Summary of the Revised Version

We sincerely thank the referees for their careful and constructive comments. We have substantially revised the manuscript to clarify the scope of the temporal-difference loss regularization (TDLR) framework, improve its logical flow, make the benchmark comparison more transparent, and more carefully distinguish model-derived functional representations from physically verified hydrological processes.

The major revisions are summarized as follows:

1. Reframing of process identification and identifiability

   We have revised the manuscript throughout to clarify that TDLR does not establish physically unique process functions from streamflow data alone. The extracted representations are now explicitly presented as robust and testable mechanistic hypotheses, conditional on the prescribed bucket structure, mass-balance constraints, input data, and output-range control. We further distinguish optimization/re-extraction stability from physical identifiability. The conclusion now explicitly states that the true catchment-scale response functions cannot be uniquely determined from the available observations.

2. Clarification of the role of TDLR in the closure problem.

   We reframed TDLR from solving or uniquely identifying the closure problem to conditionally searching for admissible catchment-scale functional representations as a hypothesis-generation step. The revised framework emphasizes an iterative workflow in which candidate representations are evaluated against hydrograph reconstruction, mass balance, repeated extraction, independent observations, and catchment information.

3. Improved benchmark fairness and information-use comparison.

   We added pure LSTM benchmarks both with and without observed Qt−1, allowing a direct comparison between TDLR and LSTM under the same information setting. We also added a fully recursive/free-running TDLR experiment using estimated Qt−1. The revised Table 1 reports both NSE and KGE, and the related discussion is consolidated in Sect. 4.1.

4. More cautious interpretation of the mass-balance residual.

   We now explicitly describe Lt as a residual that may contain interbasin groundwater exchange, measurement error, and structural mismatch. We added a ±20% precipitation sensitivity analysis and explicitly state that the observed residual cannot uniquely distinguish inter-basin groundwater flow (IGF) from persistent measurement bias or structural error. We therefore present IGF as a mechanistic hypothesis rather than an identified process and specify the independent observations required for direct testing.

5. Correct the Jiji catchment delineation and retrain all Jiji models.

While re-examining the Jiji dam catchment maps and information, we found that the original delineation and the corresponding meteorological forcing dataset omitted one headwater subcatchment. We corrected the delineation, re-collected the input data, and retrained all deep learning-related models and frameworks for Jiji. We substantially revised the Jiji results across all figures and tables in the revised manuscript. Correcting the delineation of Jiji and retraining all models improved hydrograph reconstruction accuracy across all Jiji models, but did not reverse the performance ranking in Table 1 or the main conclusions. The qualitative patterns of the extracted representations at Jiji are consistent with the original version.

6. Redraw, add, and rearrange the figures, tables, and sections.

a. Figure 1 has been redrawn to add the geological condition.

b. Figure 4 has been redrawn to add KGE and align the presented period.

c. Adding the free-run TDLR with estimated Qt−1 and KGE metric in Table 1 and rearranging Table 1 and related discussion in Sect. 4.1.

d. Fig 5(a) and (b) in Sect.4.2.1 are from the original Fig 5(f), and the new comparison between the external AET product and calculated AET is Fig 5(b).

e. To discuss canopy storage dynamics in Sect. 4.2.2, Fig 6(a), (b), and (c) are from the original Fig 5(e), Fig 1(d), and Fig B1

f. In Sect. 4.2.3, Fig 7(a), (b), and © are the original Fig 5 (c), Fig 5(d), and Fig 1 (c) to discuss the surface water routing.

g. In Sect. 4.2.4, Fig 8(a), (b), and (c) are the original Fig 5(a), Fig 5(b), and Fig 1(b) for runoff-generation and subsurface storage dynamics discussion.

h. In Sect. 4.2.5, Fig. 9 is the original Fig. 7. It shows the functional representation comparison between the ones extracted from TDLR and the dPL-HBV model.

i. In the revised manuscript, we added the 95% confidence interval of the curve fitting in Fig 5(a), Fig. 6(a), Fig 7(a) and (b), Fig. 8(a) and (b).

j. In Sect. 4.3, we added the sensitivity test for Lt in Fig. 10 to show that the residual signal is still significant even considering the precipitation measurement error, supporting the hypothesis generation for IGF.

k. Added Fig 12 in Sect. 4.5; it presents the workflow between AI and export collaboration with mechanistic hypotheses generation with TDLR. In the added Sect. 4.5, we discussed the concept of the extracted functional representation treated as data-driven mechanistic hypotheses generation.

