Enhancing physically based and distributed hydrological model calibration through internal state variable constraints

Frédéric Talbot[1], Jean-Daniel Sylvain[2], Guillaume Drolet[2], Annie Poulin[1], Richard Arsenault[1]

[1] Hydrology, Climate and Climate Change Laboratory, École de technologie supérieure, Université du Québec, Montréal, H3C 1K3, Canada

[2] Direction de la recherche forestière, Ministère des Ressources naturelles et des Forêts, Québec, G1P 3W8, Canada

Correspondence to: Frédéric Talbot (frederic.talbot.2@ens.etsmtl.ca)

RC2 (https://doi.org/10.5194/egusphere-2024-3353-RC2)

This manuscript titled “Enhancing physically based and distributed hydrological model calibration through internal state variable constraints” investigates the effectiveness of various calibration approaches within the Water Balance Simulation Model (WaSiM) to enhance the representation of hydrological variables. The study assesses three configurations: Baseline (BL), Physical Groundwater Model (GW), and Physical Groundwater with Recharge Calibration (GW-RC), which has an add-in of recharge calibration across 34 catchments in Southern Quebec, Canada. The research provides valuable insights into the importance of multi-variable calibration frameworks in developing robust models capable of adapting to anticipated hydrological shifts due to climate change. However, it is too long!

Dear Dr. Modiri,

We thank the reviewer for the thorough evaluation and insightful feedback on our manuscript. The comments are invaluable in refining our paper and we are committed to enhancing its clarity and impact. As for the length of the paper, please see our suggestions below as how we propose to shorten it.

Major Comments:

- Abstract:

While the abstract effectively conveys the general research objective and findings, it lacks specific quantitative data. It relies heavily on vague terms and subjective assessments, making it difficult for readers to grasp the magnitude and significance of the improvements achieved fully.

Generally, I suggest you revise it.

We appreciate the critique concerning the lack of specific quantitative data in our abstract, which could limit the reader's understanding of the improvements and findings. Acknowledging this, and in response to both your major and minor comments, we will undertake a thorough revision of the abstract to add quantifiable data that underscore the scope and impact of our research findings. Specifically, we will:

1. Eliminate redundancy, such as the repeated phrase "on the representation of hydrological variables,", on lines 9 and 11.

2. Clarify the study's objectives to provide clear direction regarding the purpose of our research.
3. Include a concise summary of the key findings.
4. Remove vague terms and replace them with specific and measurable outcomes from our study.
5. Include quantifiable results to support the key findings and the study's conclusion.

Methodology:

The authors should provide more details on the selection criteria for the 34 catchments used in the study. While some information is given in section 2.1, a more comprehensive explanation of why these specific catchments were chosen would strengthen the methodology.

This is a good point. We acknowledge the need for a more detailed explanation in Section 2.1 of our manuscript. The catchments were selected based on several criteria to ensure the integrity of the hydrological processes studied: they are free from dams and reservoirs, located away from large urban areas to maintain natural hydrological conditions, and each has a hydrometric station with comprehensive data from 1981 to 2010 for robust model calibration and validation. We also aimed for geographic diversity to cover various climatic conditions across Quebec and, where possible, included catchments covered by the PACES project for data consistency. We plan to revise section 2.1 to provide clarity on our methodological choices.

The selected basins are relatively medium-sized (between 100 and 10,000 square kilometers), which may limit the generalizability of the findings to larger or smaller basins.

We acknowledge the concern regarding the potential limitations in generalizability to very large or very small basins. The chosen size range, from 100 to 10,000 square kilometers, encompasses a broad spectrum that represents a significant portion of catchments typically analyzed in regional hydrological studies. To address this limitation, we will include a discussion on the implications of catchment size in the limitations section of our paper, clarifying the scope of applicability of our findings and suggesting directions for future research that might explore the model's performance across differently sized catchments.

