Response to the reviews

We thank the Editor and the reviewers for the general appreciation of our work and their efforts to provide us with a constructive review which will help improve the manuscript.

Below is a point-by-point response to the reviewers' comments. We marked our replies in a blue font, while original reviewers' comments are presented in a black font.

We hope that the Editor and the reviewers are satisfied with our discussion on the individual points and that our revised manuscript can be considered for publication in HESS.

On the behalf of all co-authors,

Yours sincerely,

Andrea Ficchi and Davide Bavera, on behalf of all co-authors

Text of the reviewers' comments and response

Anonymous Referee #1

Summary

This paper introduces a new objective function that appends to the existing KGE' a new component based on information theory. This new JDKGE metric is then tested in a variety of ways, covering three modeling experiments. For the first, the authors calibrate the GR6J model for 240 catchments in France (spatially lumped, daily time step) with the new JDKGE, as well as KGE, KGE', NSE, NSE_log and KGE_NP. They then compare the resulting distributions of performance scores on a number of separate metrics (KGE', the three KGE components r / alpha / beta, the Jensen-Shannon Divergence, and three flow percentile biases). These comparisons are both visual and based on statistical comparison of the resulting distribution. The second experiment involves calibrating GR6J and OS LISFLOOD for 45 basins worldwide, using the KGE' and JDKGE as objective functions. Distributions are compared as CDFs of a smaller number of metrics (KGE', JSD, lower and upper percentile). The third experiment involves multi-objective calibration of GR6J for the 240 French basins, using KGE' and JSD as complementary objectives. The resulting pareto fronts are compared to the JDKGE single optimum, with results (flow duration curves, averaged seasonal flows) shown for three basins. The paper concludes that in most cases the JDKGE criterion improves the simulation of certain flow percentiles, without much associated change in the KGE' (i.e., improvements come at limited/no cost elsewhere). There is a caveat that this is not the case everywhere, which the authors put down to poor suitability of the tested structures for these specific cases.

We thank the reviewer for their thorough and careful evaluation of our manuscript, and for the detailed and constructive comments provided both here and in the annotated PDF. We have carefully considered all the points raised and we are significantly extending our analyses and discussion in the revised manuscript to address all the points. We are confident that the resulting revisions will substantially strengthen the manuscript.

Opinion

I am a bit in two minds about this paper. On the one hand, the authors set out to improve low flow simulations, settled on a method, and show that this in principle has the intended effect. In that light, this paper can be a useful reference to spur further work. On the other hand, the paper to some extent seems rather ad-hoc to me. There are a number of things that stand out to me as potentially benefitting from further attention:

[1] The justification for the new metric seems very minimal to me. There are two reference to support the idea that using/adding metrics derived from information theory might be helpful. There is some explanation of the benefits of using something derived from the JSD, but no discussion of alternatives that could have been considered but were ultimately rejected. There is also no explanation for why adding the new component to the KGE in the way that was done (as an equally-weighted fourth term under the square root) is the best option. This makes it difficult to understand if the current implementation was a deliberate choice (meaning that alternatives were assessed but judged unhelpful) or more a lucky first attempt (meaning that follow-up work is possible that investigates alternative information-theory-based components, different construction of objective function, etc).

We thank the reviewer for this important comment regarding the need of enhancing the justification of the proposed metric and the need to better position it with respect to (any other potential or tested) alternative formulations.

We agree that the first version of the manuscript did not sufficiently demonstrate that the proposed formulation of JDKGE results from: (i) a theoretical analysis, and (ii) a structured and deliberate empirical exploration of several alternatives, rather than an ad-hoc or fortuitous choice. In fact, the proposed metric is the outcome of a principled design and an extensive comparative analysis of a large set of alternative calibration functions. In the first submitted version of our manuscript, we had not fully documented all the empirical tests carried out in order to limit the manuscript length, to focus mainly on presenting the newly developed metric and its benefits, while reporting only well-established alternative metrics and alternatives bringing competitive results. Now we will extend this part in the revised manuscript to highlight the theoretical considerations and empirical tests leading to the JDKGE formulation, as summarized below.

