udp UNIVERSIDAD DIEGO PORTALES

Response to the Editor – Manuscript: *"Combining uncertainty quantification and entropy-inspired concepts into a single objective function for rainfall-runoff model calibration"* by Pizarro, Montanari & Koutsoyiannis

Below, we list editor's comments verbatim in **bold**, followed by responses to these comments in blue. We *italicise* the revised additions to the manuscript.

Editor: Nunzio Romano

Obs. 1: Dear authors, Your article has undergone a thorough review by two reviewers, who generally rated it quite well, although with some slight differences of opinion. In particular, Reviewer #1 made several comments directly in the text of the article, suggesting that certain concepts should be clarified and made more explicit, especially to the benefit of a wider readership. Reviewer #2, on the other hand, provided some general comments along with primary criticism concerning the presentation and discussion of the results. Indeed, the discussion of the interesting results is precisely the weakest part of this article, I believe, and certainly needs much improvement. The article is returned for substantial revisions, to be uploaded together with a detailed point-by-point response to all the comments received from the reviewers and the editor.

Ans. 1: We wish to thank the reviewers and editor for their helpful comments. We are very appreciative and believe that these comments and suggestions have strengthened the quality of this paper. A detailed point-by-point response to all the comments received from the reviewers and the editor is presented below.

Response to reviewer 01 – Manuscript: *“Combining uncertainty quantification and entropy-inspired concepts into a single objective function for rainfall-runoff model calibration”* by Pizarro, Montanari & Koutsoyiannis

Below, we list reviewer’s comments verbatim in **bold**, followed by responses to these comments in blue. We *italicise* the revised additions to the manuscript.

Thank you again for your helpful feedback!

Reviewer 01: Keith Beven

Obs. 1: This is a really nice paper but needs a few issues of clarification in both the introduction and in the presentation of the methodology before publication in line with the comments in the MSS.
Ans. 1: Thank you for your comments and suggestions. Introduction and Methods sections were revised and improved accordingly.

Obs. 2: Line 24, “…treatment of error sources (see, e.g., Blazkova and Beven 2002; 2004; Krzysztofowicz 2002)…”, highlighting “Krzysztofowicz 2002”: Beven 2009 Environmental Modelling - An Uncertain Future provides a more general review
Ans. 2: Cited works in Line 24 were updated, incorporating Beven (2018).

References:
Beven, K. (2018). *“Environmental modelling: an uncertain future?”*. CRC press.

Obs. 3: Line 31, “…understandable to end users (Beven, 2024)…”, highlighting “Beven, 2024”: perhaps mention recent Vrugt attempts to provide more flexible likelihood functions (albeit with more parameters)
Ans. 3: Thank you for the suggestion. We incorporated Vrugt et al. (2022) alongside Beven (2024).

References:
Beven, K. (2024). *“A brief history of information and disinformation in hydrological data and the impact on the evaluation of hydrological models”*. Hydrological Sciences Journal, 69(5), 519-527.

Vrugt, J. A., de Oliveira, D. Y., Schoups, G., & Diks, C. G. (2022). *“On the use of distribution-adaptive likelihood functions: Generalized and universal likelihood functions, scoring rules and multi-criteria ranking”*. Journal of Hydrology, 615, 128542.

Obs. 4: Line 51, "…possibly statistically incoherent and potentially unreliable parameter and predictive distributions…": This is misleading - you know very well that incoherence argument could only be made in respect of ideal cases when the likelihood is known - and that GLUE could also be used with the correct likleihood to give the same result if that information is known (Beven et al. 2008). But real cases are non-ideal in that respect - as the nonsense that can come from applying formal likelihoods shows (Beven and Smith 2015).

Ans. 4: We shared the reviewer's vision in terms of real cases are non-ideal and, therefore, uncertainty quantification is a must being the core of RUMI formulation. The sentence in question was modified and reads now:

"However, GLUE has faced criticism in terms of the subjective decisions required in its application and how these affect prediction limits (informal likelihood function, lacks of maximum likelihood parameter estimation, and omission of explicit model error consideration). This subjectivity might lead to not being formally Bayesian (for that reason, GLUE includes the term "generalized" in its name)"

Obs. 5: Line 56, "…with formal Bayesian approaches…": But GLUE can also be used with MCMC/DREAM and has been - please recognise that defining a likelihood and searching the space are two seperate issues.