Response to referee 2

This manuscript presents a temporal difference loss regularization framework, TDLR, that trains an LSTM within a conceptual bucket structure to infer catchment-scale hydrological response functions. The topic is interesting and relevant. The manuscript has some strengths: the framework is well motivated and the repeated-training experiment is a useful check on numerical stability. However, I think the current version sometimes interprets

model-derived internal variables more strongly than the evidence supports. The results show plausible and stable latent process representations under the chosen model structure, but they do not yet fully establish that these are identifiable or physically unique hydrological process functions.

I therefore recommend major revision.

Major comments

1. *My main concern is that the identifiability of the inferred process functions is not yet well established. Many internal quantities, including canopy storage change, subsurface storage change, evapotranspiration partitioning, runoff partitioning, and routing parameters, are inferred mainly from streamflow error and the imposed bucket constraints. It is therefore possible that different combinations of internal process responses could lead to similar hydrograph performance. The repeated training with 30 random seeds in Sect. 4.5 is helpful, but I would interpret it mainly as evidence of optimization stability under the same data, architecture, loss function, and model structure. It is not yet strong evidence of physical identifiability. The authors also note in Sect. 4.1 that accuracy alone cannot fully verify identifiability. The paper would be stronger if the authors softened wording around "identified process functions" and added some targeted checks, such as synthetic experiments with known process functions, validation against independent ET/soil moisture/groundwater/baseflow data, or sensitivity tests with alternative bucket structures.*

Response:

We thank the referee for this thoughtful comment, and we agree that the original wording overstated the identifiability of the extracted functions. We have revised the manuscript to explicitly distinguish optimization/re-extraction stability from physical identifiability. The 30-seed experiment is now interpreted only as evidence that the extracted representations are stable under repeated optimization with the same data, architecture, constraints, and bucket structure. We no longer treat this stability alone as evidence of physical uniqueness.

More importantly, the revised manuscript now explicitly states in the Introduction, Sect 4.2, 4.5, and Conclusion that the extracted functions should be regarded as candidate mechanistic representations rather than physically unique solutions. We also identify the conditions under which their credibility can be weakened, including measurement uncertainty, inadequate bucket structure, non-stationarity, insufficient information, and contradiction by independent observations. Regarding the suggested targeted checks, we followed the referee's second suggestion: we use independent AET products, spatial

rainfall information, wind–precipitation relationships, and catchment geomorphological/geological information as external evidence. These data were not used during LSTM training. Their agreement with the extracted representations, therefore, provides partial external corroboration, while we explicitly acknowledge that such agreement does not establish unique structural identifiability. We did not add a synthetic experiment in this revision, and we do not claim that the present three-catchment study establishes physical uniqueness. Instead, we have deliberately narrowed the scientific claim to stable and testable mechanistic hypotheses under prescribed conditions.

2. *The residual term L_t needs a more cautious treatment. In Sect. 3.1 and Sect. 4.3, L_t is defined as a mixture of interbasin groundwater flow, measurement error, and structural mismatch. The manuscript also acknowledges that IGF cannot be definitively separated from these other components. Later, however, the residual patterns are interpreted quite strongly as evidence for IGF or preferential subsurface pathways. This interpretation may be plausible, but it is not unique. A systematic residual could also reflect precipitation bias, discharge uncertainty, PET/AET uncertainty, missing processes, or model structural error. I would suggest referring to L_t more neutrally as a water-balance closure residual unless independent groundwater evidence is provided. Even a simple uncertainty analysis for precipitation, discharge, and ET would make this section more convincing.*

Response: We thank the referee for pointing out that the interpretation of the mass-conservation residual as IGF was too strong. We have revised the manuscript in three ways. First, we soften the interpretation. The IGF is now presented as one possible mechanistic hypothesis for the residual signal, not as an identified process. Second, we added an external AET product to verify the extracted AET calculation and reduce the degrees of freedom in the water budget assessment, so that the ET component of the residual is externally constrained. Third, we also added a sensitivity test in which the current-day precipitation is perturbed by ±20%, which is recognized as the common range of the precipitation measurement perturbation in sparsely gauged mountainous catchments (Hrachowitz and Weiler, 2011). The resulting uncertainty envelope is shown in Fig. 10 in the revised manuscript. Specifically, for the input precipitation series of LSTM, the ±20% perturbation was applied to each concurrent timestamp.

