 threshold. However, the log transformation, despite further reducing kurtosis with respect to the Box-Cox transformation, also introduces a consistent shift toward negative L-skewness, with values occasionally reaching as low as 0.5.

Following the findings in Section 3.1, L-moments are recalculated also for the DS residual time series. Fig. 4C shows how much of the variability in L-moment ratios values for untransformed residuals was attributable to seasonality rather than inherent distributional characteristics. For untransformed DS residuals, L-kurtosis values drop significantly, often falling below the 0.4 threshold, although they remain higher on average than those for transformed DS residuals after transformation. Similarly, L-skewness values demonstrate a reduced range (also after transformation), further stabilizing the distributions.

Still inside the theory of L-moments, Hosking (Hosking, 1999), suggested, for symmetric distribution, for which all odd order moments are by definition equal to zero, to plot instead λ_6/λ_2 vs λ_4/λ_2 to better capture the behavior of the tails.

Fig. 5 clearly shows how looking at higher order L-moments ratios allow for further distinction between the behavior of symmetric distribution: in particular for DS residuals after transformation, it can be observed how the two transformations, despite both approximating on average a normal distribution, do it in different ways.

3.3. How much does model selection matter in shaping distribution marginal properties?

In Section 3.2, the analysis focused on how potential transformations affected the marginal distributional properties of residuals. However, limited attention was given to the variability of these properties across different models, as Fig. 4 only presented the median L-moment ratios for each catchment. Building on the diverse set of models included in the dataset, this section provides a more detailed investigation of such variability. Specifically, the analysis explores the spread of L-skewness and L-kurtosis values around their median, aiming to evaluate the extent to which the choice of a hydrological model influences the shape properties of residuals. For each catchment, the interdecile range (Q90–Q10) is calculated for both L-skewness and L-kurtosis, for each catchment, across the 78 models. The results, displayed in Fig. 6, are again organized by the different transformations already introduced in Sections 2.2 and 3.1.

Fig. 6 highlights how the variability across models depends on the transformation applied. Untransformed residuals, which exhibited on average L-moments ratios with much heavier than normal tails, generally show substantially lower variability compared to the residuals after log-transformation, but comparable to the Box-Cox transformation. Indeed, for both asymmetry and heavy tails, after log-transformation residuals exhibit the greatest variability among all transformations considered. The behaviour changes considerably for DS residuals, whether untransformed or not. Here, the variability in L-skewness and L-kurtosis, especially the latter, is significantly reduced compared to residuals that retain their seasonal component. In certain cases, such as for the Melt and New Year's regimes, the variability is nearly eliminated, indicating that the influence of model choice on the variability of shape properties is minimal.

Residuals that retain their seasonal component therefore introduce additional challenges: if different models produce markedly varying levels of asymmetry and heavy tails, extreme caution is required in practical applications. The variability in residual properties suggests that switching between models could require the use of different error models based on different assumptions; nevertheless, practitioners often rely on familiar models they trust and understand, and rarely consider alternative models, which may limit their ability to account for such variability.

3.4. Is heteroskedasticity reduced by a data transformation?

A widely discussed property of hydrological model residuals is their relationship with discharge values, whether observed or simulated. The structure, well-documented in hydrological literature (Evin et al., 2014; McInerney et al., 2017; Schoups & Vrugt, 2010; Sorooshian & Dracup, 1980) is well known to be heteroskedastic. Addressing heteroskedasticity is a primary motivation for introducing transformations,

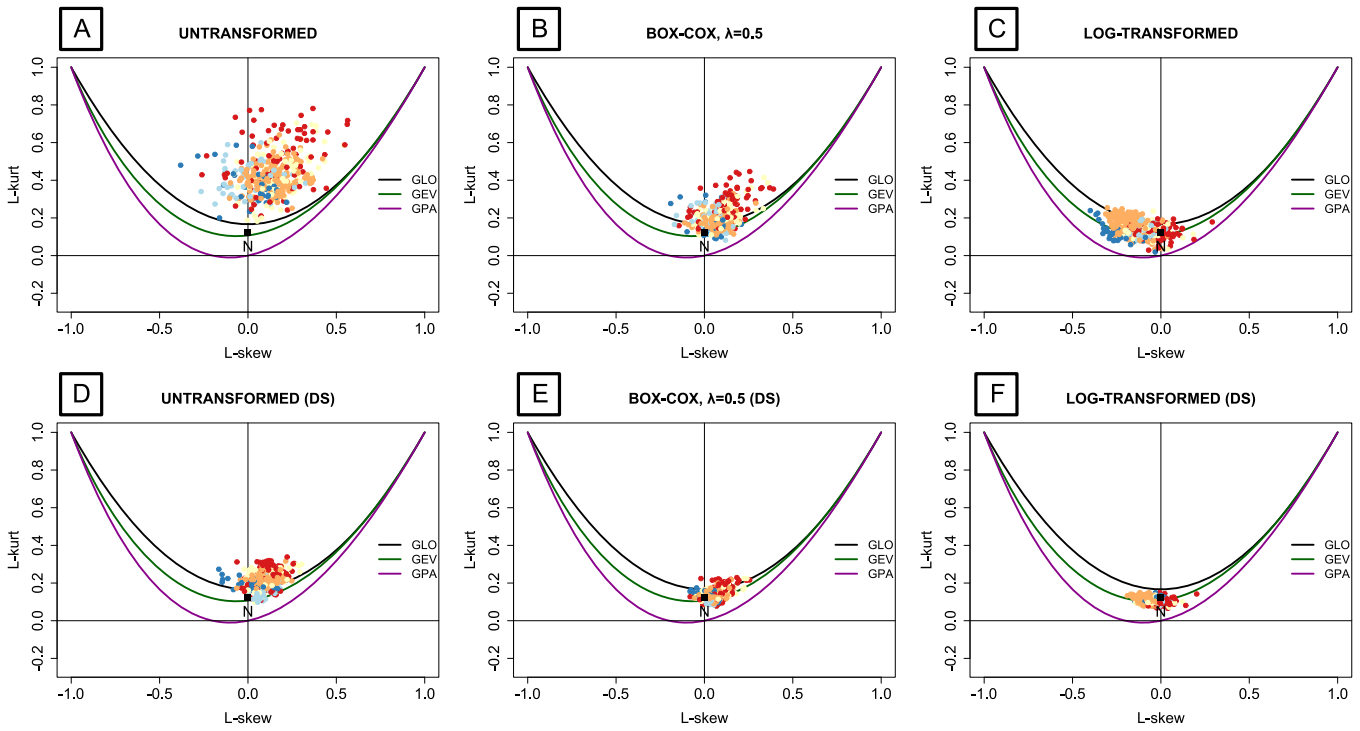


Fig. 4. L-moment diagrams (median L-kurtosis vs median L-skewness across models) for different transformations: (A) Untransformed; (B) Box-Cox (λ=0.5); (C) Log-transformed; (D) Untransformed deseasonalized; (E) Box-Cox (λ=0.5) deseasonalized; (F) Log-transformed deseasonalized. Expected values for the distributions GLO (generalized logistic), GEV (generalized extreme), GPA (generalized Pareto), and Normal (N). Points coloured based on cluster (Intermittent in red, Weak Winter in orange, Strong Winter in yellow, Melt in light-blue, and New Year's in blue). For reference about the expected values for the theoretical distributions see Hosking, 1991.