In that respect it might also be inclusive to mention the use of limits of acceptability within GLUE to get around these problems (going back at least to Freer et al 2004, also Vrugt and Beven, 2018 using DREAM_LOA; Beven and Lane, 2022 and in CURE). Still some subjectivity but involving more thoughtful decisions about the data. There is also a UPH20 paper (Beven et al., https://doi.org/10.5194/piahs-385-129-2024)

Ans. 5: We recognised that defining likelihood functions and searching the solution space in calibration are two different and separate issues. We wrote it explicitly in the main text which reads now:

"Building on previous work (see, e.g., Blasone et al. 2008), researchers have compared GLUE with formal Bayesian approaches. At this regard, both formal Bayesian approaches as well as GLUE can be used with advanced Monte Carlo Markov Chain (MCMC) schemes such as the Differential Evolution Adaptive Metropolis (DREAM, Vrugt et al. 2008). Important to note is that defining likelihood functions and searching the solution space at calibration are two independent issues. One way to get around these problems relies on the limits of acceptability which are typically used (but no mandatory) with GLUE (see, e.g., Freer et al, 2004; Vrugt and Beven, 2018; Beven and Lane, 2022; Page et al., 2023; Beven et al., 2024), involving more thoughtful decisions about the data (even though still with subjectivity)."

References:

Beven, K., & Lane, S. (2022). On (in) validating environmental models. 1. Principles for formulating a Turing-like Test for determining when a model is fit-for purpose. Hydrological Processes, 36(10), e14704.

Beven, K., Page, T., Smith, P., Kretzschmar, A., Hankin, B., & Chappell, N. (2024). UPH Problem 20–reducing uncertainty in model prediction: a model invalidation approach based on a Turing-like test. Proceedings of IAHS, 385, 129-134.

Freer, J. E., McMillan, H., McDonnell, J. J., & Beven, K. J. (2004). Constraining dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures. Journal of Hydrology, 291(3-4), 254-277.

Page, T., Smith, P., Beven, K., Pianosi, F., Sarrazin, F., Almeida, S., ... & Wagener, T. (2023). The CREDIBLE Uncertainty Estimation (CURE) toolbox: facilitating the communication of epistemic uncertainty. Hydrology and Earth System Sciences, 27(13), 2523-2534.

Vrugt, J. A., & Beven, K. J. (2018). Embracing equifinality with efficiency: Limits of Acceptability sampling using the DREAM (LOA) algorithm. Journal of Hydrology, 559, 954-971.

Obs. 6: Line 139-140, "For each point, a sample is established comprising neighbouring simulated river flows, defined by m_1 Flows smaller and m_2 flows larger than the point's discharge, both with the smallest differences.": Not very clear - what is meant by neighbouring? Illustrate with a simple figure... It is also then not clear how the m1 and m2 samples are related to u in RUMI

Ans. 6: Thank you for rising this issue. As in the original BLUECAT paper, neighbouring was taken in terms of the magnitude of stream flows. Figure R01.1 was generated and incorporated into the manuscript. The figure in question reads now:

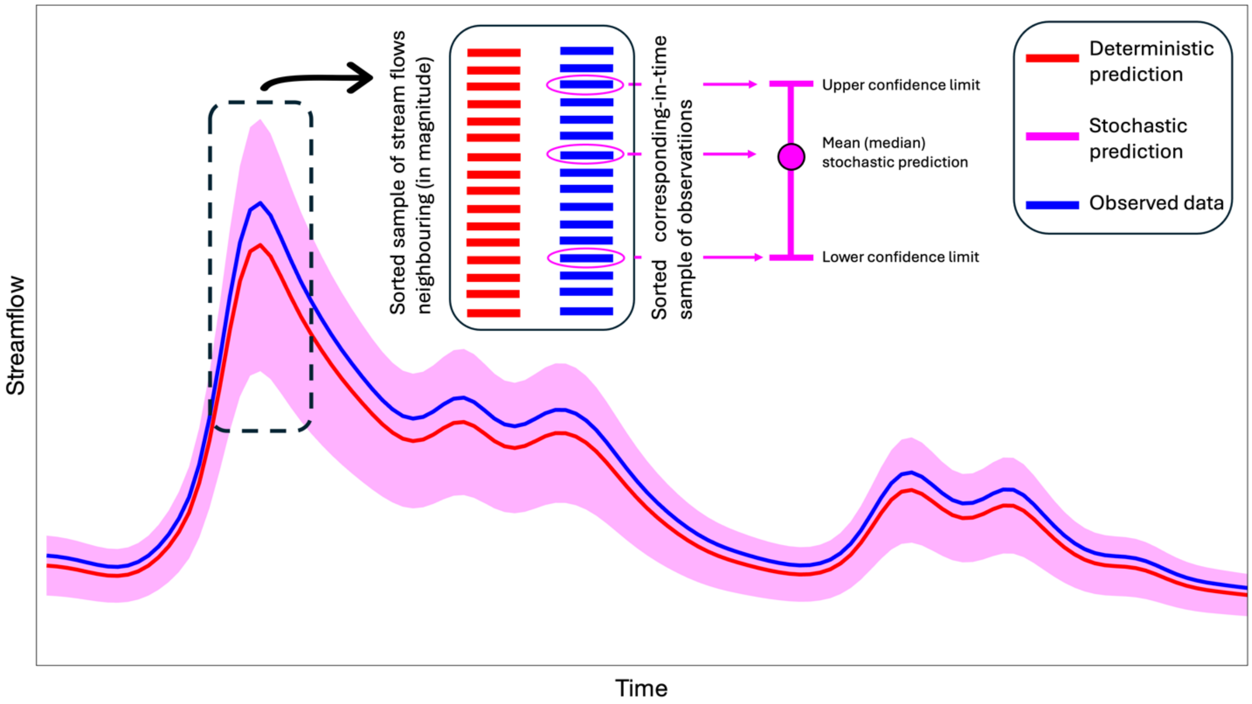

Sorted sample of stream flows neighbouring (in magnitude)
Sorted corresponding-in-time sample of observations
Upper confidence limit
Mean (median) stochastic prediction
Lower confidence limit
Deterministic prediction
Stochastic prediction
Observed data
Streamflow
Time

Figure R01.1. Conceptual illustration of BLUECAT methodology. Blue colour represents observed (streamflow) data, whereas red and pink colours are deterministic and stochastic predictions respectively.

Obs. 7: Line 145, "…Worth mentioning is…": It is worth mentioning ... is better English (also below)

Ans. 7: The sentence in question (and others similar) was (were) updated.

Obs. 8: Line 164, "…normalised amount of information…": Critical here is how H(Y|X) is estimated - needs to be more explicit. And for the limits in u (see comment above)
Ans. 8: Caption in Figure 2 (original submission) gave relevant information in this regard that is reported again as follows (see also Fig. 2d summarising RUMI computation methods):

"… Marginal and conditional entropies are computed empirically with bins. The filled cyan band is the area between the 97.5 and 2.5 percentiles of simulation estimated by BLUECAT."

We acknowledge that writing explicitly (and again) this information in the main text might be insightful for the reader, we added the following sentence to the main text:

"Additionally, and with the intention to avoid any additional assumption, marginal and conditional entropies are computed empirically with bins."

Obs. 9: Line 167, "…proper…": Is this proper in the sense of Vrugt?
Ans. 9: "Proper" was in the sense of Jorquera & Pizarro (2023) and Montanari & Koutsoyiannis (2025). Considering that this word does not add any additional information, we decided to remove it from the main text. The sentence in question reads now:

"… Furthermore, an uncertainty measure (in line with Jorquera and Pizarro (2023) and Montanari and Koutsoyiannis (2024) uncertainty quantification proposal) of the stochastic model computed with BLUECAT…"

References:
Jorquera, J., & Pizarro, A. (2023). *"Unlocking the potential of stochastic simulation through Bluecat: Enhancing runoff predictions in arid and high-altitude regions"*. Hydrological Processes, 37(12), e15046.

