Thank you very much Dr Gauch your review. We provide below answers to the reviewer's comments. 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 $VCN30_5$ 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.

Summary

The paper describes the outcomes of an effort to have human annotators detect non-natural/incorrect streamflow measurements. The authors find remarkable inconsistency in different annotators' responses.

Overall, I think this paper is well-written and to the point. I appreciated the discussion and suggestions for future related studies.

I think that a few changes in the analysis of results could further improve the paper. Please find my detailed comments below.

Thank you for your kind words about the article, and for all your constructive suggestions.

Major Comments

1. I would guess that it is not always clear which exact timesteps around an anomaly should be annotated. I think it might therefore be interesting to see if the consistency increases noticeably when we consider a windowed approach: e.g., consider two annotations to agree when they fall within a 1-week window.

Indeed, the inter-evaluators agreement might increase by including more flexibility in the start and end dates reported by the evaluators. We tried windows of plus and minus 3 and 7 days for allowing the intersection of anomalies, and then we computed the inter-evaluator agreement. The overall median of agreement rate increased to 15 % (+/- 3 days) and 17.4 % (+/- 7 days) vs. 13 % in the manuscript (+/- 0 days). Although the inter-evaluator agreement increases with these extended windows, it remains surprisingly low. We propose to add a sentence about this experiment in the discussion of the results in section 4.1:

"This is true for the identification of the timing of anomalies, although including a margin of three or seven days on the start and end date slightly increased the inter-evaluator agreement to 15 and 17 %, respectively. The subjectivity of evaluators also reflects in the attribution of a type of anomaly as shown by the high proportions of "other" and disagreement types of anomaly identified (Figure 5, a), and by the low percentage of self-association of type of anomaly (Figure 4)."

That being said, choosing the adequate length of the windows around the time steps is tricky because of the duration of the different types of anomaly (linear interpolation and noise are longer than point anomaly, several weeks vs. few days).

2. I appreciate the effort to measure the impact of removing annotations through the "change rate", but it is hard to put the values into perspective. For that, I would propose the following baseline: report the change rate that you get when randomly removing the same fraction of time steps (ideally, do this a few times and report the average change rate).

We agree that such a baseline could help interpreting the changes rates of the hydrological indicators. We computed them, following your suggestions, for each station:

    a. Count the number of days reported as anomalies (n) in a time series

    b. Attribute anomaly to n random time steps for which discharge was observed

    c. Compute the hydrological indicators for this randomly altered time series

    d. Repeat this 100 times by time series and average the change rates

We plotted the results using the same layout as in the manuscript (figure 1 below, which complements figure 7 of the manuscript). Results show that randomly removing time steps from streamflow time series (see the baseline boxplots) has negligible impacts on hydrological indicators, much lower than those observed when removing actual anomalies reported by the evaluators.
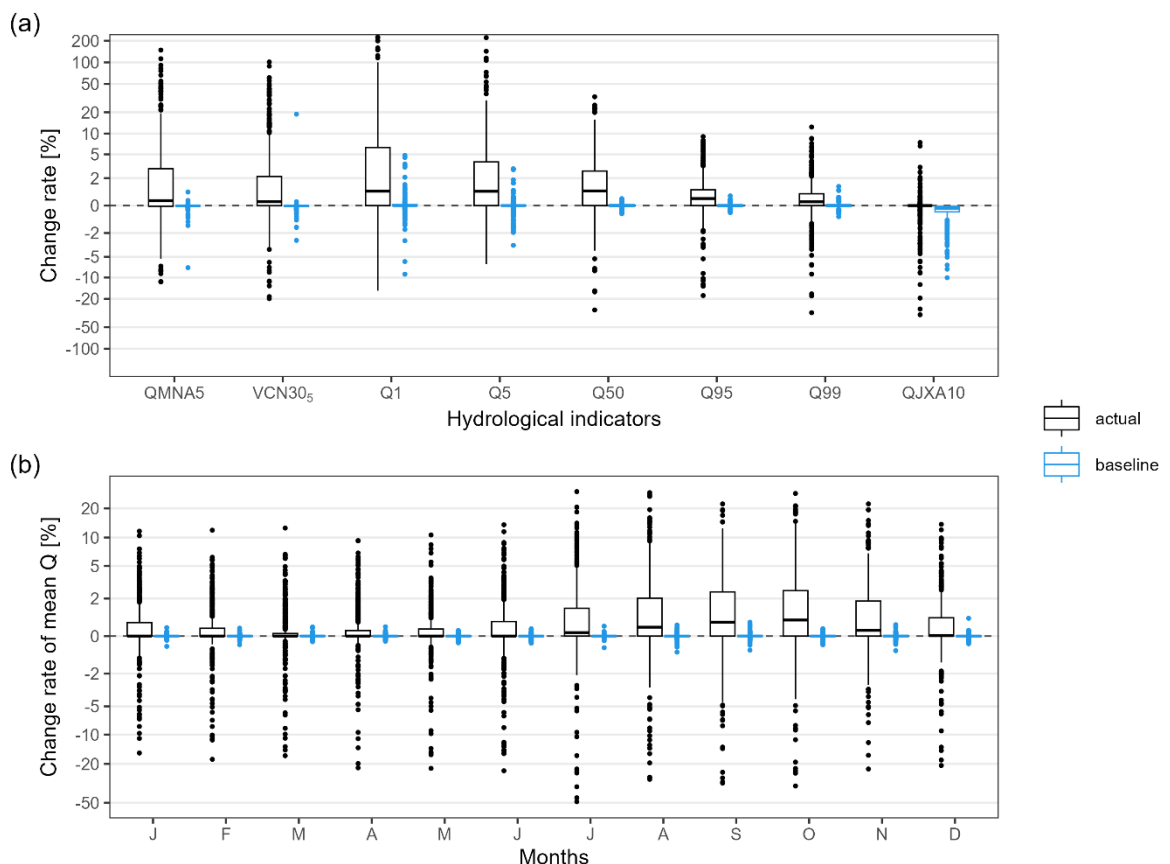

Figure 1: Change rates of hydrological indicators after cleaning the anomalies from the streamflow time series relative to indicators calculated with the initial streamflow time series. (a) is the distribution of the change rates for quantiles (1, 5, 50, 95, and 99th), low-flow indicators (QMNA5 and VCN305), and a high-flow indicator (QJXA10), and (b) is the distribution of the change rates of monthly mean streamflows. Colors refer to the type of experiment:

actual (black) relates to changes rates when removing actual anomalies reported by evaluators, baseline (blue) refers to randomly removed time step from the time series. Statistics are displayed by boxplots where the lower and upper hinges correspond to the 25th and 75th quantiles, the vertical bar is the median, the upper and lower whiskers extend up to 1.5 times the interquartile distance from the 25th and 75th quantiles, and the dots are the outliers beyond the end of the whiskers.

We propose to mention this in the manuscript in section 4.2.

"Indeed, the change rates were higher for hydrological indicators from July to November and for Q1, Q5, and QMNA5 (Figure 7), which could have implications from an operational point of view. We assessed the impact of removing random time steps from streamflow time series in the same proportion than reported by evaluators (results not shown). We observed that the change rates for these indicators when removing anomalies were larger when using evaluators' feedbacks than when removing random time steps."

3. It might be interesting to see if there is a noticeable difference in the responses/consistency of novice vs. advanced vs. senior annotators.

This is an interesting suggestion. We checked the relationship between the subjective statistics (percentages of anomaly reported and inter-evaluator agreement) with the level of experience of the evaluators. The percentage of anomaly reported seems to decrease as the experience of the evaluator increases; however, the variability between individuals is too high to conclude for a strong relationship (Figure 2). One interpretation could be that senior evaluators tend to neglect anomalies they believe will not affect a hydrological analysis, while novice evaluators tend to be more exhaustive and try not to miss any potential anomaly.

