We thank Dr Gelfan for his comments about the paper.

Reviewers' comments are shown in black. Authors' responses are shown in green

> During the review process, we identified an error in the computation of one of the hydrological indicators ($VCN30_5$). We corrected the error. The change rate values of the VCN30 are now more consistent with those of the QMNA (both low-flow indicators) than before. We will update the figure 7 and the values of changes rate of the $VCN30_5$ in the results section of the manuscript. We will also update the appendix C.

The study is a new attempt to reveal non-natural records of different origins, including erroneous ones, in streamflow time-series. The authors developed a comprehensive protocol for visual inspection of river flow data and involved 43 experts to detect anomalies in 674 streamflow time series in France using the protocol. The study showed a huge variability in the assessments of experts and confirmed the prevailing a priori ideas about the predominance of subjective factors when deciding on the presence of anomalies. Nevertheless, even with such uncertain results, the authors were able to formulate several recommendations, among which two seem to me to be the most important: (1) analyze as few types of anomalies as possible; and (2) allow experts to supplement the detected anomalies with confidence estimates.

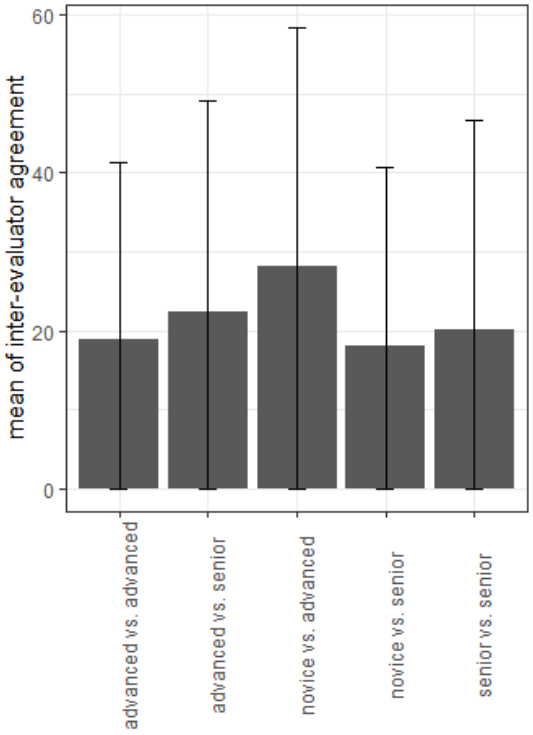
Overall, I believe that the manuscript addresses relevant scientific issues and contains results that could make a useful contribution to future studies. The scientific methods and assumptions are valid and clearly outlined. The presentation is well structured and clear. I find the study to be interesting and recommend the manuscript for publication after minor revisions.

> Thank you for you encouraging words about the manuscript.

Compared to Martin Gauch's excellent review already published, there is very little I could add. I fully agree with the major comments 2, 3, and 5 of this review; namely, following these comments, I also recommend the authors: to compare the obtained "change rates" with the values that would have been obtained by randomly deleting the same amount of data from the analyzed series; to evaluate the inter-evaluator agreement within certain categories of experts; and to assess whether the quality of hydrological simulations will change when evaluating the performance criterion on the cleaned series.

> As many of the comments match those of Martin Gauch, we refer to our comments about the impact of random sampling of anomalies on hydrological change rates. (Answers to RC1 Martin Gauch's comments 2, 3, and 5)
>
> Regarding the inter-evaluator, we feel that your comment goes a little further and aim at assessing if evaluators agree more with other evaluators of the same level of expertise (as times series were analyzed by two evaluators with potentially different levels of expertise). The figure below illustrates the mean (+/- standard deviation) inter-evaluator agreement for each combination of level of expertise. There is no evidence for a better combination of level of expertise that maximize the inter-evaluator agreement, even for the combination of 2 senior hydrologists (figure below).

During the experiment, we avoided the combination of 2 novice evaluators for a station, this is why “novice vs. novice” is missing from the graph.



In addition to the technical comments below, I would like to make two more general notes, and I’ll be grateful if the authors comment on these issues in their response.

The first one concerns to the organization of the related studies. It seems logical to me to make one preparation. Before the main study begins, ask experts to weigh in on one or a few (but not many) reference streamflow time-series where some of the data has been substituted with fictitious data that the organizers are aware of. This stage will provide a preliminary general sense of the potential levels of expert agreement and the accuracy of their expert judgments.

We totally agree, the fictitious data you suggest to add to time series could be part of the inter-calibration of the evaluators phase that we suggested in the manuscript. Since our initial aim was to clean a large dataset of streamflow time series, the study of the subjectivity of the individuals and of the distribution of the anomalies came afterward. We can mention this as a recommendation in the discussion section (L344).

“A phase of inter-calibration of evaluators, and even better with the data producer when possible, is highly recommended as it could reduce the subjectivity of such an exercise. This calibration phase could be completed by assessing the ability of the evaluator to detect fictitious anomalies in streamflow time series.”

The second general comment relates to my personal view on the perspective of visual detection of anomalies in the streamflow time-series. Given the inevitable high level of subjectivity in expert judgments (associated, first of all, with the experts’ experience), I believe that expert assessments

would become more effective if not the entire series of observations were subjected to visual analysis but only its suspicious parts, previously identified using popular quantitative algorithms (k-nearest neighbors, clustering based algorithms, machine learning algorithms, etc.). This will make it possible to reduce subjectivity and increase the information content of expert analysis.

> An algorithm that identify suspicious periods seems a more achievable goal than to precisely identify time steps with anomalies, though the risk of removing data of interest remain relevant. We propose to mention that in the manuscript (L344).
>
> *"An automatic detection of anomalies could avoid these issues of subjectivity and weariness. As a first step, an automatic detection could identify suspicious parts of streamflow time series that would afterwards be the subject of a visual inspection, instead of inspecting the whole time series."*

Technical comments

Line 90: "available length of the time series greater than 25 years..." as it follows from line 96

> We will rephrase for more consistency, thanks.
>
> *"(3) available length of the time series greater than 26 years at a daily time step between 1976 and 2019"*

Line 138: It is unclear to me what the reason was to limit an evaluation time. It seems to me that it is more important to get a thoughtful assessment than a quick response.

> Evaluators were free to take all the time needed to inspect the time series. We provided this duration for information. We will clarify that in the manuscript to avoid any confusion.
>
> *"We estimated the time needed to evaluate one station to be approximately 10--15 min per evaluator, although we haven't set a time limit."*

Line 167: "...are the duration of anomaly considering the intersection and the union..., respectively."

> Nice catch! We will correct this sentence, thanks.

Fig. 3b: It is not entirely clear how the inter-evaluator agreement between an expert who analyzed data from 111 stations and another expert who processed data from a much smaller number of stations (say, 10) was established. Please clarify

> A station was always analyzed by 2 evaluators, therefor it was always possible to compute their agreement rates. Since, each evaluator analyzed from 5 to 111 stations, resulting in 5-111 agreement rates (one by station analyzed) we can draw their distribution (as displayed by the boxplots in figure 3b).

I suggest including the main recommendations formulated in subsection 5.3 and related to visual inspection of streamflow time series into the conclusions.

> We will write a short paragraph about the lessons learned from the visual inspection in the conclusion section.

“This study also provided recommendations for future campaigns of visual inspection of time series. We strongly suggest setting up a phase of inter-calibration of evaluators in order to assess their subjectivity, as well as adding a confidence rate to the reported anomalies in order to identify more doubtful periods. Ideally, the development of automatic detection of anomalies, or at least doubtful periods, could greatly improve data cleaning stage.“