First, from a theoretical point of view, information-theoretic divergence measures were chosen as they provide a theoretically-grounded framework for comparing observed and simulated streamflow distributions in a hydrologically-meaningful way. Unlike moment-based metrics, divergences quantify differences across the full probability distribution, including tail behaviour.

They do this in a data-driven manner and without relying on parametric assumptions such as Gaussianity. Among divergence and distance-based alternatives, the Jensen-Shannon Divergence (JSD) was ultimately preferred on both theoretical and empirical grounds (see also second point below). Theoretically, JSD offers key advantages over the Kullback-Leibler Divergence (KLD) and related measures such as the Jeffreys Divergence (JefD), due to their sensitivity to zero-probability regions (critical when working with empirical observed and simulated streamflow distributions) and lack of the boundedness and symmetry (the latter applies to KLD and not JefD). Despite the theoretical advantages of JSD, KLD and JefD were carried forward to empirical testing anyway. Moreover, optimal transport metrics such as the Wasserstein Distance (WD) were also considered, but JSD was preferred over WD because of its boundedness, its computational tractability within iterative calibration frameworks, and its focus on differences in probability mass allocation rather than magnitude displacement, making the JSD more sensitive to distributional shape differences across flow regimes, and to hydrologically-meaningful differences in streamflow distributions (also thanks to our JSD discretization).

Second, from an empirical point of view, during our work we evaluated more than 30 alternative formulations, including:
(a) modified KGE-type metrics incorporating a different additional fourth component (not limited to JSD),
(b) combinations of 2 to 4 (KGE-type) components, and
(c) different formulations (e.g., components combined inside or outside the square-root, i.e., the Euclidean distance structure).

The candidate metrics with four components (point (a) above) included different information-theoretic divergence measures, namely the KLD, JSD and JefD, as well as distance-based metrics (which also replies to a comment made by the reviewer in the detailed annotations on the manuscript pdf). Among these options, the JSD consistently provided the most robust and balanced improvements, particularly for low-flow regimes, while maintaining, or slightly improving, performance under high-flow conditions. Moreover, different definitions of the original and modified KGE components and their combinations (point (b)) were tested, including the ratio of coefficients of variation which was preferred over the ratio of standard deviations. Similarly, we tested multiple formulations of the candidate new objective function (point (c)), including combinations of KGE and divergence measures, also aggregating the KGE and divergence as separate metrics (with the JSD outside the KGE's square root) into a single objective function. While some alternatives showed improvements in specific aspects, we found that the current formulation, based on an equally weighted four-component structure within the KGE framework, provided the best overall trade-off across flow regimes and performance criteria.

Thus, given all these considerations and following the reviewer's suggestion, we will strengthen the revised manuscript by:

- expanding the discussion in the Introduction and Methods (Section 2.1) to better justify the choice of divergence-based metrics from a theoretical perspective, also supported by additional literature and discussion of key references (e.g., Weijs et al., 2010; Vrugt, 2024; Pizarro et al., 2025);
- extending the discussion of the rationale for selecting a JSD-based metric over alternative candidates;

- adding a summary table of the main alternative formulations tested;
- including additional results for a representative set of alternatives (in part in the main Results section and in part in an Appendix, to maintain a clear and concise structure of the main manuscript), documenting the comparative performance of all the tested alternatives.

Even if many alternatives were theoretically and empirically evaluated, fully exploring all possible formulations is beyond the scope of a single paper. Our objective here was to introduce an extensively-validated and very promising new metric to be used as calibration function for hydrological models. In the revised manuscript, we will show more clearly that the chosen new metric (JDKGE) results from a systematic exploration, guided by theoretical literature-based considerations. However, we will also explicitly acknowledge in the discussion that further research could still investigate alternative information-theoretic measures, different weighting schemes, or objective function structures. We hope that our work will stimulate further research and discussion on the applications of information-theoretic metrics in hydrology.

