We would like to thank the reviewer for the feedback and the suggestions to improve the manuscript. Here, we respond to each comment (in bold).

The manuscript presents an interesting idea to distinguish different baseflow components. To my understanding, the main methodological assumptions are correct and could potentially make an interesting contribution for the journal. However, in my opinion, the draft is not well structured and written making it difficult to read, uncertainties regarding the selected dataset and the separation of regimes need to be addressed and the discussion and conclusion should be revised accordingly prior to publication. Below there are some suggestions that might help to improve the manuscript:
We appreciate the numerous and very helpful comments from Reviewer 2 and will revise the manuscript according to our response below.

Major comments:
- The overall readability should be improved by favoring short and straight forward formulation: The use of a multitude of abbreviations and the inconsistent usage of wording (e.g. with regard to the term storage) make the paper tough to read: e.g. the introduction needs to be re-written, in my opinion it lacks structure and conciseness; there are many incomprehensive formulations and inconsistencies e.g. the sentence in line 30 to 31 does not make sense to me, are you talking about the magnitude of "sustained streamflow" or more generally of the existence of streamflow? "sustained streamflow and hence freshwater availability" – most freshwater is stored in aquifers; And why does streamflow need to be estimated from BFI? This first general introduction is just very confusing. line 34: What are stored sources? clearly, discharge is coming from "stored sources" whenever it is not raining, the BFI is often interpreted as the contribution from groundwater . . . as you state in line 38. Line 39: you write about water from groundwater, soil and "other delayed source" Which other sources do you mean? Please mention them! There are multiple more examples in the following lines, please try to be more concise in your wording and restructure the introduction!
Thanks for this comment. We will revise the introduction with explicit focus on consistent use of terms (e.g. storage, sources, delay).

- The way you report the selection of catchments is critical: you state that human influence on these "headwater" catchments is negligible; (line 254-255): the term headwater catchments is a little misleading for basins of up to 955km2; most of the area in Germany and Switzerland is densely populated, thus human influence might matter: especially when overall magnitudes are small e.g. distinguishing between long delay and baseline delay you will need to make clear that we are not looking at human influence or potential feedbacks from evapotranspiration and vegetation during extended dry periods. The MAG is a regulated basin with huge dams for hydropower and thus highly damped discharge, which makes me a little suspicious if the other selected catchments are suitable for the analysis; please remove MAG and consider double-checking your catchment selection!
We will remove the term "headwater" and "often negligible" and will describe the catchment and regime characteristics (e.g. also human influences) with more details. We will carefully check all catchments for potential human influences in order to discuss the effect of human influences on the outcome of the analysis.

- the reasoning for classifying the different regimes, especially when distinguishing between RLWR and RUPR, needs to be further discussed: looking at figures 5 and 6 one could argue that the variation within the groups RLWR and RUPR is larger than the difference between their medians, so from a process point of view (in the end that's what you want to capture) the separation based on mean and max elevation might not be suitable. In Figure 7 you even argue that different elevation classes might be more representative.

We will investigate the variation within the two rainfall-dominated groups in order to check the reliability of the suggested catchment classification. Our catchment classification has been hypothesis-based, i.e. catchment elevation is a metric to distinguish important drivers of different delayed contributions. We will discuss the value of this approach (see also additional figures below).

HYBR represents a mixture of snow and rainfall dominated catchments, but obviously as suggest by your results, it is not, can you discuss why?

We will add more discussion about the specific streamflow response patterns in the HYBR catchments. We will use more information about recession characteristics (as suggested below) to analyze the role of catchment storage in HYBR catchments.

There might not be an easy solution to these issues, but maybe they can be discussed more detailed. The snow-dominated catchments are significantly smaller than all other catchments (Table 2), please mention that explicitly and update your interpretation accordingly (e.g. line 294 "higher flashiness during summer flows" might be an artefact of catchment size);

We will highlight that SNOW catchments are in particular smaller than the other study catchments and will revise all statements regarding the flashiness of the catchments.

maybe you can provide some basic streamflow statistics of the dataset e.g. in Table 1 potentially add magnitude and variation of q5, q50 or recession characteristics with respect to the selected catchment grouping.

Yes, we will extend Table 1 to present more flow and recession characteristics.

- also the discussion would benefit from restructuring and improved consistency: e.g. line 409 where can I see "a shift in catchment response" at around 2000m?