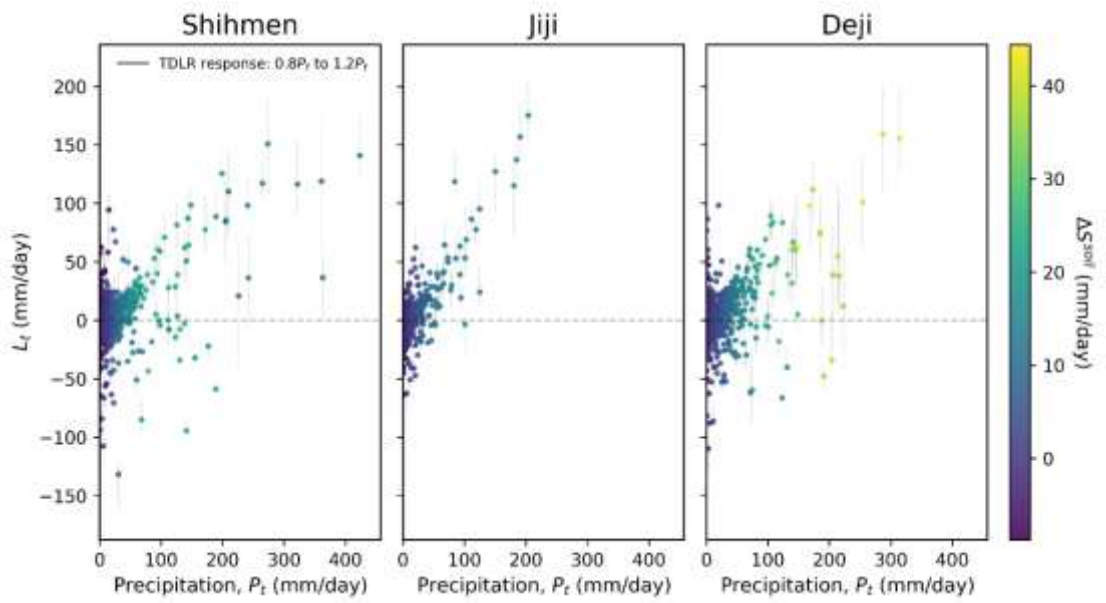

Figure 10 (in revised manuscript). Daily relationship between precipitation, Pt, and the mass-conservation residual, Lt, over the entire record for the three catchments. Point colors indicate the storage change of the integrated subsurface bucket, ΔSsoil. The gray bars show the sensitivity of Lt to a concurrent ±20% perturbation of the precipitation, recognized as the common range of the precipitation measurement (Hrachowitz and Weiler, 2011): each day's precipitation is varied between 0.8Pt and 1.2Pt, the functional responses are recalculated with the trained LSTM, and the resulting range of Lt defines the bar. Specifically, for the input precipitation series of LSTM, the ±20% perturbation was applied to each concurrent timestamp.

The analysis in Section 4.3 indicates that precipitation uncertainty affects the residual, particularly during high-precipitation events, while the broader catchment-specific residual patterns remain visible beyond this sensitivity range. Most importantly, we have revised the interpretation so that IGF is presented as one possible mechanistic explanation and a testable hypothesis, rather than as an identified process. The revised manuscript explicitly states that water-balance information alone cannot uniquely distinguish groundwater exchange from persistent measurement bias or model structural error. We also identify the observations needed for direct testing, including groundwater levels, hydraulic gradients, environmental tracers, hydrochemical information, and spatially distributed baseflow estimates. Please see Section 4.3 in the revised manuscript, from lines 547 to 559.

3. Sect. 4.2 and Fig. 5 are central to the paper, but the interpretation needs some tightening. The LSTM uses a 64-day sequence of multiple meteorological variables, while many process interpretations are based on one-dimensional projections against current precipitation. These projected curves are useful summaries, but they are not the full learned functions. This matters because the projected relationships are shaped by the observed covariation among precipitation, wind speed, radiation, humidity, temperature, seasonality, and antecedent wetness. For example, the canopy response in Shihmen is discussed in terms of wind-rainfall interaction, but other correlated factors may also contribute. The authors do not necessarily need to resolve all of this in one paper, but the limitation should be made

clearer. Conditional partial dependence, controlled perturbation tests, or two-variable response surfaces would help. If these analyses are not added, the physical interpretation of the projected curves should be phrased more cautiously.