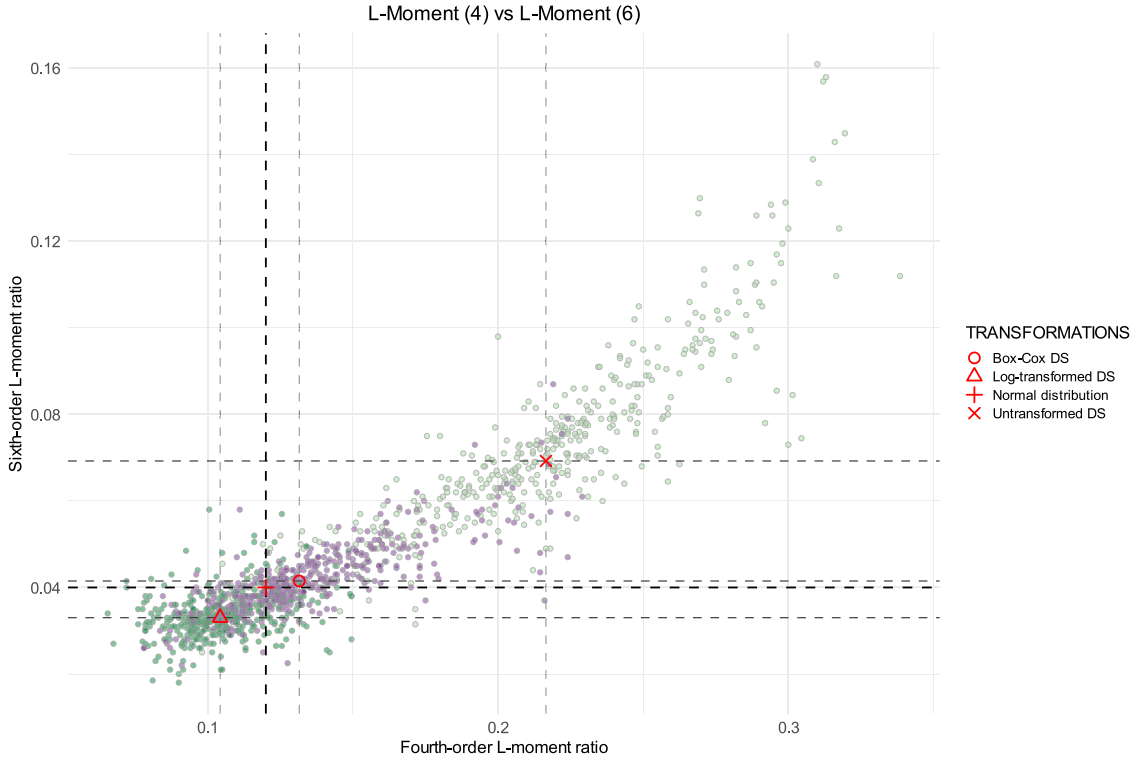


Fig. 5. L-moment diagram (λ_6/λ_2 vs λ_4/λ_2) for deseasonalized untransformed, Box-Cox (λ=0.5) and logarithmic residuals. In red the median L-moments across catchments and models, together with the expected value for a normal distribution.

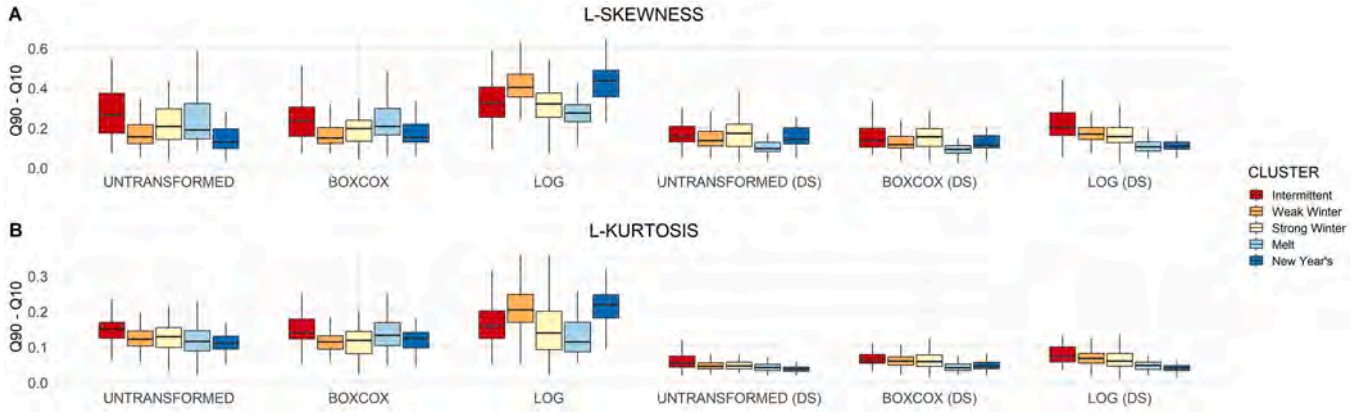


Fig. 6. Interdecile range boxplots of L-moment ratios for untransformed residuals and residuals after transformation, with and without seasonal component, for L-skewness (A) and L-kurtosis (B) across models.

which aim to stabilize variance and approximate a homoscedastic structure, as shown in previous studies (McInerney et al., 2017; Shabestanipour et al., 2023), which highlight the benefits of transformations in improving the interpretability of residuals and reducing variance irregularities. Building on these observations, this section investigates the consistency of heteroskedasticity patterns across catchments, the influence of hydrological clustering, and the effects of removing seasonal components from residual time series. To enable meaningful comparisons, a consistent methodology was applied across all catchment and model combinations:

1. Normalization: each residual time series is normalized by dividing it by its standard deviation. In this way, catchments-model pairs having the same heteroskedastic structure, differing at most for a scaling factor, overlap. In other words, Fig. 7 shows the shape of the heteroskedastic structure, not the absolute values of standard deviation for different ranges of simulated discharges;
2. Rearranging: the simulated discharge values are sorted in increasing order; the same sorting is also used to re-order the residual time series;
3. Standard deviation evaluation: using a moving average (window's size of 200 values), the variance of the newly ordered residual time series is evaluated (in this way a smooth curve is obtained instead of a series of few points for increasing ranges of simulated discharges).

