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.

Obs. 4: 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.

Ans. 4: 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.

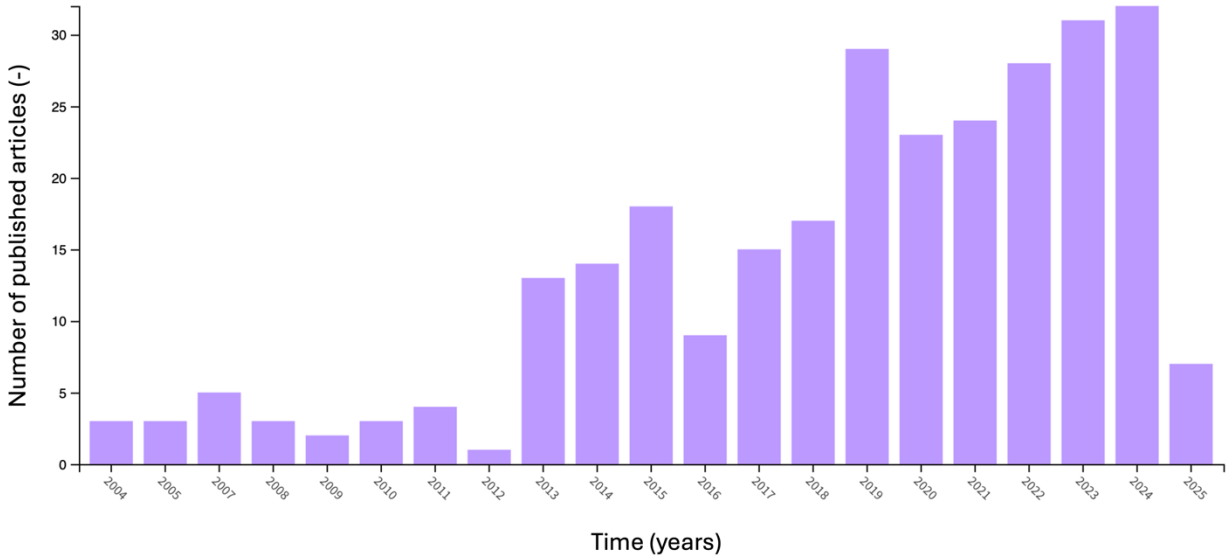

Figure R02.1. Web of Science search of the term "GR4J". Number of published papers as a function of time (from 2004 to the present).