References

Weijs, S. V., Schoups, G., and van de Giesen, N. (2010). "Why hydrological predictions should be evaluated using information theory", Hydrol. Earth Syst. Sci., 14, 2545–2558. https://doi.org/10.5194/hess-14-2545-2010

Vrugt (2024), "Distribution-Based Model Evaluation and Diagnostics: Elicitability, Propriety, and Scoring Rules for Hydrograph Functionals": https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2023WR036710) …

Pizarro et al. (2025) "Combining uncertainty quantification and entropy-inspired concepts into a single objective function for rainfall-runoff model calibration" https://hess.copernicus.org/articles/29/4913/2025/

[2] The comparison does not seem as clear to me as it can be. Line 204-205 explains that the JSD component in the JDKGE uses log-transformed flows. This means the JDKGE is a function of both regular and transformed flows (i.e., f(Q,log(Q)). The objective functions used for comparison however only have access to a single set of flows. KGE, KGE', NSE and KGE_NP are all f(Q), whereas NSE_log is f(log(Q)). It is thus unclear from the presented comparisons if the improvements provided by JDKGE are a result of the inclusion of the JSD component, or the fact that JDKGE is a merged metric of regular and transformed flows, or a combination of both. A comparison with JDKGE*, where the JSD component only has access to regular flows seems a very useful thing to add, because this would indicate if the improvements come from the merger of information theory or from extra weighting of low flows resulting from the log(Q) transformation.

We appreciate the reviewer's concern regarding the clarity of the comparison between JDKGE and other objective functions. Indeed, as correctly pointed out by the reviewer, the JDKGE metric is a function of both regular and log-transformed flows, while the other objective functions reported so far in the manuscript have only access to either original or transformed flows and not a combination of both. We agree that a comparison with JDKGE*, where the JSD-based component only has access to regular flows seems very useful to add. This was one of the tests we made at the start, when including the JSD component in the KGE, but given that the results on low-flows were further improved by the inclusion of the logarithm, at no significant expense for high flows, the latter option was chosen. From our first tests and analyses, we have seen that the improvements provided by JDKGE are a result of a combination of both the inclusion of the JSD-based component and the access to both regular and transformed flows. We will further evaluate this and include a systematic performance comparison of JDKGE and JDKGE*. We would like to clarify already that the introduction of the JSD component is not solely helping due to the use of log-transformed flows, but also by including the JSD on regular flows, as it adds a measure of the distributional distance between the predicted and observed flows, rather than only the first moments and correlation (as the KGE does). We are now further analysing the comparison of the JDKGE and JDKGE*, including different weights for the components, and we will add a summary of these results in the revised manuscript to show the advantages of JDKGE of the combination of changes (inclusion of the JSD and use of both regular and transformed flows) with respect to the original KGE. In particular, in the revised manuscript, we will include the JDKGE* among the selected metrics for which results will be reported systematically in the main body of the manuscript (not only in the summary of all tested options in the Appendix; see response to point [1]). The new Results figures showing the performance distributions for different selected calibration functions (current Figures 3 and 4) will include the JDKGE*, as suggested by the reviewer.

[3] While reading the manuscript, it was unclear to me if the results that were shown are for calibration data (i.e., the outcomes of data fitting) or for evaluation data (i.e., an estimate of how the model performs for unseen data). This clarification is needed and if, as I expect based on line 442-444, all results shown are for calibration only I would strongly recommend to add an assessment of JDKGE using unseen data. Particularly in the light of the rather short calibration windows, evidence that JDKGE helps the optimizer find better parameter sets that describe the general behaviour for the catchment, rather than parameter sets that better fit the specific conditions of the calibration data, is needed.

We acknowledge the importance of clarifying this point. Indeed, as correctly noted by the reviewer (Lines 442–444), all results presented in the original manuscript are based on calibration data only, i.e., the outcomes of data fitting. We made this conscious decision for two main reasons.

First, we used the full dataset for calibration to maximise the length of the calibration periods, avoiding splitting the data into separate calibration and evaluation sets. This was particularly relevant for the French catchment set, where shorter data records are available due to the high temporal and multi-scale resolution of that dataset, and which enabled hourly model calibration tests to validate the new metrics across sub-daily temporal resolutions (we will include a summary of this in an Appendix).