This analysis is conducted for both untransformed residuals and

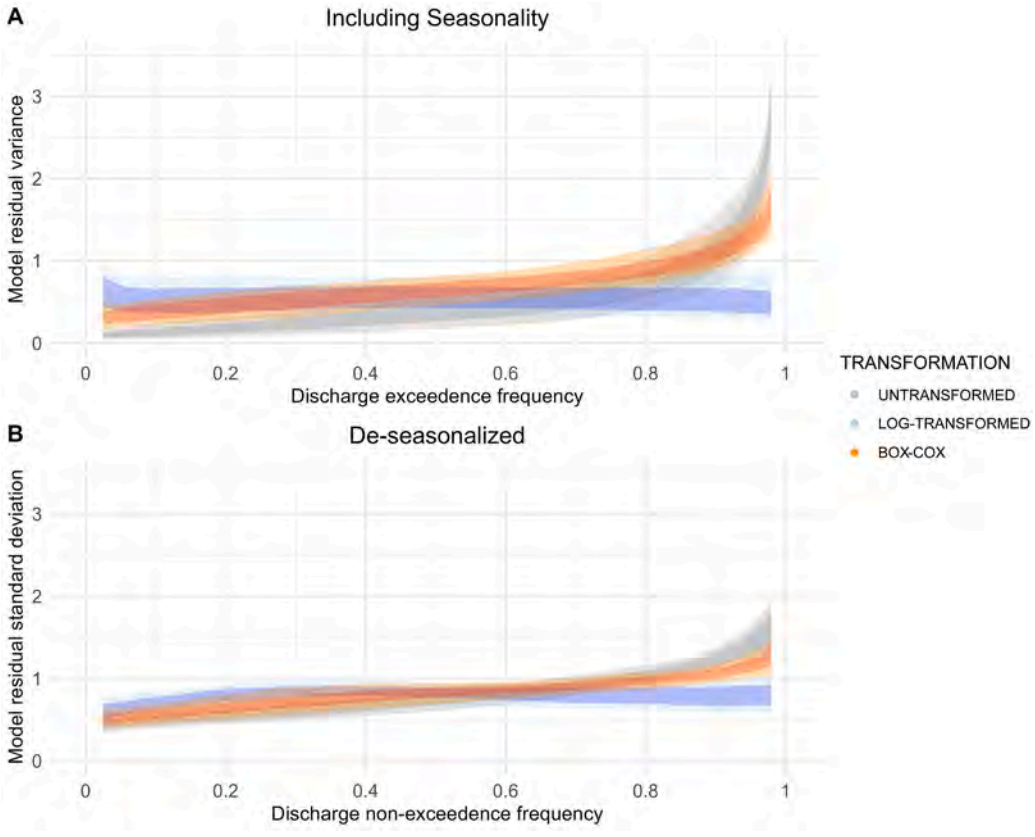


Fig. 7. Model residual standard deviation as function of simulated discharge for untransformed residuals, Box-Cox (λ=0.5) and logarithmic transformation: (A) with seasonality; (B) after deseasonalization.

residuals after transformation, with and without seasonal components. The results, summarized in Fig. 7, depict shaded areas indicating variance interquartile ranges (25th–75th percentiles, darker shades) and interdecile ranges (10th–90th percentiles, lighter shades) across all catchments and models.

For untransformed residuals, a clear pattern emerges: variance increases linearly at lower quantiles before transitioning to exponential-like growth at higher discharge values. Removing seasonality reduces the overall variance and flattens the initial trend, as reflected in the narrower y-axis range. In contrast, after the log-transformation, residuals display nearly constant variance across discharge quantiles, with only slight deviations at extreme quantiles. As for untransformed residuals, removing seasonality further narrows the y-axis range and marginally reduces variance. The effect of the Box-Cox transformation is instead in the middle of the two, displaying heteroskedasticity mitigation, but resulting less effective than the log-transformation.

Variability among hydrological clusters, despite not being reported here, is relatively small, with median trends closely aligned across clusters, except for the Melt regime, which diverges significantly at higher discharge quantiles in all cases (see Appendix A for additional plots). Additionally, differences between hydrological clusters are less pronounced after the application of transformations, reflecting more uniform behaviour across the ensemble. It is important to note that while Fig. 7 represents ensemble averages across models and catchments, individual cases may deviate significantly from these central tendencies, particularly for untransformed residuals where variability is greater. As such, careful attention is warranted when analysing specific cases to account for potential deviations from the general trends.

4. Temporal Correlation Characteristics of the Residuals

4.1. How persistent are the model errors?

Discharge data are widely recognized for their long memory and strong autocorrelation structures, which persist over time-lags much longer than the typical time steps used in hydrological modelling (e.g., daily or hourly). This phenomenon, extensively documented in the hydrological literature, is attributed to the slow dynamics of catchment processes and storage mechanisms (Sorooshian & Dracup, 1980). A key question arises: do similar autocorrelation characteristics persist in model residuals? Specifically, when a model produces an error, does that error persist across subsequent time steps? Furthermore, how much of this autocorrelation is influenced by the seasonal component of the signal?

Residual autocorrelation has been observed in previous studies (McInerney et al., 2020; Shabestanipour et al., 2023; Sorooshian & Dracup, 1980), often analysed using tools such as the autocorrelation function (ACF) and partial autocorrelation function (PACF). These methods measure the correlation of a time series with lagged versions of itself, typically using the Pearson correlation coefficient. However, the application of Pearson correlation in this context has limitations. Pearson's coefficient assumes linearity, normality, and homoscedasticity, assumptions frequently violated by residuals, particularly untransformed or seasonal ones. Barber et al. (2020) documented the poor performance of Pearsons correlation when used with skewed data and they introduce improved estimators of the Pearson correlation coefficient for highly skewed hydrological data. As outlined in Sections 3.2 and 3.3, untransformed residuals often deviate significantly from normality, exhibiting heavy tails and skewness. Even after transformation and deseasonalization, these deviations, while reduced, are not entirely eliminated. Fig. 8 highlights challenges associated with Pearson correlation. Scatterplots at lag-1 for residuals before and after transformation are applied show numerous outliers and unequal dispersion around the 1-to-1 line, particularly in seasonal data. This uneven spread reflects non-linear relationships and heteroskedasticity, limiting the validity of Pearson correlation. To address these issues, alternative correlation measures may be necessary for residual analysis. For example, Kendall or Spearman rank correlation, which evaluates rank order rather than magnitude, may offer robustness to outliers and better captures non-linear relationships.

Fig. 9C and Fig. 9D display a comparison between the median autocorrelation function for the large-ensemble (all catchments and models) calculated with Pearson's correlation (continuous line) and Spearman's correlation (dashed line). The most striking observation is the persistence of autocorrelation across multiple lags for all transformations. This indicates that model errors tend to propagate, with

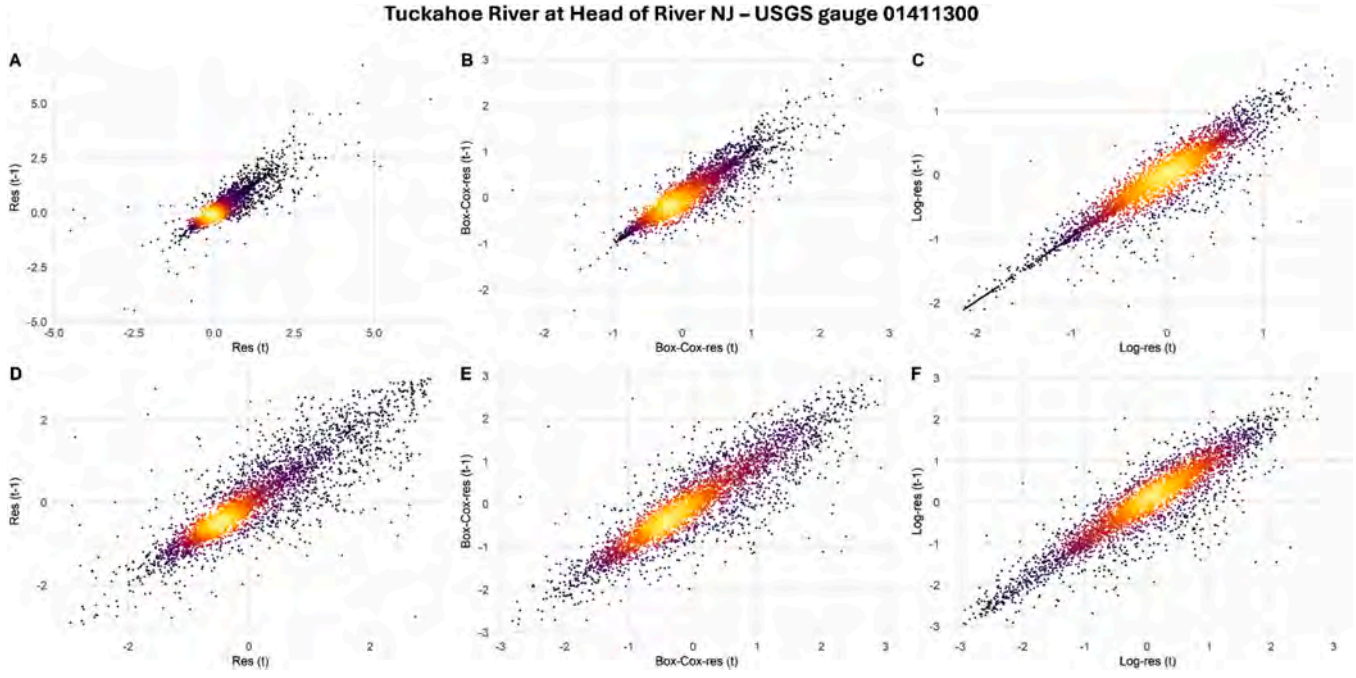


Fig. 8. Scatterplots (t vs t-1) for untransformed residuals (A), Box-Cox (λ=0.5) residuals (B), logarithmic residuals (C) (USGS gauge 01411300 - model no.1). The corresponding deseasonalized residuals are shown in panels D, E, and F. Colour scale indicating points density.