Montanari, A., & Koutsoyiannis, D. (2025). *"Uncertainty estimation for environmental multimodel predictions: The BLUECAT approach and software"*. Environmental Modelling & Software, 106419.

Obs. 10: Line 176, "…it is possible…": It is not possible????????? You can draw it (as in Fig 2a) but how can that actually arise? Also Figure 2c implies perfect data (that unrealistic ideal case again) - but you certainly do not have that here, if only because of daily discretisation and input interpolation errors..
Ans. 10: We thank reviewer's comment and share his view in terms of ideal and non-ideal available data. However, the intention behind the Figure 2 was to illustrate conceptually the reasoning behind RUMI formulation. We strongly believe this figure helps to clarify RUMI computation steps and line of thought (even though they can be extremely conceptual).

Obs. 11: Line 292 – 293, "…The latter with the intention to quantify the metric uncertainty…": Not clear what you mean here
Ans. 11: The sentence in question was rewritten with the intention to avoid confusion. It reads now:

"The latter with the intention to quantify the sensibility of RUMI as a function of those additional methodologies."

Obs. 12: Line 294, "high-quality input data and length of the time series": But that was surely not the case for some of these catchments? Using daily data itself introduces discretisation issues that will add to simulation errors.
Ans. 12: Used data were taken from the CAMELS-CL database (which, indeed, followed a quality control analysis), ensuring high-quality data. Additionally, in section "2.2 Data", an additional filter was applied with the intention to consider: a) near-natural hydrological regimes catchments; b) catchments with no more than 25% of missing data; and, c) standardisation of the time series length, finally being from 1990 to 2018.

Time discretisation was out of the scope of the presented work, even though it was highlighted as possible future research at the end of the conclusion section (see Ans. 15).

Obs. 13: Line 295, "…upgrade from the deterministic to the stochastic model)…": Do you mean in terms of sampling the m1 and m2 values to get at the local distribution of errors?
Ans. 13: Not the local distribution of errors but the local distribution of the variable under analysis (at each time step). See Ans. 6 and Figure R01.1.

Obs. 14: Line 297, "…might be…": might be? you have not said how it was estimated?
Ans. 14: See Ans. 8. Additionally, codes for RUMI computation in R and Matlab were provided with this submission and, therefore, any additional metric computation doubt can be resolved looking at the codes. The latter gives transparency and foster reproducibility of results.

Obs. 15: Line 305, "…additional research…": add testing with higher resolution data to reduce discretisation issues in smaller catchments
Ans. 15: Thank you for this suggestion. We added it to the conclusion section, reading now:

"Possible additional research is mentioned as follows: (a) Testing the RUMI-based approach with other rainfall-runoff models (lumped, semi-distributed, and distributed hydrological models); (b) Testing the RUMI-based approach under other hydroclimatological catchment characteristics and in a higher number of catchments; (c) Testing alternative uncertainty quantification methods; (d) Exploring the impact of varying data quality on RUMI performance to establish guidelines for data requirements; (e) Testing with higher resolution data to reduce discretisation issues; and, (f) Exploring the applicability of the RUMI in other disciplines such as meteorology, environmental science, and ecology where modelling and uncertainty quantification are critical."

Response to reviewer 02 – Manuscript: *"Combining uncertainty quantification and entropy-inspired concepts into a single objective function for rainfall-runoff model calibration"* by Pizarro, Montanari & Koutsoyiannis

Below, we list reviewer's comments verbatim in **bold**, followed by responses to these comments in blue. We *italicise* the revised additions to the manuscript.

Thank you again for your helpful feedback!

Reviewer 02: Salvatore Grimaldi

Obs. 1: The manuscript proposes an innovative metric (named RUMI) to support the calibration of hydrological models, based on a combination of two approaches: BLUECAT and Mutual Information (MI). The authors evaluate the performance of the proposed metric through an extensive case study analysis, comparing RUMI to an established metric, Kling-Gupta Efficiency (KGE).