The rationale behind using ERA5 reanalysis data instead of ground-based observations for meteorological inputs should be further elaborated. Since ERA5 recorded underestimating winter precipitation and bias in convective precipitation, it would be better to employ another dataset.

We acknowledge the concerns regarding ERA5's underestimation of winter precipitation and biases in convective precipitation. However, research, such as the study by Tarek et al. (2020), has demonstrated that ERA5-driven hydrological simulations perform comparably to those driven by observational data across Eastern Canada. This study (available at https://doi.org/10.5194/hess-24-2527-2020) showed that, for a broad set of 3138 North American catchments, the results using ERA5 were equivalent to those using traditional meteorological observations in terms of hydrological modeling accuracy over Eastern Canada.

Additionally, we opted for ERA5 because it offers comprehensive spatial coverage that can be particularly advantageous in regions with sparse weather station networks. While it is true that meteorological stations have their biases, which could introduce different limitations, the uniform coverage of reanalysis data like ERA5 provides a consistent baseline for our study.

Given these points, we justify our preference for ERA5 while recognizing that exploring the impacts of using different meteorological datasets could be a valuable avenue for further research. This point will be expanded in our methodology section to better articulate the rationale behind our choice.

- Model Configurations:

While the three configurations (BL, GW, GW-RC) are described, readers would benefit from a more detailed explanation of how they differ in their treatment of groundwater processes. The authors should consider discussing the potential limitations of each configuration and how these might impact the results.

We acknowledge that the distinctions and potential limitations among the three model configurations (BL, GW, GW-RC) were not adequately detailed, particularly in their treatment of groundwater processes. In response to your comments, we will revise Section 2.4 of our manuscript to include a clearer comparison of these configurations. Specifically, we will enhance the description of how each configuration manages groundwater processes, and we will discuss the implications of these methodologies on the results.

- Calibration and Validation:

The split-sample approach for calibration and validation is appropriate, but the authors should discuss any potential impacts of climate non-stationarity on this approach, given the study's focus on climate change adaptation.

We appreciate the comment on the need for a detailed discussion on the potential impacts of climate non-stationarity on our split-sample approach for calibration and validation. We will include a cautionary note regarding the potential limitations due to climate non-stationarity in section 4.4, acknowledging that despite our efforts, the model might not be as robust as anticipated under varying climate conditions. This addition will help clarify the implications and limitations of our approach in the context of climate change adaptation.

Given that you modified the lower and upper boundaries of the model parameter by 10% (L310), a direct comparison with the calibration results of the BL configuration using default parameters might not be entirely fair. Simulating WaSim for all configurations using the adjusted parameter range would be beneficial to ensure a more consistent evaluation.

We understand this comment. To clarify, the GW-RC configuration employs a two-step calibration process. First, it initially uses groundwater recharge data to constrain the parameter range. This begins with a pre-calibration phase applying an objective function that gives more weight to groundwater recharge metrics—20% for the standard deviation of recharge and 10% for mean annual recharge. Second, after determining a set of parameters from this pre-calibration, we adjust these values by ±10% to define a new, narrower parameter range for the final calibration. This adjustment is applied only to 5 of the 17 calibration parameters. We also make sure that the new parameter range remains within the original bounds. If a boundary exceeds the original range, we adjust it to maintain the parameter within its initial limits. The parameter constraints are based on groundwater recharge values. Since configurations BL and GW do not integrate recharge in their calibration, they cannot utilize the constrained parameter range as described in this study.

Employing the same constrained parameter range across all configurations would mask the specific impact of including recharge in the calibration, as it would diminish the ability to distinctly evaluate the benefits of this approach. Furthermore,

the developed method reduces the degrees of freedom of the GW-RC model, and as such, it is penalized compared to the other models, and thus the results obtained are conservative. If we also shrink the parameter range for the BL model, for example, then the streamflow score can only be reduced as the obtained parameter set would be a subset of the parameter space of the original model. Perhaps the processes would be better represented, however it would not be possible to estimate these bounds without using the recharge data, and thus it is not possible to implement this model in a fair manner.