Response: We agree and acknowledge that our process interpretations based on these responses to current precipitation do not fully represent the multi-dimensional functions learned by LSTM. The projected relationships are indeed shaped by the observed covariation among the meteorological drivers and antecedent conditions, and we have revised the manuscript to make this limitation explicit.

We had made three major changes. First, we additionally provide the relationships between the extracted representations and each meteorological driver, including precipitation, humidity, radiation, temperature, pressure, and wind speed, in Appendix A. Across these projections, the current precipitation was found to provide the dominant information for responses in these three catchments, motivating the use of precipitation as the primary projection axis. Second, we also explicitly state that the projected curves are one-dimensional summaries of the learned multi-dimensional representation, conditioned on the observed covariation of the drivers, and that the individual memory (antecedent-condition) effects of each extracted representation are not separately tested in the present study. Third, more specifically, the main reason the information is dominated by current precipitation is that our study area is not only humid and warm but also has intense precipitation. We stated this limitation in Lines 343 to 349. We also revised the physical interpretations accordingly, including the wind–rainfall interpretation of the canopy response at Shihmen, which is now phrased as a mechanistic hypothesis consistent with the wind–precipitation covariation, rather than an isolated causal effect.

We agree that conditional partial dependence or controlled two-variable perturbation experiments would provide a stronger causal interpretation. These analyses are beyond the scope of the current proof-of-concept study. We therefore revised the language to emphasize that the projected curves are one-dimensional summaries of the learned representations, from which mechanistic hypotheses are generated, and the further analyses suggested by the referee are identified as future work.

4. The evaluation setup is not yet transparent enough. Sect. 3.2 indicates that observed previous-day streamflow is used during training, while simulated previous-day streamflow can be used during reconstruction. Sect. 4.1 then presents reconstruction results using an initial streamflow value. This leaves some uncertainty about whether the reported results are teacher-forced one-step predictions, fully recursive continuous simulations, or something in between. Please provide exact training, validation, and testing periods, clarify warm-up and initialization, and report whether normalization and hyperparameter tuning are based only on training data. It would also be useful to report fully recursive simulation performance separately from any teacher-forced setting. The benchmark comparison in Sect. 4.4 also needs more detail. The performance gap between TDLR and dPL-HBV is large, especially for Jiji, so readers need enough information to judge whether the

comparison is fair. The claim would be stronger if the authors described the dPL-HBV setup, parameter ranges, forcing inputs, spin-up, optimization budget, and whether all models use the same data split and evaluation protocol.

Response: We thank the referee for pointing out the insufficient transparency of the evaluation presentation and the training budget revealing. We rewrote Sections 2. and 4.1 and added Appendix C with the full training and evaluation details.

We address each point in turn.

(1) Simulation mode. The hydrograph reconstruction without previously observed streamflow has been added in Section 4.1. We now explicitly distinguish the two modes. In the teacher-forced setting, TDLR and the pure LSTM with Q(t−1) use the *observed* previous-day streamflow at every step (one-step-ahead prediction). In the fully recursive (free-running) setting, newly added to Table 1 and Sect. 4.1, TDLR is initialized with the observed streamflow on the day before the whole recorded period/description of initialization, and thereafter uses only its own *simulated* previous-day streamflow throughout the whole recorded period. The two settings are reported separately, as the reviewer suggested; dPL-HBV and the pure LSTM without Q(t−1) are compared against the recursive setting, since neither uses observed antecedent streamflow.

(2) Framing of the comparison. Following this and Referee 1's related comment, we reframed the comparison between the fashion with and without previously observed streamflow. One is a teacher-forcing fashion comparison. Another, like dPL-HBV, serves as an advanced parameterization with LSTM compared to the pure LSTM without teacher-forcing. The new point of the comparison is to highlight that the fixed prior functional forms of hydrological fluxes might limit the deep learning model's use of information because the fixed equation form may not be adequate for the specific catchment behavior. Sect. 4.1 now reframes the comparison into with- and without-observed-Q(t−1) groups, so that the performance gap between TDLR and dPL-HBV is interpreted within a like-for-like information setting. Regarding Jiji specifically, we also note that the catchment delineation was corrected in this revision and all Jiji models retrained, which substantially improved dPL-HBV's accuracy there (see the revision summary). The revised Table. 1 is shown below.