The topic is particularly interesting, and the proposed approach is promising due to its potential for providing more effective calibration and incorporating uncertainty evaluation through the BLUECAT component.

While I am inclined to recommend publication, I have several comments and suggestions for the authors, detailed below.
Ans. 1: Thank you for your comments and suggestions provided.

Obs. 2: The main issue to address concerns the paper's structure and organization.

Although the abstract clearly conveys the manuscript's aim and content, the subsequent sections may leave the reader disoriented.
Ans. 2: Manuscript sections were rethought and updated accordingly.

Obs. 3: Introduction:
The Introduction should be revised to focus more on the calibration problem, highlighting available metrics and their limitations. The new proposed metric should then be introduced, emphasizing its innovative aspects and added value. The two components, BLUECAT and Mutual Information, could be briefly mentioned at the end.
Ans. 3: Introduction section was rethought, focusing on calibration issues, available metrics and their limitations, and uncertainty assessment.

<u>Obs. 4</u>: Section 2.1:
The information on where the GR4J model is available could be moved to the Appendix, as it is not directly relevant to the manuscript. Instead, it would be more helpful to provide additional details about the model itself. Currently, we only know that it has four parameters and two storage modules. The calibration strategy is omitted and should be described, particularly how the two metrics (RUMI and KGE) are applied in the calibration process. It is also essential to confirm that the model is suitable for daily-scale applications and all spatial scales. For example, the reader might question whether the GR4J model performs well at a daily scale for a 35 km² watershed, which could undermine the relevance of the evaluation and comparison.

<u>Ans. 4:</u> Thank you for the comment. We strongly believe model information is important to mention in this methods section and mainly because it assures reproducibility of results (for instance, it was used MARRMoT in Matlab and not the airGr package in R). Additionally, the GR4J model is a widely used rainfall-runoff model with many model applications worldwide (see, e.g., Figure R02.1 with a Web of Science search with the term "GR4J" and Figure R02.2 with known places where the airGr package is used). From these figures is possible to visualise the number of research studies (and, therefore, the general knowledge of the model) and the widely used application of the GR4J model in different locations and catchment characteristics.