To enhance clarity and provide a comprehensive understanding, we will expand the description of this calibration methodology in Section 2.4.3 of the manuscript, ensuring a detailed explanation of the process.

Given that the manuscript focuses on WaSim performance, including the computational cost and time associated with each configuration is crucial. This information will be highly valuable for other researchers, allowing them to estimate the resource investment required to achieve comparable improvements in water balance closure.

We recognize the importance of detailing the computational resources required for each configuration of the WaSiM model as highlighted in the comment. Accordingly, we will enhance Table 5 to incorporate the computational demand for each configuration, providing clarity on computational cost in CPU-years (totalling 35 CPU-years on 4.5 GHz CPUs, for your benefit).

Table 5. Summary of configurations

Settings	BL	GW	GW-RC
Groundwater Modelling	Conceptual within unsaturated zone sub-model	Physically based within the groundwater sub-model	Physically based within the groundwater sub-model
Calibration Parameters	17 parameters (including K_B and Q_0)	17 parameters (including Kol and K_{XY})	17 parameters (including Kol and K_{XY})
Precalibration	N/A	N/A	200 simulations at 1000 meters followed by 50 simulations at 250 meters
Calibration	1000 simulations at 1000 meters followed by 50 simulations at 250 meters	1000 simulations at 1000 meters followed by 50 simulations at 250 meters	1000 simulations at 1000 meters followed by 50 simulations at 250 meters
Objective function	Kling-Gupta efficiency	Kling-Gupta efficiency	Constrained Kling-Gupta efficiency
Computational demand	10 CPU-year at 4.5 GHz	10 CPU-year at 4.5 GHz	15 CPU-year at 4.5 GHz

CPU-year : A CPU-year is the effort of a CPU running for one year.

- Results Presentation:

Figure 4 highlights the significant shift in the proportion of surface runoff and interflow. Please elaborate on the specific factors that influenced this shift during calibration, particularly considering the inclusion of groundwater recharge in the model.

To address this comment, we will add a sentence in the results section near figure 4 by explicitly stating that the specific factors influencing these shifts, particularly the role of groundwater recharge during calibration, are comprehensively discussed in the discussion section of the manuscript.

The results presented in Figures 3-10 are generally clear, but some figures (e.g., Figures 5, 6, 7) could benefit from additional explanation in the text to help readers interpret the complex information presented.

We agree with this point. In response, we will conduct a thorough review of the text in the results section to ensure that it effectively communicates the major points and provides clearer guidance on interpreting the data presented in these figures.

A more in-depth discussion of the spatial variability in model performance across the 34 catchments would enhance the study's insights, especially when compared with PACES.

This comment highlights a valuable aspect that can indeed enhance the study's insights significantly. We propose to add a few sentences in the discussion to address spatial variability across the 34 catchments. This section will explore the performance variations, providing a deeper understanding of the model's effectiveness in various hydrological settings.

My understanding differs from your conclusion in Figure 10. None of the configurations are aligned with PACES, except for a case in Noire. I would say that the lowest difference is between GW-RC and PACES. In general, I found PACES recharge different than the applied three configurations in this research, according to Figure C1.

We appreciate the observations regarding Figure 10. You're correct in noting the discrepancies between the model configurations and PACES. We will revise the relevant text to ensure it clearly states that the GW-RC configuration provides "the lowest difference" with PACES instead of "align more closely".

- Climate Change Implications:

While the study mentions the importance of the findings for climate change adaptation, a more specific discussion on how the improved model configurations might be applied in climate change impact assessments would strengthen the paper's relevance.

We thank the reviewer for this suggestion. In response, as stated above, we will expand the discussion section of our manuscript to emphasize how the refined model configurations aim to improve hydrological process representation, thereby enhancing model robustness in the face of climate change. Additionally, we will incorporate a cautionary note about the potential limitations of our models due to climate non-stationarity. This will help clarify the expected robustness of our models under varying climatic conditions and outline the broader implications for climate change adaptation.