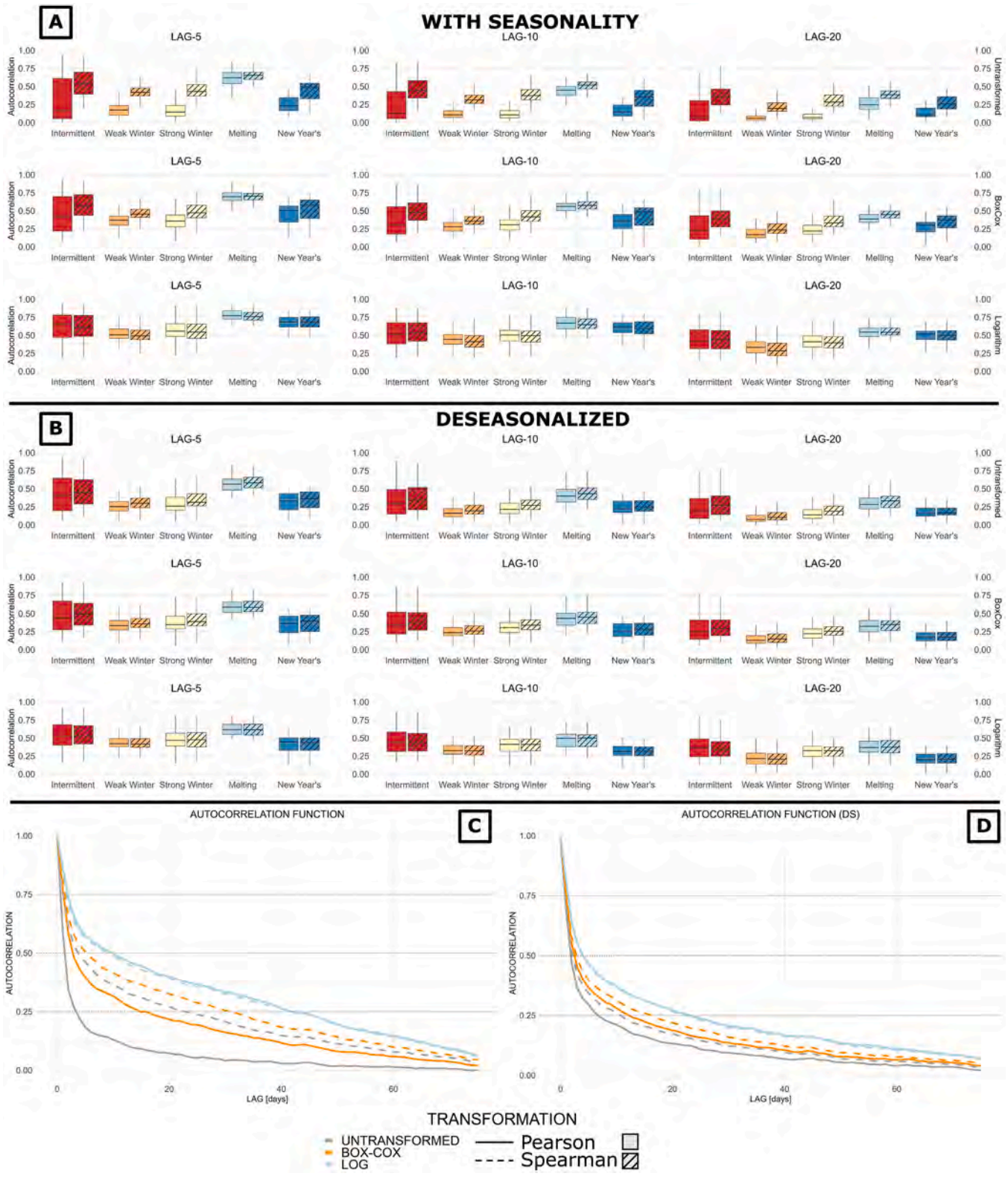


Fig. 9. Median autocorrelation function for the large-ensemble for residuals with (C) and without (D) seasonal components for untransformed, Box-Cox and logarithmic transformations up to lag-75. Boxplots (A and B) represent the distribution at lags 5, 10, and 20 across clusters and transformations. The clusters follow the color scheme described in Section 2.2. Continuous lines (and boxplots) for Pearson correlation, dashed lines (and boxplots) for Spearman correlation.

consecutive overestimations or underestimations persisting over several simulation time steps. Additionally, the different transformations enhance overall correlation. Specifically, moving from untransformed residuals to residuals calculated after the Box-Cox and logarithmic transformations, a noticeable increase in autocorrelation is observed, particularly at lower lags. A similar pattern is also evident for DS residuals, though the effect is less pronounced. The differences between the two correlation metrics are most noticeable for untransformed

residuals where the rank correlation is substantially higher than the sole linear correlation. The increase is observable also for the Box-Cox transformation (to a less degree) while it is nearly negligible for the logarithmic transformation instead. In DS residuals, a general reduction in autocorrelation for the Box-Cox and logarithmic transformations at lower lags is observed, although the correlation remains substantial. For untransformed residuals is instead observed an opposite trend, with a general increase in linear correlation at lower lags. Still in DS residuals, the differences between Pearson and Spearman are negligible.

Examining differences across clusters (Fig. 9A and Fig. 9B – uniform boxplots for Pearson, dashed boxplots for Spearman), it is evident that at lags 5, 10, and 20, all groups exhibit similar behaviors, with the exception of the Melt regime (and to a lower extent the New Year's regime), which consistently shows substantially higher correlation both before and after seasonality removal. In contrast, catchments belonging to the intermittent regime display much greater variability. This variability may be attributed to differences in the amount of "low-flow" days: catchments with a higher frequency of low discharge days exhibit stronger autocorrelation at lower lags due to consecutive close-to-zero values, whereas rivers that experience them with a lower frequency may tend to show lower autocorrelation.

4.2. Temporal Correlation's upper and lower tail dependency

As residual values transition from low to high, their correlation structure may vary substantially. While the correlation coefficient provides useful insights into the overall rank correlation across different lags, tail dependency offers a complementary perspective by focusing on extreme values. Tail dependence coefficients, which quantify the strength of dependence in the upper and lower tails of a distribution, can be estimated non-parametrically as:

$$\lambda_L = \lim_{u \to 0^+} P\left(Y \le F_Y^{-1}(u) \middle| X \le F_X^{-1}(u)\right) \tag{5}$$

$$\lambda_U = \lim_{u \to 1^-} P\left(Y > F_Y^{-1}(u) | X) F_X^{-1}(u)\right) \tag{6}$$

Where X and Y are two variables with marginal quantiles $F_X^{-1}(u)$ and $F_Y^{-1}(u)$, and $P(Y|X)$ denotes conditional probability.