Table 1. Comparison of simulation accuracy among the proposed TDLR framework, dPL-HBV, and the pure LSTM. We used Nash-Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970) and Kling-Gupta efficiency (KGE) (Gupta et al., 2009; Knoben et al., 2019) as metrics to assess simulation accuracy over the entire record. Bold values show the TDLR/model computes current streamflow with the previous observed streamflow, Qt-1, as input; All other models simulate the current streamflow from meteorological driving forces only. Numbers in parentheses (64/365) denote the LSTM input sequence length in days. Note that dPL-HBV, as a storage-based model, cannot ingest observed Q(t−1); comparisons are therefore most meaningful within each group.

	TDLR with observed Qt-1	*Pure LSTM added Qt-1 as input (64/365 days)*	*Free run TDLR with estimated Qt-1*	*dPL-HBV (static/dynamic)*	*Pure LSTM (64/365 days)*
Shihmen (NSE)	**0.863**	**0.86/0.85**	0.807	0.833 / 0.802	0.84/0.86
Jiji (NSE)	**0.867**	**0.86/0.87**	0.612	0.611 / 0.676	0.81/0.81
Deji (NSE)	**0.823**	**0.82/0.84**	0.711	0.709 / 0.656	0.74/0.72
Shihmen (KGE)	**0.930**	**0.82/0.86**	0.867	0.860 / 0.857	0.89/0.91
Jiji (KGE)	**0.931**	**0.88/0.90**	0.801	0.688 / 0.800	0.80/0.90
Deji (KGE)	**0.898**	**0.79/0.84**	0.826	0.701 / 0.728	0.78/0.76

(3) Data splits. Details of the training budget are provided in Appendix C in both tabular form and accompanying text. In addition, the training, validation, and testing periods are identical across the benchmark evaluations. For Shihmen, the development period is 8 March 2011–19 August 2020, and the test period is 20 August 2020–31 December 2022 (864 test days). For Jiji, the corresponding periods are 7 April 2016–26 August 2021 and 27 August 2021–31 December 2022 (492 test days). For Deji, they are 6 January 2012–19 October 2020 and 20 October 2020–31 December 2022 (803 test days). The data splits information has been added in Lines 119 to 130. Within each training and validation period, TDLR and the pure LSTM reserve approximately 10% of the samples for validation using high-flow-stratified random sampling, while the remaining samples are used for parameter optimization. Consequently, training and validation samples are interspersed within the same development period rather than occupying separate continuous calendar intervals. dPL-HBV uses the complete development period for training because it employs a fixed epoch budget without validation-based early stopping. All models are scored over the same calendar test dates, and longer LSTM histories use only pre-test observations and do not delay the test starting date. The details of training budgets are shown in Table 3.

Table 2. Data split.

Catchment	Train + Validation	Test period	Test samples
Shihmen	2011-03-08–2020-08-19	2020-08-20–2022-12-31	864

Jiji	2016-04-07–2021-08-26	2021-08-27–2022-12-31	492
Deji	2012-01-06–2020-10-19	2020-10-20–2022-12-31	803

Table 3. Training budgets information

Model	Seeds	Optimizer	Batch size	Epochs	Training sequence and warm-up
TDLR	30	RMSprop (0.0001)	64	Early-Stop 20	65-day input; no training warm-up
Pure LSTM	30	RMSprop (0.0001)	64	Early-Stop 20	64- or 365-day input; no training warm-up
Static dPL-HBV	1	Adadelta	1	50, fixed	365-day loss window + 365-day state warm-up
Dynamic dPL-HBV	1	Adadelta	1	50, fixed	365-day loss window + 365-day state warm-up

5. *The manuscript sometimes states that TDLR learns process functions "without relying on prior functional forms" and presents the method as a new paradigm for solving the closure problem. I agree that the method avoids some fixed empirical process equations, but it still imposes substantial prior structure: the bucket layout, process ordering, Muskingum routing, parameter bounds, non-negativity constraints, 64-day input window, and time-invariance assumption. A more precise framing would be that TDLR learns flexible process-closure relationships within a prescribed mass-balanced conceptual structure. This is still a meaningful contribution, but it would avoid overstating the method. Similarly, conclusions about the general non-transferability of fixed functional forms should be limited to the three tested catchments unless broader tests are added.*

Response: We agree with the referee. The TDLR framework is not a general-purpose, bucket-free data-driven approach, but it can conditionally extract functional representations. We have adopted the referee's suggested framing in the revised manuscript: to avoid overclaiming about TDLR, we emphasized that TDLR learns flexible process-closure relationships within a prescribed mass-balanced conceptual structure. It flexes the functional representation of hydrological fluxes within the given bucket and across LSTM output ranges, compared with the conventional hybrid model. The related statements are in Lines 631 to 633 in the revised manuscript.