3.2 Known places where airGR is used

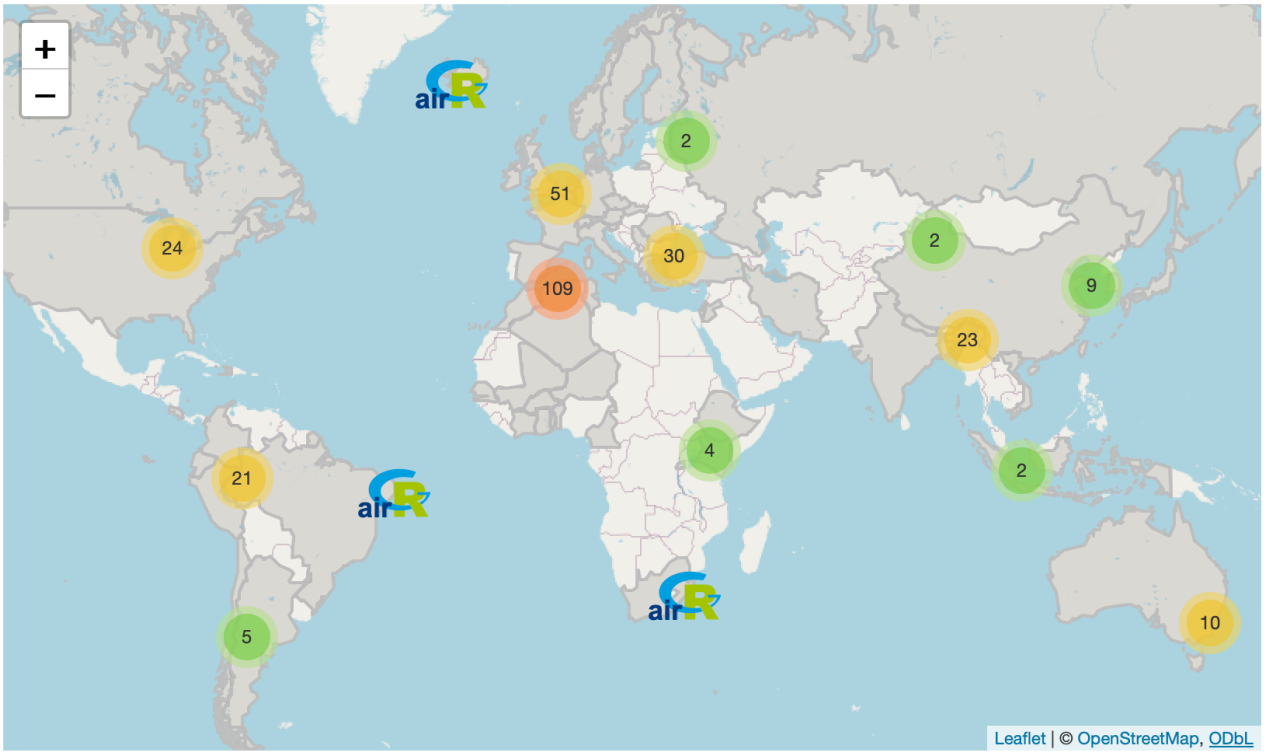

Figure R02.2. Known places where airGr is used. Source: https://hydrogr.github.io/airGR/

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:

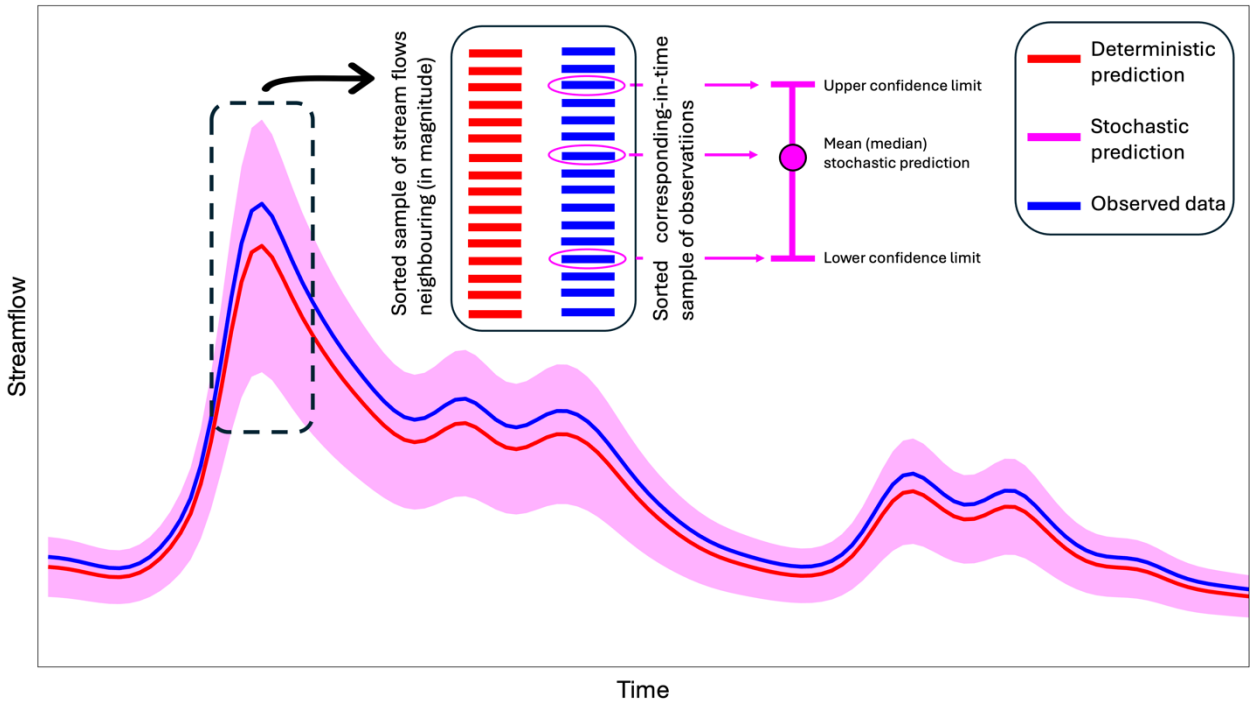

Figure R02.3. Conceptual illustration of BLUECAT methodology. Blue colour represents observed (streamflow) data, whereas red and pink colours are deterministic and stochastic predictions respectively.

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.