Fig. 10 presents the upper and lower tail dependence coefficients as functions of lag for untransformed residuals and residuals after transformations (for Weak Winter and Melt regimes). These two regimes are selected because they display the most different behaviours. Each of the four panels displays the median upper (top half) and lower (bottom half) tail dependence, along with the interquartile ranges (Q25–Q75). The analysis incorporates all catchments within each cluster and all models for each catchment. Several key observations can be made: (I) Similar to the overall rank correlation described in Section 4.1, both upper and lower tail dependence exhibit persistence across the time series up to lags equal or shorter of 25–35 days for most cases; (II) the tail dependence structure is largely symmetric, with only a slight tendency for the upper tail dependence to be higher than the lower tail dependency; (III) for both seasonal and DS series, the tail dependence coefficients after logarithmic transformation are generally higher, or equal to those of untransformed residuals, while after Box-Cox transformation in several cases tail correlation is instead reduced; (IV) the effect of deseasonalization is more pronounced for smaller lags, but not so much for larger ones. Although it is impractical to present results for all clusters here, a full collection is available in Appendix A.

These findings, along with those reported in Section 4.1, underscore how residual correlation structures are not uniform across the entire range of values. Consequently, capturing the full behaviour of residual autocorrelation requires approaches that extend beyond single-coefficient summaries (e.g., individual correlation coefficients). One possibility is modelling autocorrelation through bivariate distributions, such as copula models. Fig. 11 shows scatterplots analogous to those in Fig. 8, focusing on lag-1 autocorrelation, but for pseudo-observations, as used in copula analyses. Pseudo-observations are rank-based transformations of data, typically derived from empirical cumulative distribution functions (ECDFs), that map original values to a (0,1) scale, enabling non-parametric dependence modelling.

A key observation is that the pronounced differences evident in Fig. 8 are no longer present in Fig. 11. Instead, the residuals exhibit a relatively consistent structure, regardless of whether they were transformed or not. However, the removal of seasonality exerts a clear influence: while values previously clustered along the 1-to-1 line, they become more dispersed after deseasonalization.

Using the VineCopula package in R (Nagler et al., 2024), copula models are fitted to catchment-model combinations for DS cases (not shown here), without imposing predefined structures, allowing the algorithm to select the best-fitting model. The t-copula emerged as the most suitable model in over 98% of cases, reflecting its capacity to capture the observed symmetry in upper and lower tail dependencies. However, goodness-of-fit (GOF) tests revealed that the models are statistically non-significant in most instances. While the t-copula captures

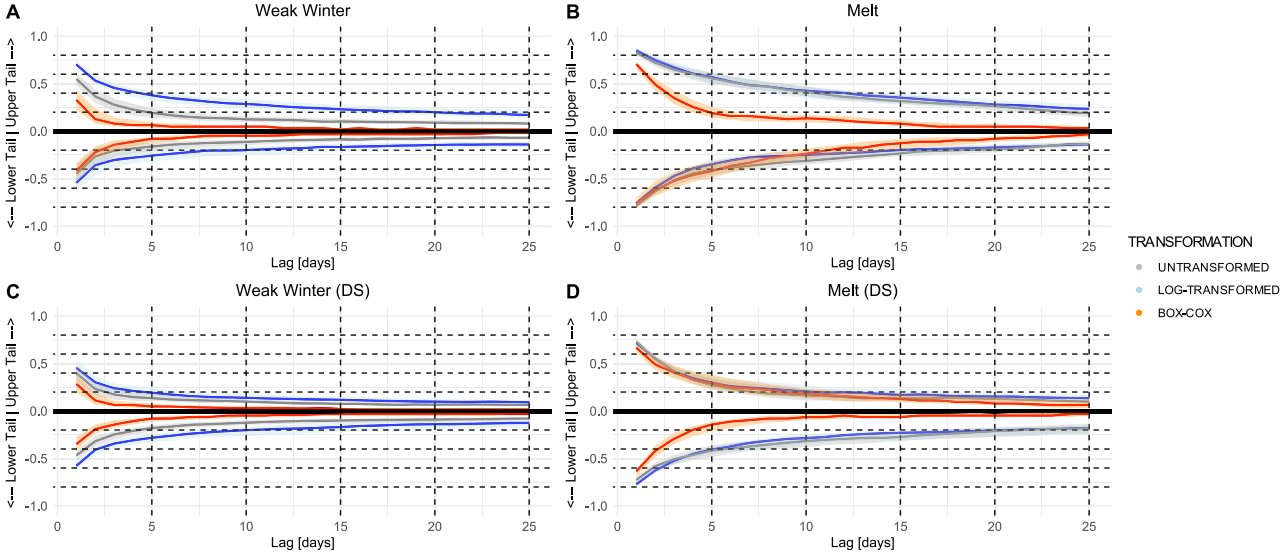


Fig. 10. Upper and lower temporal tail correlation up to lag 25 for the different transformations, with and without seasonality, for Weak winter and Melt regimes.

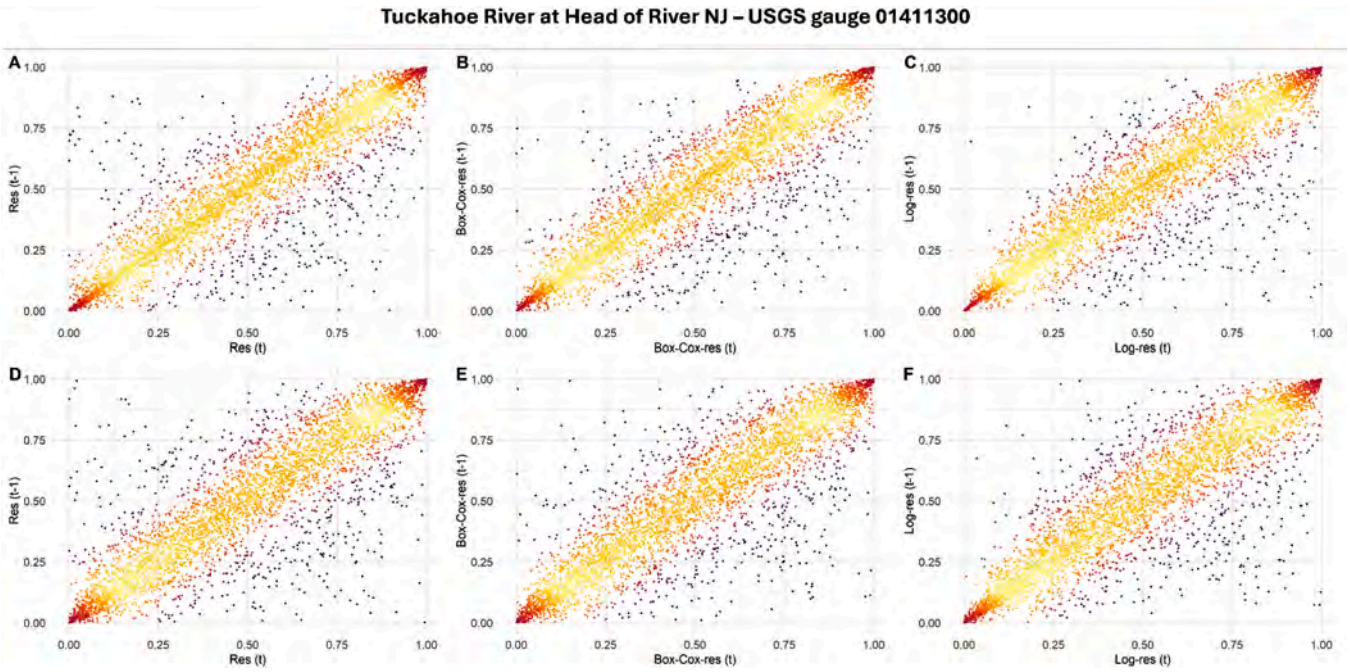


Fig. 11. Scatterplots (t vs t-1) for pseudo untransformed residuals (A), pseudo Box-Cox (λ=0.5) residuals (B), pseudo logarithmic residuals (C) (USGS gauge 01411300 - model no.1). The corresponding deseasonalized residuals are shown in panels D, E, and F. Colour scale indicating points density.

the general structure of the data, it falls short of fully representing its complexity. Significant deviations persist, notably instances where low ranks are followed by high ranks (or vice versa) across successive time steps.