Second, this choice is consistent with recent literature. Arsenault et al. (2018) and Shen et al. (2022), suggest that calibrating over the full available time-series period and skipping validation entirely is the most robust split-sample decision, consistently outperforming strategies that calibrate on a portion of the available data and validate on independent (unseen) data (split-sample tests). This is also in line with the operational needs of systems such as EFAS/GLOFAS. We acknowledge, however, that this conclusion may be nuanced and context-dependent, as the literature on this topic is varied and the value of split-sample testing has been recognised for decades in the hydrological literature (e.g., Klemeš, 1986).

We therefore agree that the reviewer's concern is legitimate, and we will include additional results from a split-sample evaluation in the revised manuscript, demonstrating JDKGE's ability to generalise well to unseen data. These results will be presented in a dedicated figure and discussed in the context of the existing literature on calibration and evaluation strategies.

We performed preliminary tests and can already confirm that the split-sample evaluation does not change the overall message of the manuscript, further validating the robustness and generalisability of JDKGE.

References

Arsenault, R., Brissette, F., and Martel, J.-L. (2018). The hazards of split-sample validation in hydrological model calibration, Journal of Hydrology, 566, 346–362, https://doi.org/10.1016/j.jhydrol.2018.09.027, 2018

Klemes, V. (1986). Operational testing of hydrological simulation models. Hydrological Sciences Journal, 31(1), 13–24. https://doi.org/10.1080/02626668609491024

Shen, H., Tolson, B. A., and Mai, J. (2022). Time to Update the Split-Sample Approach in Hydrological Model Calibration, Water Resources Research, 58, e2021WR031 523, https://doi.org/10.1029/2021WR031523, publisher: John Wiley & Sons, Ltd.

[4] The calibration length seems rather short to me. Eight years for the French basins, 4+ years for the global basins. Particularly in intermittent regimes, this means the results really are a snapshot of model performance rather than a more generally applicable assessment of performance. An analysis of the robustness of conclusions w.r.t. calibration data length would be very welcome.

We appreciate the reviewer's concern regarding calibration length and welcome the opportunity to clarify this point. We agree that longer calibration periods are generally desirable for robust conclusions, consistent with the literature cited in the manuscript.

We would like to clarify that the “at least 4 years” mentioned for the global basins refers only to the minimum data threshold required for GloFAS calibration stations. In practice, almost all observed streamflow time series used in this study are considerably longer, with an average length exceeding 10 years.

While some calibration periods, particularly in the French catchment dataset, are relatively short (8 years), accepting this was a deliberate design choice motivated by two considerations.

First, this enabled us to evaluate the performance of various calibration functions when data availability is limited, a common constraint in real-world operational settings, particularly in large-scale global systems like GloFAS. In the GloFAS operational setting, where the LISFLOOD hydrological model is employed, in some areas the model needs to be calibrated using all the available good-quality time series, even if just over 4 years, in order to maximize spatial coverage of the calibrated domain including some data-scarce regions which otherwise would be completely uncalibrated (thereby reducing the need for regionalization techniques and helping improve the accuracy and reliability of the model). By assessing the performance of different calibration functions under these data-limited conditions, we can identify approaches that allow for effective and robust use of local data even in such data-scarce contexts.

Second, the French dataset was specifically developed as a multi-resolution benchmark (Ficchì et al., 2016), enabling hourly and sub-daily calibration tests, which have allowed us extending the validation of JDKGE across temporal resolutions beyond the daily scale (results will be summarised in an additional Appendix).

We agree that an analysis of robustness with respect to calibration data length would add value. To respond to this reviewer’s comment, in the revised manuscript we will extend the experiments and analyses reported in Section 3.2. We will do this by also adding at least 20 additional catchments to the global sample (currently 45 catchments) with longer observed time series available than the current average (at least 15 years), which will be included in the modelling experiments using JDKGE vs KGE’ for both LISFLOOD and GR6J. This will increase the average time series length and allow an analysis of the robustness of conclusions with respect to calibration data length, as suggested by the reviewer. We will report and discuss this in a concise new paragraph of the Results, providing sufficient insights into the sensitivity of conclusions to calibration data length without extending the manuscript further.