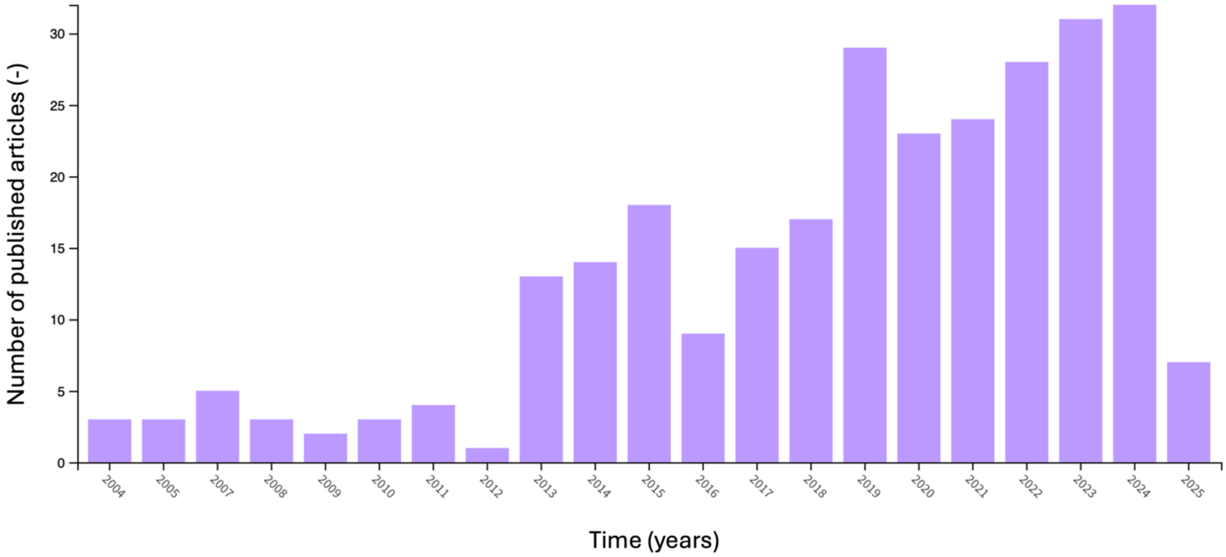
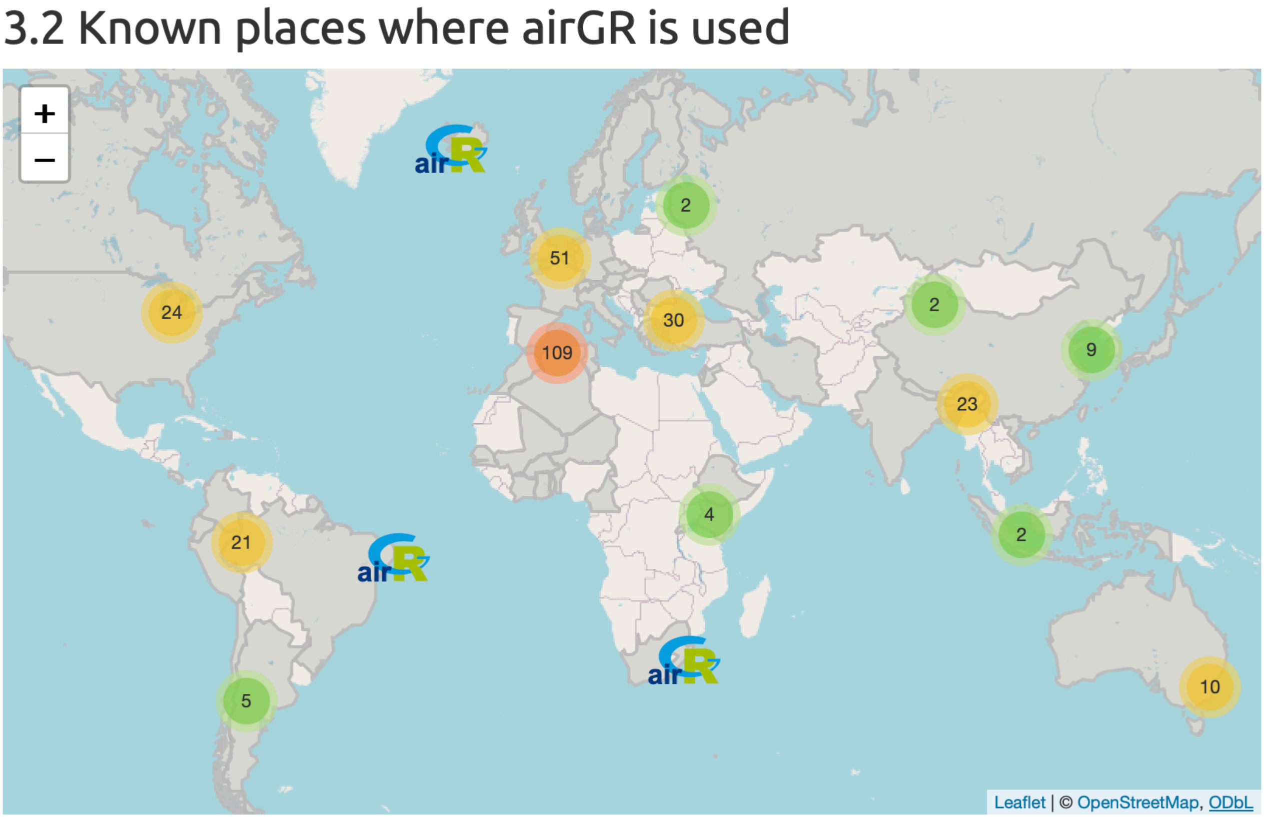


Figure R02.1. Web of Science search of the term "GR4J". Number of published papers as a function of time (from 2004 to the present).



3.2 Known places where airGR is used

Figure R02.2. Known places where airGr is used. Source: https://hydrogr.github.io/airGR/

Despite the latter, we added the following sentence in section 2.1. The sentence in question reads now:

"The GR4J model has four parameters and two storage components. Its primary purpose is to represent processes such as vegetation interception, time delays within the catchment, and water exchange with neighbouring catchments (for detailed information of the GR4J model, see Perrin et al., 2003; and the official website of the developers: https://webgr.inrae.fr/eng/tools/hydrological-models)."