Minor Comments:

- Abstract:
  - It exhibits some redundancy, such as the repetition of "on the representation of hydrological variables" in lines 9 and 11.
- The abstract could benefit from a more precise statement of the study's objectives and a more concise summary of the key findings
- Vague Language:
  - "significantly refines the model's ability to depict subsurface processes"
  - "minimal emphasis on recharge"
  - "small and targeted calibration adjustments"
  - "marked improvement"

- “enhancing the precision”

- Lack of Quantifiable Results:
  - No specific metrics are mentioned (to quantify the improvement in model performance.
  - No specific values are given for the improvement in groundwater recharge representation.
  - No indication of how the “minimal emphasis” on recharge was defined or quantified.

As mentioned in our response to the major comments, we will thoroughly revise the abstract to address the highlighted issues.

- Methodology:
  - Figure1: Visualising the selected case studies within a coarser-level basin delineation would be beneficial. This would provide context, as the presence of a river traversing the study area can significantly influence catchment behaviour.

    To address the suggestion, we propose to revise Figure 1 to add a new panel zoomed in on the Matane catchment to provide more detailed information about the case study.
  - Table 5: Table’s style is totally different from the other presented tables.

    We will adjust the formatting of Table 5 to align with the styling of the other tables presented in the manuscript and to journal standards.
  - Could you elaborate on the rationale behind conducting 1000 simulations at 1000 m resolution and only 50 at 250 m resolution? What factors influenced the selection of these specific numbers?

    This is a good question. Our decision to conduct 1000 simulations at 1000 m resolution and only 50 at 250 m resolution was based on preliminary testing on catchments Bonaventure and Matane, which demonstrated that this configuration provides the best balance between computational cost and result accuracy. We tested 75 and 100 simulations at 250 m resolution, but the results were comparable to those obtained with 50 simulations, making the additional computational expense unjustified.

    The key reason we could limit the 250 m simulations to 50 runs is that the calibrated parameters from the 1000 m simulations transferred effectively to the finer resolution, requiring only a slight refinement. We hope this clarifies our approach, and it will be added to the revised version of the paper.
  - Given the widespread familiarity of the KGE metric within the research community, a detailed definition in section 2.5.1 may be redundant.

    We agree that a detailed definition of the KGE metric in Section 2.5.1 may be redundant given its widespread familiarity within the research community. To address this, we propose removing the detailed definition and merging Sections 2.5.1 and 2.5.2. This will eliminate redundancy and contribute to a more concise manuscript.
  - Furthermore, as per comment CC1 received on November 23rd, the assigned weights in section 2.5.2 (L349) require more comprehensive scientific justification and supporting literature.

This is a valid point. Since this methodological step is novel, there is no specific literature directly supporting the assigned weights. To determine these values, we conducted multiple tests with various weight combinations on two test catchments. Our results showed that assigning 20% to the standard deviation of recharge and 10% to the mean recharge provided the best trade-off, ensuring recharge values remained realistic while maintaining acceptable KGE scores. For calibration, a weight of 4% on the recharge standard deviation was sufficient to preserve adequate recharge estimates while achieving strong KGE values, basically providing an incentive to ensure proper process representation without sacrificing too much performance on the streamflow simulation. This will be explained in the text.

- The manuscript would benefit from considering alternative objective functions besides KGE for streamflow. As suggested in this paper (https://gmd.copernicus.org/articles/11/1873/2018/), using SPAtial EFficiency (SPAEF) could enable the evaluation of multiple hydrological components when you utilise distributed hydrological models. This would provide a more comprehensive assessment of model performance.

This is a good point. We did not use a spatial objective function like SPAEF because we lacked sufficient spatially distributed observations to properly calibrate the model. Applying SPAEF could be an interesting avenue for future studies, particularly when using remote sensing data for calibration. A sentence will be added to section 4.4 of the discussion.