5. Discussion

5.1. Difference between regimes and general effect of transformations

This study presents a comprehensive analysis of the statistical properties of conceptual rainfall-runoff models residuals, with the goal of improving the understanding of residuals structures and their implications for uncertainty estimation.

Regarding the marginal distribution properties (Section 3.2), the introduction of transformations and normalization clearly shows beneficial effects in reducing tail heaviness. At the same time, some regime-specific differences emerge, particularly for the log transformation (which displayed L-kurtosis values closer to those expected for a normal distribution). In this case, a shift toward negative L-skewness values is observed; however, a closer inspection reveals that this effect is confined to the Weak Winter and New Year's regimes. After deseasonalization, catchments' residuals in the New Year's regime also approximate a symmetric distribution, while those belonging to the Weak Winter regime still tend to display negative asymmetry values. In terms of variability across models (Section 3.3), no marked differences among hydrological regimes are evident. Nonetheless, deseasonalization effectively reduces the variability introduced by model selection.

The heteroskedastic structure (noting that in Section 3.4 only its shape was considered) is highly consistent across regimes (with the sole exception of the Melt regime, see Appendix A), showing a nonlinear increase with simulated discharge quantiles, although differences in the absolute values of standard deviation remain possible across regimes. In agreement with previous findings in the literature, the considered transformations adequately address heteroskedasticity, effectively approximating a homoscedastic structure, with the logarithmic transformation providing the best approximation. A previous study (McInerney et al., 2017) suggested that the Box–Cox transformation could improve upon the logarithmic one; however, in that case the authors proposed as Box-Cox parameter λ values between 0 and 0.2, which are closer to the logarithmic transformation than the $\lambda = 0.5$ adopted in this study. Overall, reducing λ progressively moves the residuals toward a homoscedastic structure, suggesting that in specific cases a sensitivity analysis on the Box–Cox parameter could help optimize the transformation. Deseasonalization in this context reduces overall variability but does not produce noticeable differences across hydrological regimes (median values for all regimes are provided in Appendix A).

Autocorrelation (Section 4.1) is present in the residuals regardless of the transformation applied. For the non-normalized residuals, particularly at lower lags, some differences across hydrological regimes are evident (especially for Pearson correlation): catchments in the Intermittent regime exhibit the largest variability among all regimes, with lag-5 correlations ranging from near 0 to values above 0.6 (Fig. 9A). Conversely, catchments belonging to the Melt regime, again most clearly at lower lags, display a systematic tendency toward higher autocorrelation values than any other regime. The use of a rank based correlation coefficient (Spearman) highlighted a noticeable increase in temporal correlation for untransformed residuals, while for both Box-Cox and logarithmic transformations, as well as for deseasonalized residuals, the difference between the two metrics is less relevant. Transformations tend to increase overall autocorrelation, while deseasonalization has only a limited effect. The same trends are noticeable also for upper and lower temporal tail correlations, where again the Melt regime (Fig. 10B and Fig. 10D) display overall higher correlation than the other regimes (for a complete collection of Figures see Appendix A).

5.2. General insights for modelling

Being able to generate realistic ensembles of residuals would fundamentally transform hydrological modelling: instead of a single deterministic trajectory, each model run could be expanded into a full ensemble, providing a direct and quantitative basis for uncertainty assessment and scenario exploration. Yet designing such stochastic models is far from trivial, because residuals do not behave like simple white noise. They carry signatures of seasonality, strong autocorrelation, cross-dependencies (in this study the investigation was limited to

the heteroskedastic structure), and non-Gaussian distributional shapes, features that are not constant but may vary across hydrological regimes. Moreover, the deviations in scatterplots in Fig. 11, suggest that the residuals may result from at least two overlapping processes: one that governs the primary structure and another introducing anomalies. Future research should investigate multi-process models capable of disentangling these effects, thereby enhancing our understanding of residual structures and their implications for hydrological predictions. This study makes a step forward by systematically mapping these properties across a uniquely large multi-model and multi-catchment dataset, allowing us to distinguish common structures from regime-specific differences. These insights are critical: they guide which aspects of residual behaviour must be captured explicitly, and which can be simplified without major loss of realism. They also reveal the limits of one-size-fits-all approaches, highlighting the need for general yet flexible stochastic frameworks, such as CoSMoS (Papalexiou, 2018), that can accommodate seasonality, variance stabilization, temporal dependence, and tail behaviour in a unified way.

5.3. Note about deseasonalization

Accounting for the persistence of seasonality in residual time series clearly improves the stabilization and generalization of residual properties. In the present study, a direct transformation approach was proposed, but alternative strategies could also be considered. Examples include calibrating rainfall–runoff model parameters at monthly or seasonal scales to reduce seasonality in residuals at the source, or developing error models with time-varying parameters. Regardless of whether a direct or indirect approach is adopted, an important issue remains: when attempting to explicitly account for seasonality, the available record length must effectively be reduced. For instance, estimating twelve distinct parameter sets, one for each month, requires partitioning the observed series into twelve sub-samples, each with only one-twelfth of the original data, which are then used for calibration. This raises a key question: how many years of data are required for robust estimation? In this study, a record length of ten years was considered. Whether this is sufficient for reliable parameter estimation is questionable. In general, this situation poses a dilemma: while seasonal approaches offer clear benefits, their applicability is constrained by data availability. There is no straightforward solution, apart from emphasizing the critical importance of long-term data collection and advocating for sustained efforts to maintain and expand discharge monitoring networks, which historically remain less extensive than those for other key hydrological variables such as precipitation and temperature.

5.4. Note about intermittent catchments

A general remark concerns intermittent catchments. As noted in Section 2.1, although the original dataset included temporary rivers, these were excluded to ensure consistency in the analysis. Arid catchments are already notoriously challenging to model, and the presence of intermittency adds an additional layer of complexity. Consequently, their statistical properties, including those of residuals, are expected to be harder to characterize and more variable, as is already partially evident from the results of the remaining catchments classified within the Intermittent regime (see, for instance, Figs. 9A and 9B in Section 4.1). A more extensive analysis would therefore be required to provide a comprehensive description. Furthermore, when developing an error model, intermittent catchments impose an additional constraint: the need to correctly reproduce zero-flow statistics. Addressing this requirement likely demands a mixed-type modelling strategy (similar to those proposed for stochastic rainfall generation), combining one component to represent non-zero flows with another designed to capture the observed degree and temporal structure of intermittency.