Obs. 5: Section 2.3:
The section title is not informative. A short introduction should be added to remind readers that RUMI is a combination of BLUECAT and MI.
Ans. 5: See Ans. 2.

Obs. 6: Line 138: The text refers to a flowchart that is missing. Please include it or clarify the reference.
Ans. 6: Thank you for rising this issue. We added Figure R02.3 to the main text to better explain how BLUECAT works. The figure in question reads now:

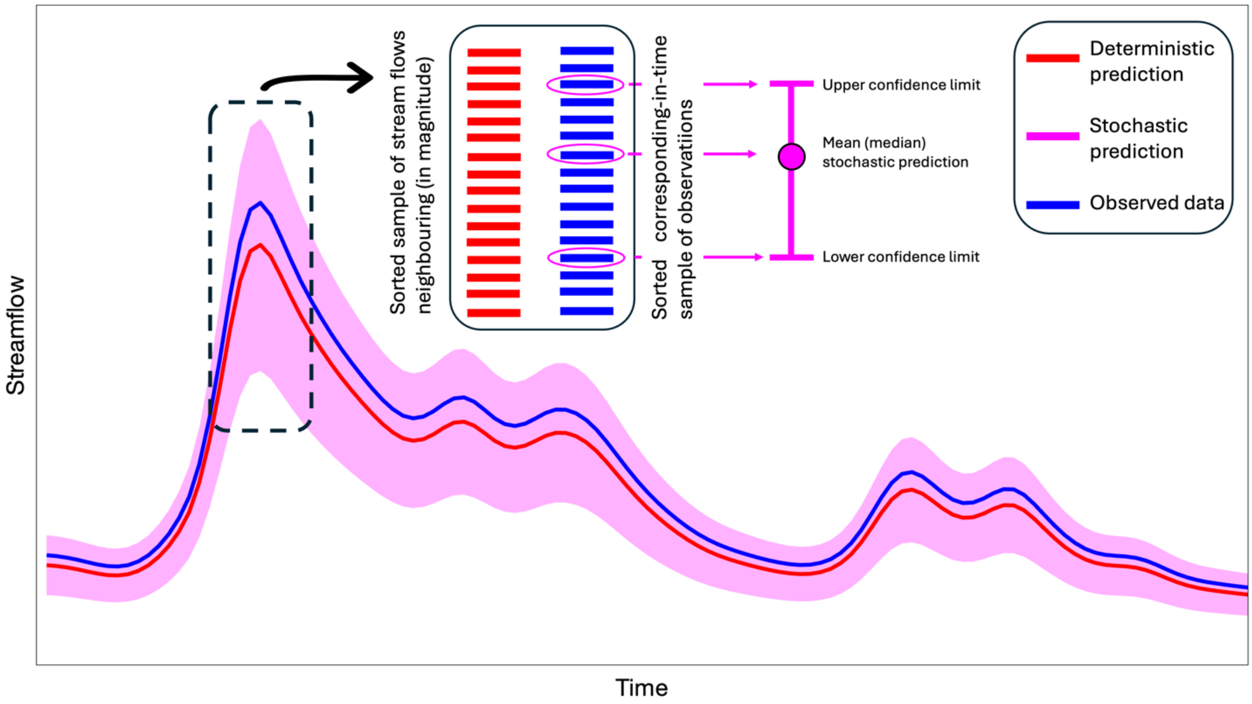

Streamflow
Time
Sorted sample of stream flows neighbouring (in magnitude)
Sorted corresponding-in-time sample of observations
Upper confidence limit
Mean (median) stochastic prediction
Lower confidence limit
Deterministic prediction
Stochastic prediction
Observed data

Figure R02.3. Conceptual illustration of BLUECAT methodology. Blue colour represents observed (streamflow) data, whereas red and pink colours are deterministic and stochastic predictions respectively.