However, to provide a more comprehensive evaluation of model performance across multiple metrics, we propose adding a table in Appendix A (Table A2) that presents calibration and validation results for different configurations. This table includes several key performance metrics beyond KGE, allowing for a broader assessment of model performance. See Table A2 below.

Table A2. Multiple metrics values during calibration and validation periods, for the three configurations.

Metric		Calibration			Validation		
		BL	GW	GW-RC	BL	GW	GW-RC
KGE	μ	0.852	0.852	0.799	0.816	0.820	0.772
	σ	0.034	0.036	0.050	0.055	0.049	0.056
Pearson Coefficient	μ	0.855	0.855	0.804	0.844	0.845	0.797
	σ	0.034	0.036	0.050	0.040	0.039	0.049
Bias ratio	μ	0.998	0.990	0.985	1.030	1.024	1.018
	σ	0.021	0.024	0.023	0.054	0.052	0.050
Variability ratio	μ	0.996	1.005	1.020	1.022	1.028	1.055
	σ	0.023	0.013	0.027	0.083	0.075	0.079
NSE	μ	0.704	0.706	0.603	0.677	0.679	0.558
	σ	0.059	0.076	0.091	0.091	0.073	0.120
RMSE	μ	20.926	20.749	24.317	22.877	22.621	26.545
	σ	11.018	10.681	13.055	12.594	11.665	13.870
Percent bias	μ	0.155	0.074	1.263	-3.563	-2.760	-1.756
	σ	1.399	2.209	2.132	5.792	5.628	5.650
MAE	μ	11.364	11.198	13.287	12.444	12.045	14.432
	σ	6.621	6.259	8.106	7.966	7.043	8.638

- Results
  - I would like to know if the same results would be obtained by switching the calibration and validation periods, as indicated in Figure 2. Given that the KGE values for all three setups are relatively close, I am uncertain about the potential benefits of using GW.

    This is a valid concern. We do not have simulations for switched calibration and validation periods, as this would require recalibrating the entire project, which would take several months of computation on our compute infrastructure (see computation time in revised table 5 above). However, to minimize the risk of overfitting to a specific time period, we used an ensemble of 34 catchments distributed across Southern Quebec. Given the diversity of catchments and the three different configurations tested, it is reasonable to hypothesize that the results would remain similar if the calibration and validation periods were reversed, as the likelihood of all configurations overfitting to a particular time period is minimal. Also, the fact that the calibration and validation scores are similar indicates that the models were not overfitted and that expected errors are a good proxy of the generalization error (Hastie et al., 2009) (https://doi.org/10.1007/b94608), and can be used directly, as described in Arsenault et al. (2018). (https://doi.org/10.1016/j.jhydrol.2018.09.027)

- In Figure3, consider adding each variable's total mean or sum of observations to enhance the visual comparison. This will allow readers to contextualise the calibration and validation boxplots by providing a reference point for the overall data distribution.

  This is a good suggestion. We have modified Figure 3 accordingly. Specifically, we have enhanced the visualization by adding a wider median line for each boxplot and incorporating a grid in each subplot to facilitate comparison between periods and configurations. These adjustments improve readability. Please see the updated figure below.

