Referee Comment of Alexander Gelfan on Strohmenger et al. “On the visual detection of non-natural records in streamflow time series: challenges and impacts”

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.

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.

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.

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.

Technical comments

Line 90: “available length of the time series greater than 25 years…” as it follows from line 96

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.

Line 167: “…are the duration of anomaly considering the intersection and the union…, respectively.”

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

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