5.5. Limits in our approach and future developments

The dataset used in this study, as noted in the introduction, is significantly larger than any previously employed for systematic residual analysis. While it provides extensive coverage of climatic and morphological catchment variability and includes a broad set of models, some caveats are necessary. Clearly, 78 models cannot encompass the full range of possible model architectures and parameterizations; nevertheless, this number was sufficient to highlight key features of model-induced variability in residual properties and to evaluate the effectiveness of the proposed transformations in reducing such variability. Moreover, these models are built from four existing models that are already conceptually quite different (PRMS, Sacramento, VIC, TOPMODEL). The simplifying equations used to represent the processes in these four base models are distinct, and their structural design is as well. The framework can then be considered quite general. Additionally, previous studies, though limited to fewer catchments and narrower sets of properties, have generally employed models outside the FUSE framework. Their results, however, are consistent with the broader findings reported here, at least for the properties shared between the studies.

Another layer of variability not explicitly addressed in this study concerns the role of the model efficiency (ME) function used during calibration. Here, parameters were tuned exclusively with the Kling–Gupta Efficiency (KGE), while other works have also employed different metrics, most notably the Nash–Sutcliffe Efficiency (NSE). However they reported residual properties similar to those observed here. A notable example is Hunter et al. (2021), who specifically investigated the influence of ME selection, although with a focus more on evaluating error-model performance than on a detailed characterization of residual properties.

Finally, an important research gap relates to the scope of the present analysis. This study focused on a purely statistical characterization of aggregated residuals and assessed model selection effects only through an ensemble perspective, considering model outputs without examining in detail how specific structural features might drive residual properties. A more process-oriented approach, directly linking model structure and parameterization to residual behavior, could provide valuable insights, particularly for guiding architectural choices toward models that generate better-behaved residuals. However, such an endeavor would require substantial additional effort beyond the scope of a single study. Future research could effectively address both the influence of calibration metrics and the process-based connections between model structure and residual characteristics.

6. Conclusions

Here we show that aggregated residuals of conceptual hydrological models exhibit consistent yet nuanced statistical behaviours across a wide range of catchments and model structures. Specifically, we assesses for the first time a comprehensive set of residual properties stratified across hydrological regimes, considering model induced variability and systematically mapping the effect of transformations. Transformations and deseasonalization are required to stabilize residual properties, with the former playing a key role in stabilizing variance and the latter in reducing heavy tails as it became apparent from the higher-order L-moment properties we investigated. Despite these adjustments, residuals retain temporal dependence, and regime-specific differences persist, most notably in skewness/kurtosis and autocorrelation patterns. These findings highlight the aspects of residual behaviour that must be accounted for explicitly in stochastic error models, while also indicating where simplifications remain defensible. By consolidating insights across a uniquely large multi-model and multi-catchment dataset, this work provides a practical foundation for the construction of generalizable residual models. Embedding these diagnostics into flexible stochastic frameworks opens the way for hydrological models to move

beyond single deterministic trajectories toward ensembles of realistic error sequences, ultimately supporting more reliable uncertainty assessment and risk-informed decision making.

Data Availability Statement

Thébault, C. (2025). FUSE simulations, HydroShare, http://www.hydroshare.org/resource/d450bfdfdb944548ac0577f04dd60529

Funding

The participation of Luca Lombardo and Alberto Viglione in this study was carried out within the RETURN Extended Partnership and received funding from the European Union Next-GenerationEU (National Recovery and Resilience Plan –NRRP, Mission 4, Component 2, Investment 1.3 – D.D. 1243 2/8/2022, PE0000005) and from the Italian Ministry of University and Research (PRIN project n. 2022AX3882 -Clim2FlEx - Mapping of climate to flood extremes). This research was supported by the Cooperative Institute for Research to Operations in Hydrology (CIROH) with funding under award NA22NWS4320003 from the NOAA Cooperative Institutes programme. The statements, findings, conclusions and recommendations are those of the author(s) and do not necessarily reflect the opinions of NOAA.

CRediT authorship contribution statement

Luca Lombardo: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Simon Michael Papalexiou:** Writing – review & editing, Writing – original draft, Supervision, Formal analysis, Conceptualization. **Cyril Thébault:** Writing – review & editing, Data curation. **Martyn P. Clark:** Writing – review & editing, Data curation. **Richard M. Vogel:** Writing – review & editing. **Alberto Viglione:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare no conflicts of interest relevant to this study.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.advwatres.2025.105165.

References

Addor, N., Newman, A.J., Mizukami, N., Clark, M.P., 2017. The CAMELS data set: catchment attributes and meteorology for large-sample studies. Hydrology and Earth System Sciences 21 (10), 5293–5313. https://doi.org/10.5194/hess-21-5293-2017.

Barber, C., Lamontagne, J.R., Vogel, R.M., 2020. Improved estimators of correlation and R2 for skewed hydrologic data. Hydrological Sciences Journal 65 (1), 87–101. https://doi.org/10.1080/02626667.2019.1686639.

Beven, K. (2012). *Rainfall-Runoff Modelling: The Primer. Rainfall-Runoff Modelling: The Primer: Second Edition* (Vol. 15). https://doi.org/10.1002/9781119951001.

Bourgin, F., Ramos, M.H., Thirel, G., Andréassian, V., 2014. Investigating the interactions between data assimilation and post-processing in hydrological ensemble forecasting. Journal of Hydrology 519, 2775–2784. https://doi.org/10.1016/j.jhydrol.2014.07.054.

Box, G.E.P., Cox, D.R., 1964. An Analysis of Transformations. J. Royal Stat. Soc. Series B (Methodological), 26 (2), 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x.

Brunner, M., Melsen, L.A., Newman, A., Wood, A., Clark, M.P., 2020. Future streamflow regime changes in the United States: assessment using functional classification. Hydrology and Earth System Sciences 24, 3951–3966. https://doi.org/10.5194/hess-24-3951-2020.

Clark, M., Nijssen, B., Lundquist, J., Kavetski, D., Rupp, D., Woods, R., et al., 2015. A unified approach for process-based hydrologic modeling: 1. Modeling concept. Water Resources Research 51. https://doi.org/10.1002/2015WR017198.

Clark, M.P., Slater, A.G., Rupp, D.E., Woods, R.A., Vrugt, J.A., Gupta, H.V., et al., 2008. Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. Water Resources Research 44 (12). https://doi.org/10.1029/2007WR006735.

Clark, M.P., Kavetski, D., Fenicia, F., 2011. Pursuing the method of multiple working hypotheses for hydrological modeling. Water Resources Research 47 (9). https://doi.org/10.1029/2010WR009827.

Clark, M.P., Vogel, R.M., Lamontagne, J.R., Mizukami, N., Knoben, W.J.M., Tang, G., et al., 2021. The Abuse of Popular Performance Metrics in Hydrologic Modeling. Water Resources Research 57 (9), e2020WR029001. https://doi.org/10.1029/2020WR029001.

Duan, Q., Sorooshian, S., Gupta, V., 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resources Research 28 (4), 1015–1031. https://doi.org/10.1029/91WR02985.

Evin, G., Kavetski, D., Thyer, M., Kuczera, G., 2013. Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration. Water Resources Research 49, 4518–4524. https://doi.org/10.1002/wrcr.20284.

Evin, G., Thyer, M., Kavetski, D., McInerney, D., Kuczera, G., 2014. Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity. Water Resources Research. https://doi.org/10.1002/2013WR014185.

Greenwood, J.A., Landwehr, J.M., Matalas, N.C., Wallis, J.R., 1979. Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form. Water Resources Research 15 (5), 1049–1054. https://doi.org/10.1029/WR015i005p01049.

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of Hydrology 377 (1), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003.

Hosking, J.R.M., 1990. L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. Journal of the Royal Statistical Society: Series B (Methodological) 52 (1), 105–124. https://doi.org/10.1111/j.2517-6161.1990.tb01775.x.