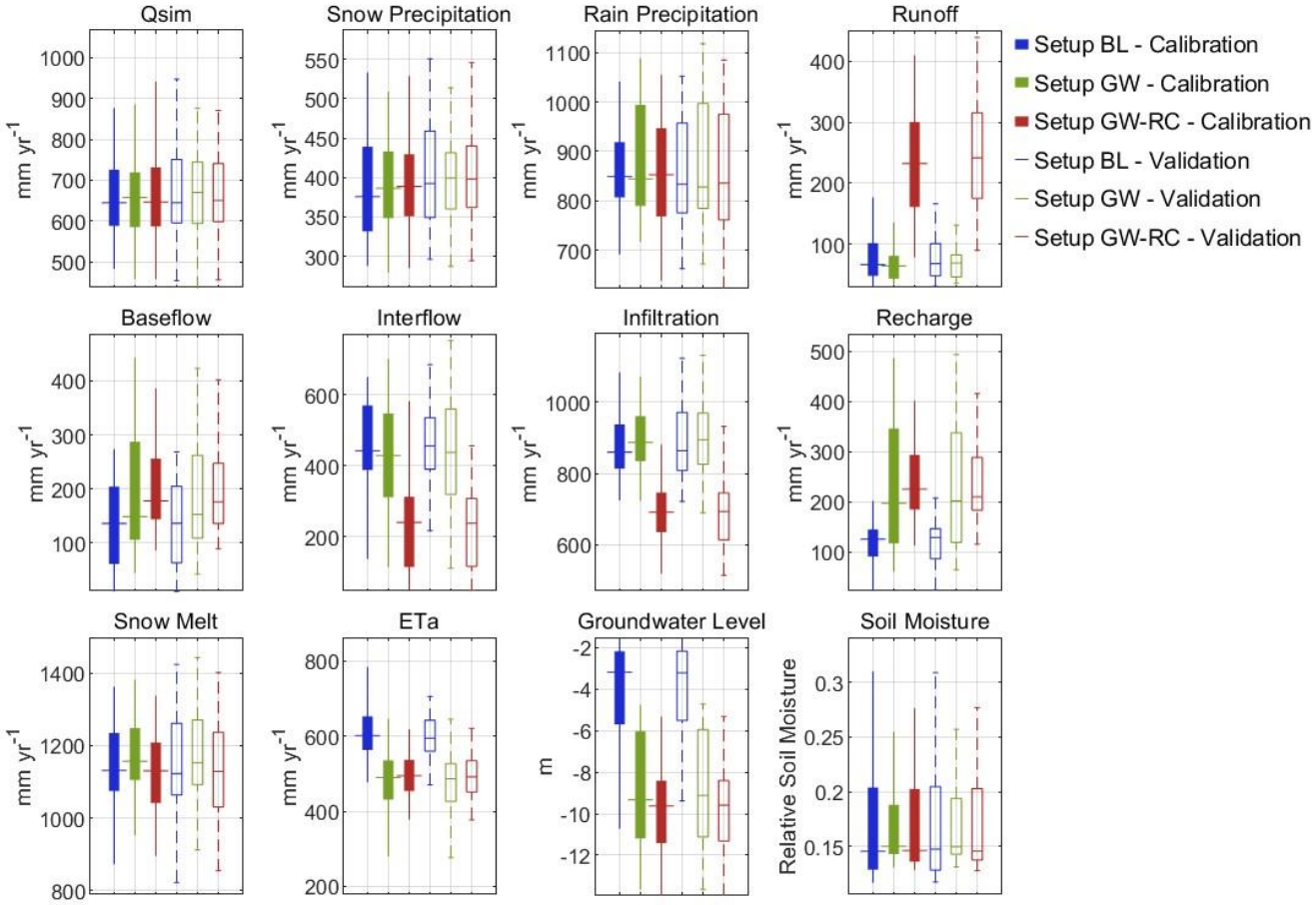


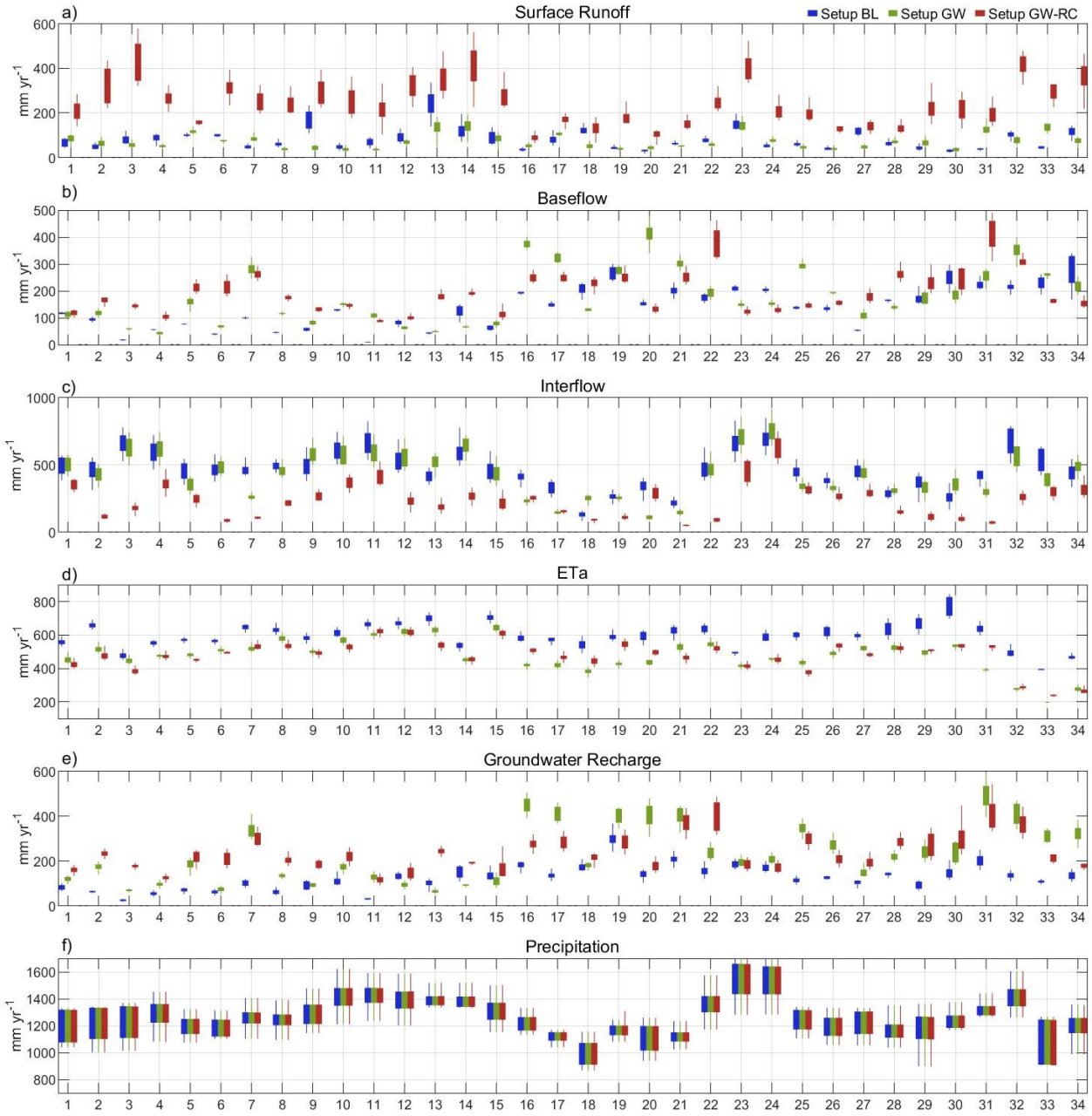
Figure 3. Boxplots illustrating annual totals (means for groundwater level and soil moisture) variability of model internal variables. These boxplots detail the variability of key hydrological variables modeled with the different configurations, for calibration and validation periods and for all catchments.

- The manuscript should provide an explanation for the lack of differentiation in baseflow between configurations GW and GW-RC, as noted in L430

  This is a relevant point. While the baseflow of configurations GW and GW-RC appears similar, the differences are statistically significant. This outcome is expected, as both configurations use the same groundwater module, with GW-RC differing only in its calibration method, which accounts for the small variations observed. However, when compared to BL, both configurations exhibit similar baseflow behavior, indicating that the choice of model configuration primarily drives the differences in baseflow across the three setups. To clarify this, we propose adding a brief explanation at L430.