For this experiment, we refined the level of experience of the evaluators in our database. Thus, the proportion of each level of experience of evaluators slightly changed from 14 to 24% for novice, and 46 to 38 %, for advanced and senior (see L104). This will be adjusted and additional analysis will be mentioned in section 4.1.
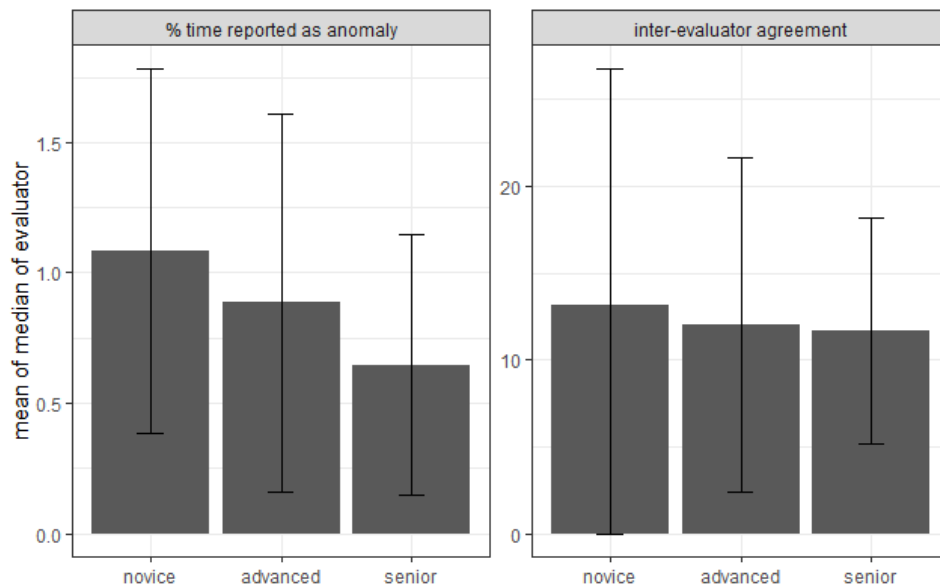
Figure 2: Percentage of time reported as anomaly (left) and inter-evaluator agreement (right) with respect to the level of experience of the evaluators. The height of the column and the error bar indicate the mean and the standard deviation within the groups, respectively.

4. Reproducibility/data availability: For purposes of reproducibility, I would appreciate the download to include:
    a. the streamflow data (and ideally also the supplemental information like precipitation, temperature, GR5J simulation, etc.), such that others can reproduce and build upon the analyses. The French-only HydroPortail link is not helpful and it's unclear what data one would need to collect from there.
    b. the code used in the study (e.g., to calculate the various statistics)
    c. information on the evaluators (background academic/operational, level of experience)

Unfortunately, we are not allowed to provide for download data that we did not produce and that we do not own.

Regarding hydrological data, we would like to mention the existence of the hub'eau API to collect streamflow time series over France from the HydroPortail. It is a French service that aim at simplifying access to water data:

https://hubeau.eaufrance.fr/page/api-hydrometrie

A tutorial is available here (unfortunately also in French, but translation tools seem to be efficient in providing an English version):

https://hubeau.eaufrance.fr/page/api-qualite-cours-deau-tuto

We will mention the hub'eau API in the data availability section.

Regarding the SAFRAN meteorological data we used, we are also not allowed to share it. However, the data producer, Météo-France, generally provides these data for research purposes. In the present case, to run the GR hydrological models the use of SAFRAN is not

compulsory: any open access European dataset (e.g. ERA-Interim) could have been used, as the performance of the model is not the core of this work and we expect larger datasets still to have consistent precipitation signals for the evaluation phase. We only used SAFRAN because this is the best dataset at our disposal and usually use it in most of our studies over France.

Based on the fact that we are not allowed to provide neither hydrological not meteorological data, we prefer not to provide GR5J simulations. However, the GR5J model is available in an open access mode in airGR (https://cran.r-project.org/web/packages/airGR/index.html) and could be used by anyone to reproduce the experiment.

We provided additional information on the background and the level of the evaluators in a new file (evaluators_profile_Explore2) available on the data deposit: https://entrepot.recherche.data.gouv.fr/dataset.xhtml?persistentId=doi:10.57745/SO2WOV

Thanks for the suggestion; we apologize for not being able to positively answer to the entirety of this demand.