References

Ficchì, A., Perrin, C., and Andréassian, V. (2016). Impact of temporal resolution of inputs on hydrological model performance: An analysis based on 2400 flood events, Journal of Hydrology, 538, 454–470, https://doi.org/https://doi.org/10.1016/j.jhydrol.2016.04.016.

[5] In addition, there are numerous things that I think would strengthen the paper substantially:

We thank the reviewer for all the constructive comments. We responded to all the constructive points raised below in detail, mentioning our planned revisions and additions to take into account the useful suggestions received, that will help strengthen the article.

[6] All analysis is performed on distributions of scores only, without any mention of (a lack of) regional patterns. I expect it would be instructive to see if changes obtained from using JDKGE are uniformly spread in space or if there are specific regions where the metric enhances simulations more than others. Either case would be instructive.

We thank the reviewer for this thoughtful suggestion.

In general, we agree that analysing the spatial distribution of performance changes would add an instructive dimension to the evaluation of JDKGE, potentially revealing whether improvements are uniformly distributed or concentrated in specific regions or climatic contexts. However, a comprehensive spatial analysis of performance patterns is not feasible within our global catchment set and goes beyond the scope of this manuscript, whose primary goal is to introduce and validate JDKGE as an improved calibration metric across diverse, large catchment sets.

Given the sample size of the global catchment set, these maps would be informative at the catchment level, but will not support generalisations on broader regional or global patterns, also given the multivariate nature of the problem of clustering performance changes. While we acknowledge that exploring regional patterns of performance changes would be a valuable direction for future research, we think that it would deserve an in-depth evaluation on a larger catchment sample with a level of multi-variate analysis that goes beyond the scope of our manuscript. Accordingly, other authors that proposed and tested new metrics for hydrological model calibration, such as Gupta et al. (2009), Kling et al. (2012) and Pool et al. (2018) tested their newly proposed metrics on a relatively small number of catchments (e.g., 49 Austrian basins or a single basin or 100 US catchments) without investigating regional patterns, likely for similar good reasons as the ones we mentioned.

References

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

Kling, H., Fuchs, M., & Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, *424–425*, 264–277. https://doi.org/10.1016/j.jhydrol.2012.01.011

Pool, S., Vis, M., & Seibert, J. (2018). Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency. Hydrological Sciences Journal, 63(13–14), 1941–1953. https://doi.org/10.1080/02626667.2018.1552002

[7] Related to the previous point, scores such as NSE and KGE are known to be (sometimes highly) conditional on the data used to calculate them. This sampling uncertainty (Lamontagne et al., 2020; Clark et al, 2021; Vrugt & De Oliveira, 2022) can be estimated and is often rather large. An assessment of sampling uncertainty would show how the changes in evaluation metric values obtained by switching to JDKGE compare to the uncertainty in the scores themselves. If the benefits of using JDKGE are consistently larger than the uncertainty in the scores, this would be a helpful line of evidence. Vice versa, if the benefits of using JDKGE are small compared to the uncertainty in the scores themselves this would be worth knowing too.

We thank the reviewer for raising this important point and for the relevant references. We fully agree that sampling uncertainty in performance metrics such as NSE and KGE is a well-documented and non-trivial issue. We also agree that assessing whether the performance improvements obtained using JDKGE are consistently larger than the inherent uncertainty in the scores themselves is a very valuable line of evidence that would help strengthen the key messages of the manuscript and the comparative evaluation of JDKGE and KGE calibrations.

We will therefore carry out an assessment of sampling uncertainty in the revised manuscript, following one of the approaches outlined in the references suggested by the reviewer (e.g., Clark et al., 2021). We will compare the sampling uncertainty of the KGE and JDKGE metrics with respect to the metrics obtained by calibration for the global catchment set given the longer time series of this set with respect to the French catchment set. Linked to this, we note here that we are also expanding this catchment set to include more catchments with longer time series (>15 years) than the current average one.