Obs. 7: Section 2.4:
Both the section title and its content should be revised for clarity. This section contains mixed information on data (which might be better placed in Section 2.2), the comparison metric KGE, and the evaluation methods. These topics should be better organized and clearly separated.
Ans. 7: We restructured the methods section, adding a new subsection called: *"Calibration and validation strategies"*.

Obs. 8: Section 3 (Results):
This section could be better organized by providing a more detailed explanation of the plots. Consider using more informative and communicative visualizations.
Ans. 8: Text and plots were revised accordingly throughout the manuscript.

Obs. 9: Lines 220–230: These lines are somewhat confusing. It might be more effective to first show the validation dataset and then focus on two specific years.
Ans. 9: We appreciate this comment and showing first results of the whole analysed dataset or an example of simulation is a matter of writing style. We believe many readers will be more interested to see how is a RUMI-based simulation (at first glance) and then, to see results of the complete analysed dataset.

Obs. 10: Table 2: Is this table necessary? Perhaps violin plots in Figure 4 would suffice. If kept, the table should include comments on specific values (e.g., 1755 and other notable values).
Ans. 10: The main text was updated accordingly, including comments on extreme (high and low) values.

Obs. 11: Table 3: The meaning of the values presented is unclear and needs further clarification.
Ans. 11: Additional information was added to the main text to clarify what is presented in Table 3. Additionally, Table 3 was also modified (table without colours) following HESS table guidelines.

Obs. 12: Figure 5: The phrase "only for illustration purposes" in the caption is confusing. Consider keeping only subplots c.1, c.2, d.1, and d.2, as they show a clear performance difference between the two methods.
Ans. 12: The *"only for illustration purposes"* phrase was removed from the figure caption. We keep our willingness to maintain all the subplots and shown cases as they show evidence for both performance similarities and differences (usually with a better performance in RUMI-based simulations).

Obs. 13: Section 4:
This section needs to be improved for clarity and coherence. Emphasize the significance of the results and their practical implications.
Ans. 13: Section 4 (Strengths and limitations) is not a usually thought "discussion" section. Indeed, it focused on the strengths and limitations of the proposed method with the intention to give clarity to the reader (from the authors' point of view). The section in question is divided in two main paragraphs, each of them focused on the strengths (first paragraph) and limitations (second paragraph). Table R02.1 shows (almost in its textual form) strengths and limitations provided in section 4:

Table R02.1. Summary of strengths and limitations provided in Section 4.

Strengths	**Limitations**
Metric: Uncertainty quantification integration in the calibration process	Methods: Use of other methodologies for uncertainty quantification
Metric: Normalisation of the metric for comparability	Metric: RUMI calculations can be computationally intensive
Case studies: 99 catchments in a pseudo-natural hydrologic regime that covers different macroclimatic zones	Methods: Need of high-quality input data and length of time series
Methods: Simplicity of the approach	Metric: the method assumes that observed and simulated stream flows can be effectively described by these measures, which may not capture all dependencies and non-linearities.
Methods: Use of 50 hydrological signatures	Metric: entropy and mutual information might be sensitive to outliers

Obs. 14: Section 5:
The Conclusions should be rewritten. The first sentence belongs at the end, as a concluding remark. The section should summarize key findings and highlight future research directions.
Ans. 14: The sentence in question was intentionally written at the beginning of the conclusion section to reinforce the message the authors want to give to the readers. Key findings and future research directions (possible additional research) are already in the conclusion section. Additionally, and following reviewer's 01 (Keith Beven) suggestion, one additional research was incorporated (See Ans. 15 of reviewer 01).

Obs. 15: Additional Suggestions:
To thoroughly evaluate the proposed RUMI, consider conducting a virtual test in which parameters are assigned to the model and then calibrated using both metrics (RUMI and KGE). The comparison would then focus on the parameter estimates rather than the streamflow time series. This is a common approach that helps isolate model-independent performance and avoids limiting the conclusions to the GR4J model.
Ans. 15: Thank you for this suggestion. Indeed, testing what is proposed by the reviewer as well as using a higher number of catchments and other rainfall-runoff models are future research directions to avoid limiting conclusions to one model or specific catchment characteristics. We will tackle these issues in future research.