Hosking, J.R.M., 1991. Approximations for Use in Constructing L-moment Ratio Diagrams. IBM Research Division, T.J. Watson Research Center.

Hosking, J.R.M., 1999. L-moments and their applications in the analysis of financial data. Research Report RC21466. IBM Research Division. Yorktown Heights, New York, pp. 13.–p.

Hrachowitz, M., Clark, M.P., 2017. HESS Opinions: The complementary merits of competing modelling philosophies in hydrology. Hydrology and Earth System Sciences 21 (8), 3953–3973. https://doi.org/10.5194/hess-21-3953-2017.

Hunter, J., Thyer, M., McInerney, D., Kavetski, D., 2021. Achieving high-quality probabilistic predictions from hydrological models calibrated with a wide range of objective functions. Journal of Hydrology 603, 126578. https://doi.org/10.1016/j.jhydrol.2021.126578.

Jakeman, A.J., Hornberger, G.M., 1993. How much complexity is warranted in a rainfall-runoff model? Water Resources Research 29 (8), 2637–2649. https://doi.org/10.1029/93WR00877.

Kavetski, D., Kuczera, G., Franks, S., 2006. Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. Water Resources Research 42, W03407. https://doi.org/10.1029/2005WR004368.

Knoben, W.J.M., Freer, J.E., Peel, M.C., Fowler, K.J.A., Woods, R.A., 2020. A Brief Analysis of Conceptual Model Structure Uncertainty Using 36 Models and 559 Catchments. Water Resources Research 56 (9), e2019WR025975. https://doi.org/10.1029/2019WR025975.

Koutsoyiannis, D., Montanari, A., 2022. Bluecat: A Local Uncertainty Estimator for Deterministic Simulations and Predictions. Water Resources Research 58. https://doi.org/10.1029/2021WR031215.

Kuczera, G., Kavetski, D., Franks, S., Thyer, M., 2006. Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. Journal of Hydrology 331 (1), 161–177. https://doi.org/10.1016/j.jhydrol.2006.05.010.

McInerney, D., Thyer, M., Kavetski, D., Lerat, J., Kuczera, G., 2017. Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. Water Resources Research 53. https://doi.org/10.1002/2016WR019168.

McInerney, D., Thyer, M., Kavetski, D., Laugesen, R., Tuteja, N., Kuczera, G., 2020. Multi-temporal Hydrological Residual Error Modeling for Seamless Subseasonal Streamflow Forecasting. Water Resources Research 56. https://doi.org/10.1029/2019WR026979.

Montanari, A., Koutsoyiannis, D., 2012. A blueprint for process-based modeling of uncertain hydrological systems. Water Resources Research 48 (9). https://doi.org/10.1029/2011WR011412.

Nagler, T., Schepsmeier, U., Stoeber, J., Brechmann, E. C., Graeler, B., Erhardt, T., et al. (2024, December 17). VineCopula: Statistical Inference of Vine Copulas (Version 2.6.0). Retrieved from https://cran.r-project.org/web/packages/VineCopula/index.html.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — A discussion of principles. Journal of Hydrology 10 (3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.

Newman, A., Clark, M., Sampson, K., Wood, A., Hay, L., Bock, A., et al., 2015. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. Hydrology and Earth System Sciences 19, 209–223. https://doi.org/10.5194/hess-19-209-2015.

Papalexiou, S.M., 2018. Unified theory for stochastic modelling of hydroclimatic processes: Preserving marginal distributions, correlation structures, and intermittency. Advances in Water Resources 115, 234–252. https://doi.org/10.1016/j.advwatres.2018.02.013.

Perrin, C., Michel, C., Andréassian, V., 2003. Improvement of a parsimonious model for streamflow simulation. Journal of Hydrology 279 (1), 275–289. https://doi.org/10.1016/S0022-1694(03)00225-7.

Quilty, J., Adamowski, J., 2020. A stochastic wavelet-based data-driven framework for forecasting uncertain multiscale hydrological and water resources processes. Environmental Modelling & Software 130, 104718. https://doi.org/10.1016/j.envsoft.2020.104718.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., Franks, S.W., 2010. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. Water Resources Research 46 (5). https://doi.org/10.1029/2009WR008328.

Renard, B., Kavetski, D., Leblois, E., Thyer, M., Kuczera, G., Franks, S.W., 2011. Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation. Water Resources Research 47 (11). https://doi.org/10.1029/2011WR010643.

Sang, Y., 2013. A review on the applications of wavelet transform in hydrology time series analysis. Atmospheric Research 122, 8–15. https://doi.org/10.1016/j.atmosres.2012.11.003.

Sang, Y., Singh, V., Sun, F., Yaning, C., Liu, Y., Yang, M., 2016. Wavelet-Based Hydrological Time Series Forecasting. Journal of Hydrologic Engineering 21, 06016001. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001347.

Schaefli, B., Talamba, D.B., Musy, A., 2007. Quantifying hydrological modeling errors through a mixture of normal distributions. Journal of Hydrology 332 (3–4), 303–315. https://doi.org/10.1016/j.jhydrol.2006.07.005.

Schoups, G., Vrugt, J., 2010. A Formal Likelihood Function for Parameter and Predictive Inference of Hydrologic Models With Correlated, Heteroscedastic, and Non-Gaussian Errors. Water Resources Research 46. https://doi.org/10.1029/2009WR008933.

Shabestanipour, G., Brodeur, Z., Farmer, W., Steinschneider, S., Lamontagne, J., 2023. Stochastic Watershed Model Ensembles for Long-Range Planning: Verification and Validation. Water Resources Research 59. https://doi.org/10.1029/2022WR032201.

Sikorska, A.E., Montanari, A., Koutsoyiannis, D., 2015. Estimating the Uncertainty of Hydrological Predictions through Data-Driven Resampling Techniques. Journal of Hydrologic Engineering 20 (1), A4014009. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000926.

Sillitto, G.P., 1951. Interrelations Between Certain Linear Systematic Statistics of Samples from Aany Continuos Population. Biometrika 38 (3–4), 377–382. https://doi.org/10.1093/biomet/38.3-4.377.

Sorooshian, S., Dracup, J.A., 1980. Stochastic Parameter Estimation Procedures for Hydrologie Rainfall-Runoff Models: Correlated and Heteroscedastic Error Cases. Water Resources Research 16, 430–442. https://doi.org/10.1029/WR016i002p00430.

Tajiki, M., Schoups, G., Hendricks Franssen, H.J., Najafinejad, A., Bahremand, A., 2020. Recursive Bayesian Estimation of Conceptual Rainfall-Runoff Model Errors in Real-Time Prediction of Streamflow. Water Resources Research 56 (2), e2019WR025237. https://doi.org/10.1029/2019WR025237.

Valdez, E.S., Anctil, F., Ramos, M.-H., 2022. Choosing between post-processing precipitation forecasts or chaining several uncertainty quantification tools in hydrological forecasting systems. Hydrology and Earth System Sciences 26 (1), 197–220. https://doi.org/10.5194/hess-26-197-2022.

Vogel, R.M., 2017. Stochastic watershed models for hydrologic risk management. Water Security 1, 28–35. https://doi.org/10.1016/j.wasec.2017.06.001.

Vogel, R.M., Papalexiou, S.M., Lamontagne, J.R., Dolan, F.C., 2024. When Heavy Tails Disrupt Statistical Inference. The American Statistician 1–15. https://doi.org/10.1080/00031305.2024.2402898.

Wang, Q.J., Shrestha, D.L., Robertson, D.E., Pokhrel, P., 2012. A log-sinh transformation for data normalization and variance stabilization. Water Resources Research 48 (5). https://doi.org/10.1029/2011WR010973.