The results will be presented either as an additional figure in the main manuscript or in an Appendix, depending on the complexity of the analysis and the overall structure (and increasing length) of the revised manuscript. This analysis will allow us to contextualise the observed performance differences between JDKGE and KGE-based calibrations relative to the uncertainty in the scores themselves, providing a more rigorous and complete evaluation of the practical value of the new metric.

[8] As far as I can tell, experiments 1 and 2 calibrate a single parameter set per model per metric per basin. One of the main conclusions in the work is that considerable improvements in low flows can obtained with minimal reduction in KGE' scores. This strongly suggests that the

main benefit of the JDKGE is that it nudges the optimization away from the very highest KGE' optimum, but that in absolute terms these differences need not be large. I think this can be tied in more with the existing literature on parameter uncertainty and equifinality.

We appreciate the reviewer's insightful comments and we confirm that their understanding of experiments 1 and 2 is correct (in experiments 1 and 2, we calibrate a single parameter set per model, per metric, per basin). Indeed, one of the main conclusions of our study is that the improvements in low-flow simulations obtained with JDKGE come with only minimal reductions in KGE' scores. This is the result of a deliberate design choice and extensive empirical testing that led us to choose this metric formulation (see also response to point [1]). From the outset, our target was to develop a new metric that would not drastically change the overall performance as assessed by the existing popular KGE metric.

The reason for this target is that we are already applying the JDKGE metric to an operational product, namely the EFAS/GLOFAS system, which provides critical flood forecasting and simulation services. It is essential that this system maintains users' trust in the model performance, especially in terms of simulating and forecasting floods, as it has a direct impact on decision-making and emergency response. Therefore, we had to find a balanced formulation that would improve low-flow simulations without compromising the existing performance of the system in high-flow events.

In other words, our goal was to refine the KGE in a way that would address a specific limitation (limited performance on low flows) while preserving the overall performance of the system. We believe that the JDKGE metric achieves this goal successfully, and we are pleased to see that our positive results confirm our expectations and they are well understood and received by the reviewers. By staying close to the KGE metric, we can ensure a smooth transition to the new metric and maintain the trust and reliability of the EFAS/GLOFAS system (and potentially of other hydrological operational models), while also providing improved performance on low-flow conditions.

In the revised manuscript, we further clarified this point in the introduction. Moreover, as suggested by the reviewer, we will enhance the discussion of current results linking them with the existing literature on parameter uncertainty and equifinality.

[9] Related, adding some evidence of convergence of the calibration algorithm would be good to add. The text (lines 625-626) already suggests that there are certain cases in the multi-objective experiment where calibration was not fully successful, and getting some indication of how often this happened in the other two experiments would be helpful.

We thank the reviewer for this suggestion. We agree that providing evidence of calibration convergence would strengthen the manuscript and increase confidence in the reliability of the results. In the revised manuscript, we will add a brief comment on the analysis of convergence across all experiments, including an indication of convergence diagnostic results, like the frequency of cases where calibration may not have fully converged. This could be presented as a summary table or figure in Appendix, and/or adding some text in the discussion of the results, reporting statistics of convergence diagnostics such as the evolution of the objective function across iterations or the frequency of cases where the optimisation did not reach a stable solution. This will complement the existing note at Lines 625–626 regarding the multi-objective experiment and provide a more complete picture of the calibration performance across all three experiments.

I hope these points highlight why I struggle to decide what to recommend for this paper. On the one hand, it outlines a potential trajectory to improve model calibration. On the other hand, I have a lot of open questions about the process presented in the paper here, some of which seem fairly fundamental to me. I hope the above and the comments in the pdf are helpful in some way.

Please note that Section 2.1.3. was beyond my ability to assess.

We appreciate the reviewer's candid reflection on the manuscript and their detailed and positive engagement with our work. Their comments, both here and in the annotated manuscript, are genuinely helpful and will contribute to making the revised manuscript a more rigorous and complete contribution to the hydrological community.