- Figure 5 reports a difference of around 200 mm/y across all variables among the three configurations for all 34 catchments. To facilitate water balance closure assessment, consider adding a subplot for precipitation data for each basin.

  This is a good point. We have modified Figure 5 to include precipitation data in the last subplot to facilitate the assessment of water balance closure. However, since all three configurations use the same precipitation data, the values remain identical across configurations. A revised version of Figure 5 is provided below.



265

Figure 5. Boxplots of annual values for key hydrological variables predicted by WaSiM for the 34 catchments and three configurations.

- The current explanation of Figure 6 was neither informative nor relevant to my perspective. It needs to emphasise the significance of the figure.

  This is a valid concern. While we recognize that it may not be of interest to all readers, we believe it is a valuable addition, as it directly links model parameters to hydrological processes, offering essential insights for WaSiM users. In order to maintain a concise article, we propose moving Figure 6 to the Appendix. This approach allows us to reduce the length of the main text and ensure that readers who are particularly interested in these modeling details can still access the figure.

- Figure7, is the x-axis long-term mean of Q, or are they for a given year? The problem is between October to December in validation period. In the rest, I see no significant differences. Maybe you could drop this figure.

  Figure 7 presents the mean annual hydrographs for the calibration and validation periods. As noted, all three configurations show discrepancies with observed streamflow between October and December during the validation period. Given that the figure does not provide significant additional insights and the manuscript is already lengthy, it will be removed to streamline the article.

  Also, since you have gaps in some of the frozen months (L130), how did you consider them in the likely monthly discharge time series?

  This is a good point. Not all catchments had missing data during the frozen months. To minimize the impact of missing data, we selected calibration and validation years to ensure most catchments had complete records. With this approach, only three out of the 34 catchments had missing data. For these specific catchments, we adjusted the calibration and validation periods to focus on years without gaps. As a result, the final analysis does not include years with missing data. A sentence to this effect will be added at L266 to clarify.

- Language and Style:
  - The manuscript is generally well-written, but there are occasional instances of complex sentence structures that could be simplified for clarity. Thank you for acknowledging the use of ChatGPT-4. The presence of long sentences with numerous commas can be indicative of revised text by an LLM-AI.

    This is a fair point. A thorough verification will be conducted to ensure the text remains clear and concise, and we will have the paper revised by native English speakers to ensure the syntax is less “LLM-y”

- Conclusion:
  - I remain uncertain about the meaning of lines 659-660.

    Line 659-660 will be revised to enhance clarity. The original sentence: *"leads to a more accurate representation of hydrological variables."*

    Will be modified to: *"leads to a representation of hydrological processes that better aligns with expected system behavior."*

This study contributes to hydrological modelling by demonstrating the importance of incorporating internal state variables, particularly groundwater recharge, into model calibration. The authors designed their model well and developed it to have three configurations and further calibrations.

The findings highlight the potential for improved representation of hydrological processes, which is crucial for water resource management and climate adaptation strategies. However, addressing the major and minor comments outlined above would further strengthen the manuscript and enhance its impact on the scientific community. Overall, with appropriate revisions, this paper has the potential to be an important addition to the literature on hydrological model calibration and process representation.

Regarding the initial point about the manuscript length, we propose the following adjustments based on the above responses. Given the widespread familiarity with the KGE metric within the research community, we suggest omitting its definition. Additionally, we will relocate Figure 6 to the appendix and remove Figure 7 from the manuscript. These modifications will result in the removal of two figures and a reduction of approximately 850 words, bringing the manuscript from 9050 words to roughly 8200 words, excluding tables, figure captions, references, appendix, and abstract.

We are eager to implement these changes and believe they will significantly strengthen the manuscript. We thank the reviewer once again for the constructive review, it is much appreciated.

Sincerely,

Frédéric Talbot, on behalf of all authors