Answer to Anonymous Reviewer, Reviewer #2:

This article presents a systematic study of the hydrometric data production process across 1800 active stations operated by Water Survey Canada. An independent (Python-based) approach intended to reconstruct the archived discharge times series based on information and data available from the Aquarius operational software. Interestingly, only 67% of the data could be reproduced (within 5%) from the stage series, the rating curves, and the rating shift curves, the other differences being explained by the significant use of temporary shifts and “overrides”. This exercise is valuable as it quantifies the frequency of operational practices that are more complex to reproduce than the simple application of rating curves (and permanent shifts). In particular, the need for suitable uncertainty computation methods is rightly emphasized.

The paper is generally very well written and well illustrated, however, I fear that its length may discourage some readers less passionate about hydrometry (including data users!) and reduce its impact. I would recommend shortening the paper (20 pages max and 10 figures max). Some technical details (eg multiple data examples) could be cut or moved to Annexes or Supplementary materials.

We thank the reviewer for their positive and constructive feedback on this work.

Similar to our response to editor and reviewer #1's feedback, we aim to relocate specified material from the main text to the appendix, consolidating figures to reduce their count. While we will push to limit the figures, we cannot guarantee a maximum cap.

Both reviewers demonstrate expertise in the field, suggesting that detailed explanations of rating curves or fundamental concepts may be redundant. However, considering a broader audience, establishing foundational knowledge remains crucial. It's been observed that colleagues utilizing discharge for data-driven modeling, such as machine learning, may lack familiarity with rating curves. Recognizing this gap, the first author noted, when doing the literature review, that many works on hydrometric stations tend to lose connection with a wider audience, due to deep diving directly into detailed technicalities.

69-70 the method (IVE) introduced by Cohn et al. 2013 does not relate to rating curves. Not sure about Whalley et al. 2001 and Huang 2018. Please check and remove if need be.

We thank the reviewer, and we will correct the reference.

405 I’m not sure the method presented by Coxon et al. 2015 is actually applied systematically by UK Env agency to establish their rating curves. I don’t think so. Kiang et al. 2018 compared 7 methods for rating curve uncertainty and only the NVE method (in Norway) and the Baratin method (in France) were applied by national hydrological services.

We thank the reviewer, and we corrected the references to methods (also per reviewer #1 request)

What about (seasonal) aquatic vegetation? Is it a problem for Canadian stations (as it is in Europe for instance) and is it managed through temporary shifts? I assume that beaver dams are another issue…

The correlation between shifts and seasons frequently follows a distinct pattern. Colder periods often experience more pronounced intervention from processes like "overrides" and "temporary shifts." Interestingly, in our limited exploration, we found a lack of discussion on vegetation within the operational database, assuming these data have been accurately transferred to the digital operational database.

Additionally, the impact of beavers tends to be localized over smaller areas. While we observed the beaver effect on some experimental catchments, the stations managed by WSC typically cover larger river segments and tributaries.

We will corroborate this information with our WSC colleagues. We'll furnish a more detailed response and update the text accordingly based on our discussions with them.

As stated end of 2.5 and elsewhere (Tab. 3), the central issue is the traceability, reproducibility of the data production process. However, reproducibility and repeatability are different things, and this could be made clearer in the paper. A first step is that discharge computation can be repeated (exactly) using available data and already established rating curves, shifts and overrides (from Aquarius especially): this doesn't seem to be the case as some important information is missing (or not easily retrieved through API), which would be a first issue of incomplete traceability (am I correct?). Another step is that discharge computation can be reproduced (with some permissible variability) from scratch by another equivalent expert: this should be OK thanks to established SOPs and well-trained operators, hopefully, but this statement is not evaluated in this study (through some comparison exercise, for instance). Actually, the problem seems to arise because the assumptions and decisions made by the hydrographers for establishing rating curves, rating shifts, temporary shifts and overrides are not available in a formal way. I think that beyond the statistical technique chosen for uncertainty estimation, this is the key issue: each data production process must be 'modelled' in a reproducible way, even expert-based operations. I agree that some solutions have been published that apply to rating curves and (partially) shifts but not to temporary shifts and overrides. This paper is a first step towards modelling these operations but much more work looks necessary to write mathematical models, especially for 'override' operations, which refer to multiple situations and data estimation techniques. The discussion and comments in the paper could elaborate on this issue more clearly.

This response effectively captures our intended message. However and initially, we refrain from delving deeply into the detailed explanations of reproducibility and repeatability. While Standard Operating Procedures (SOPs) and training ensure the repeatability of results, achieving reproducibility demands a deeper understanding and an underlying explanatory model. This model can serve in estimating discharge by non-experts and form the basis for uncertainty analysis, akin to uncertainty models for rating curves.

To emphasize these points, we aim to incorporate a discussion paragraph addressing the reviewer's comment on the distinction between reproducibility and repeatability. Additionally, if feasible within the constraints of length and conciseness, we try to mention the difference between reproducibility and repeatability within the manuscript more directly.

Another obstacle stated in the paper is the deterministic approach: the uncertainty of stage-discharge measurements must be accounted for, as well as the uncertainty of the input data (stage) and of the rating curves (and more generally the "discharge models"). It looks difficult to quantify the uncertainty of data that have been produced in a fully deterministic approach without reprocessing them. The ideal way to go is to reproduce the data in a probabilistic framework, hence the need for reproducibility…

The reviewer's perspective aligns with our initial project goal, which centered around streamflow uncertainty assessment. Originally, our focus was on contemplating potential uncertainties in streamflow, encompassing stage, measurement, and rating curve uncertainties. However, as the authors delved deeper into WSC practices, the project's focus shifted towards comprehending existing processes rather than solely estimating uncertainty. The process of uncertainty estimation necessitates an explanatory model, which, to our current knowledge, remains absent. Without this foundational model, quantifying uncertainty would be exceptionally challenging. We hope this work is a step toward that "model".

146 Aquatic? Corrected from Aquatics

176 include? Corrected from includes

230 curve corrected from curved

558 To investigate what? Corrected to "To investigate the rating curve uncertainty,"

738 Pytho corrected to "Python"

Once again, we would like to thank the reviewer for their constructive comments.

With kind regards,

Shervan Gharari, on behalf of co-authors, Paul Whitfield, Al Pietroniro, Jim Freer, Hongli Liu, Martyn P. Clark