5. It would be really interesting to see how removing bad data impacts the quality of hydrological models. Is a model that is calibrated/trained only on "good" timesteps better than one that is calibrated on all timesteps (on a validation period)? This might be out of scope for this paper, but could at least be mentioned as future work.

Absolutely! Actually, we are currently working on the impact of anomalies on models parametrization and simulations. However, in our opinion, this is a large topic that would merit a whole article. We will add a sentence to discuss this perspective in the discussion section (see also comment on L328ff below)

Minor Comments

L90 states >30 years of available records as a condition, but then L96 reports that some rivers have only 26 years of streamflow. Am I missing something?

L90 states "30 years with less than 10 % of missing data" which should be >= 27 years. You are right; one of the station is actually shorter (26 years) due to a human mistake during the station selection stage.

We will rephrase, as also suggested by RC3, for:

"(3) available length of the time series _greater than 26 years_ at a daily time step between 1976 and 2019"

L224 "little impact on the change rate". If I understand correctly, the change rate is exactly what measures the magnitude of impact. I.e., the change rate only exists after removing the anomalies (and not before). So shouldn't this rather be reworded to something like "did not result in large change rates"?

Change rates were computed as the difference between hydrologic indicators with and without anomalies included in streamflow time series. You are right, "removing anomalies" do

not affect change rates, but rather "results in" change rate modification. We can rephrase the sentence as follows:

"Removing the anomalies had little or no impact on the high-flow indicator values, while those of low-flow indicators could slightly increase."

L242f Regarding the first reason you provide for the stabilized variability: I don't think it is the variability that lowers as the number of stations increases. I think it is the quality of the estimate that improves as more samples are available. Both station quality and human randomness have some variance, which will be better estimated from larger sample sizes.

You've said it better than we could. We propose to rephrase as follows:

"This variability seems to stabilize for the evaluators who analyzed more than 40 stations, which raises two questions: Is it related to the variability of the quality of the station time series that is better estimated over 40 stations analyzed? Did the evaluators gain experience in visual inspection during the operation, so that their expertise converged toward a value similar to that of other evaluators?"

L266: This notion of annotators getting "bored" could be verified by checking whether the number of annotations also decreases across stations (those rated first vs. those rated last)

True, however, we cannot retrieve the information on when evaluators examined the time series, thus such analyze cannot be performed.

L281: Do you know how this "good-quality" label was assigned? Does this mean another human had visually deemed these records as good? Or via automatic checks?

Quality label was assigned by data producers and describes the quality of measurements during low-, mid- and high-flow periods. Each data producers has its own criteria to validate the data, based on daily, weekly or monthly expertise, statistical analyses, and hydrometric monitoring software. A good quality label means that the data producer removed the known period of the streamflow measurement disturbance such as for example sensors maintenance operations or sensor drift.

L290: The logarithmic scale is an important detail to me, because I think it makes linear interpolation much harder to notice. This might be something to add to L349f.

We feel that streamflow time series variability is so high (at short time scale) that any linear interpolation longer than a few days is easy to notice, even in logarithmic scale. However, we strongly encourage to always display time series using a linear scales at least, and we provided observed time series both using a log and a linear scale.

L328ff: I would move this brief literature review to the related work section and remove it from the discussion.

We provided this literature review to introduce model's sensitivity to potential anomalies. As you mentioned in a previous comment (comment 5), we feel that this is out of scope for this paper, and then we would like to keep theses sentences in the discussion section. However, we did not clearly state it in the manuscript. We will add a sentence to mention this perspective in the discussion section.

"Hydrological models may also be influenced by anomalies, and further studies should investigate how parameters and simulated streamflow change when removing anomalies from the training period data. "

Out of curiosity: would a map of France show any spatial patterns in the annotations?

We did plot the map of anomaly rates, as we expected to see the propagation of anomalies from upstream to downstream stations, but we did not find any spatial pattern or spatial correlation of the anomaly.

Typos
L125 rainfall—runoff

This will be corrected, thanks.