As mentioned in the revised manuscript in Sect, 4,5, the conditions and given buckets, along with prior knowledge from the hydrologist and locally specific characteristics of the catchments, form the basis for generating a hypothesis that describes catchment-scale behaviors. We adjusted the description of TDLR to highlight that it is a conditional search that extracts the functional representations for hydrologgical fluxes from data. Therefore, TDLR is now a conditional search for functional representations of the hydrological fluxes, and we state explicitly that the prescribed conditions, the bucket design, output ranges, and the hydrologist's prior and catchment-specific knowledge form the foundation on which the extracted hypotheses are generated.

Regarding the second point, we agree that conclusions about fixed functional forms must be limited to the tested catchments. We have revised the relevant statements in Sec. 4.1 and the Conclusion so that the discussion of the non-transferability of fixed functional forms for moisture fluxes is scoped to our three study catchments.

Minor comments

1. Please define and use “functional representation,” “projected relationship,” “response function,” and “constitutive relationship” more consistently.

Response: Thanks to the referee for pointing out the terminology inconsistency. We revised and unified the terminology with functional representation throughout the whole manuscript. But only maintain the projected relationship in Figure 3 and the opening of Section 4.2 to highlight that the discussion is focused on the current precipitation and target functional representations. We now use functional representation as the primary term for the LSTM-extracted process mapping. Projected relationship refers specifically to its visualization against a selected driving variable, such as current precipitation. Mechanistic hypothesis is used when a hydrological interpretation is proposed but has not been independently verified.

We have also explicitly distinguished these projected relationships from the complete sequence-based functional representation.

2. NSE alone gives a limited view of model behavior. KGE, bias, RMSE, high-flow error, low-flow error, peak timing error, and water-balance error would provide a fuller evaluation.

Response: We agree and added Kling–Gupta efficiency (KGE) as a complementary performance metric. The revised Table 1 and Figure 4 report both NSE and KGE for all model configurations. We retained NSE because it is directly connected to the original evaluation framework, while KGE provides complementary information on correlation, variability, and bias.

3. Fig. 5 would benefit from uncertainty bands and sample-density information, especially in the high-precipitation tails where the fitted curves may be less well constrained.

Response: We agree with the referee that the fitted curve has a wide confidence interval due to the limited high-precipitation sample. In the revised manuscript, we added the 95% confidence interval for non-linear regression from extracted functional representation (points) to fitted curves (visualized trend). In Appendix B, including the distinction between dense low-to-moderate precipitation regions and sparse high-precipitation tails. The fitting procedure uses binned means in dense regions and raw observations in sparse tails to prevent the dense low-variance region from dominating the fitted relationship. Based on this technique, we used Monte Carlo sampling to estimate the 95% confidence interval of fitted curves.

4. Physical interpretations such as preferential fracture flow, backwater effects, and canopy shaking should be framed as hypotheses unless supported by independent observations.

Response: We agreed with the suggestion from the referee. We reframe all the extracted function representations by TDLR as the plausible mechanistic hypotheses.

5. The manuscript would benefit from language editing and some tightening, especially where model output, fitted projection, physical interpretation, and independent evidence are discussed together.

Response: We agreed with the suggestion from the referee. We tightened Section 4.2 with Section 2.2. Make the flow from the discussion figure and result first, then independent information or evidence, and the interpretation of the mechanistic hypotheses.

Reference

Hrachowitz, M., & Weiler, M. 2011. Uncertainty of precipitation estimates caused by sparse gauging networks in a small, mountainous watershed. *Journal of Hydrologic Engineering*, 16, 460–471. DOI: 10.1061/(ASCE)HE.1943-5584.0000331.

Nash, J., & Sutcliffe, J. 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

Gupta, H.V., Kling, H., Yilmaz, K.K. and Martinez, G.F. 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of hydrology 377(1-2), 80–91.

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