No, we do not see this here, rather with the "2000m" (line 410) we are referring to another study (Pellet and Hauck, 2017). We will make this statement clearer in a revised version. Our reference on Fig. 6 is also wrong in this section (should be Fig. 7). We will rewrite the sentences accordingly.

line 420: how would you apply the framework worldwide? your case study is on data carrying a strong seasonal signal and elevation gradient. In my opinion the called paradigm shift appears a little too ambitious, as there are (as you also point out in the introduction) several approaches to capture delayed contributions from different storage settings. I don't see how the proposed approach assess (line 439) "different type and number of storages, hence various delayed contributions" While I agree that BFI does not account for single catchment features, also DFI will not identify them specifically (line 445), but you rather get a signal of delayed outflow from potentially multiple (different) sources. Potentially the climate regime itself might significantly influence Nmax, dry periods in southern Europe or norther latitudes, high-elevation catchments streamflow droughts occur on timescales of < 60 days (up to 4 months). Whereas it might not be relevant for large parts of your study region, it might lead to a biased view on snow-dominated systems and potentially when applying the proposed method elsewhere.

Thanks for this detailed comment. Review 1 has also raised the question about the transferability of the method to other regions. We have discussed the influence of N_{max} and the number of breakpoints (e.g. line 400, line 396-398, line 475-481). In the revised manuscript we will unite these discussion points and will also highlight that neither BFI nor DFI are able to identify contributing sources in terms of process understanding. DFI instead gives an estimate of the composition of different delayed contributions. Further, the transferability of the method to other regions will be discussed more detailed.

Also 5.3 starts with a confusing argumentation (lines 483, 484): recharge is crucial everywhere, fair enough, in Alpine catchment seasonal snowpack supplies summer streamflow, however according to table 1 low flow / delayed flow occurs Jan to March, also (line 485) saturated soils are not allowing groundwater recharge. The influence of global warming on melt processes and groundwater recharge is highly depended on the elevation range you are referring to (line 486). To my knowledge it is not yet clear if smaller DB (or smaller groundwater contributions in general) can be directly related to the size of subsurface storages (line 493). There is ongoing discussion if differences in magnitudes are related to variable connectivity of storage and stream, variable precipitation / evapotranspiration in different elevation / exposition or differences in storage recharge. Line 513: If DB is the groundwater contribution, why would less developed soils matter? Again, the ranges you report a quite large, however the SNOW catchments are significantly smaller. The whole argument on storage in SNOW catchments is complicated to follow, you start the argument with Alpine storages are small (but you don't mention who reports that), afterwards you mention numerous studies that report the opposite, to conclude that "high-elevation catchments have larger catchment storage than previously thought".

Reviewer 1 has comparable concerns about the role of snow, soils and groundwater in alpine catchments. Indeed, the line of argumentation in section 5.3 could be improved. We will discuss potential reasons for large(er) D_B-contributions in SNOW catchments and will compare our results with other studies analyzing (dynamic) storage in alpine catchments.

Some final thoughts on 6: Low streamflow occurrence might be highly variable comparing different years, mainly depending on climate, I'd suggest mentioning that explicitly and re-formulate less definite. Also, the high accordance to elevation gradients might be specific for the Alps, you might not find that in other regions e.g. Scandinavia, southern US;

We will discuss the role of low flow occurrence during the year, looking at the timing in particular (summer or winter low flow regimes). Indeed, the occurrence is variable comparing different years. The variation is higher for rainfall-dominated than for snowmelt-dominated catchments. We will make clear that the analysis in section 6 is based on a set of generic low flow month for different regime types (i.e. summer low

flow, winter low flows) and that in other regions or climates (outside the Alps) other months should be chosen to evaluate the low flow stability.

Minor comments:

in Figures 1, 3 & 9 the difference between light blue and blue (long vs. baseline) is not visible (in Figure 7 you even replace blue by black, which makes it much more readable, maybe change it also for Figures 1 and 3).
Will be revised with focus on consistent use of colors.

the usage of hyphens is quite arbitrary throughout the document, to my knowledge there are clear rules, please check them and change accordingly e.g. line 26 low flow stability index. . . low flow regimes, line 30 groundwater-surface-water-interactions, line 318 5-days… In Figure 1 the dark blue color refers to baseline delay class although it is obviously (the volume) below the baseline, 1b is too small
Will be revised.

In line 169: What is the “seasonal low flow period”? How long is it? Where can I see that period of 60 days in the hydrographs of Figure 1a? What exactly is AM, MAM and MQ and how do you calculate them?
We will explain “low flow period” and the index MAM/MQ with more details. The indices AM, MAM and MQ are explained in line 170-173. However, as details on calculation are missing, we will add more explanation here.

Line 387: assessed, and may; line 391: sustain low flow for sustained dry periods; line 523: winterly recession;
Will be changed.

Additional material we are considering to include in the revision:

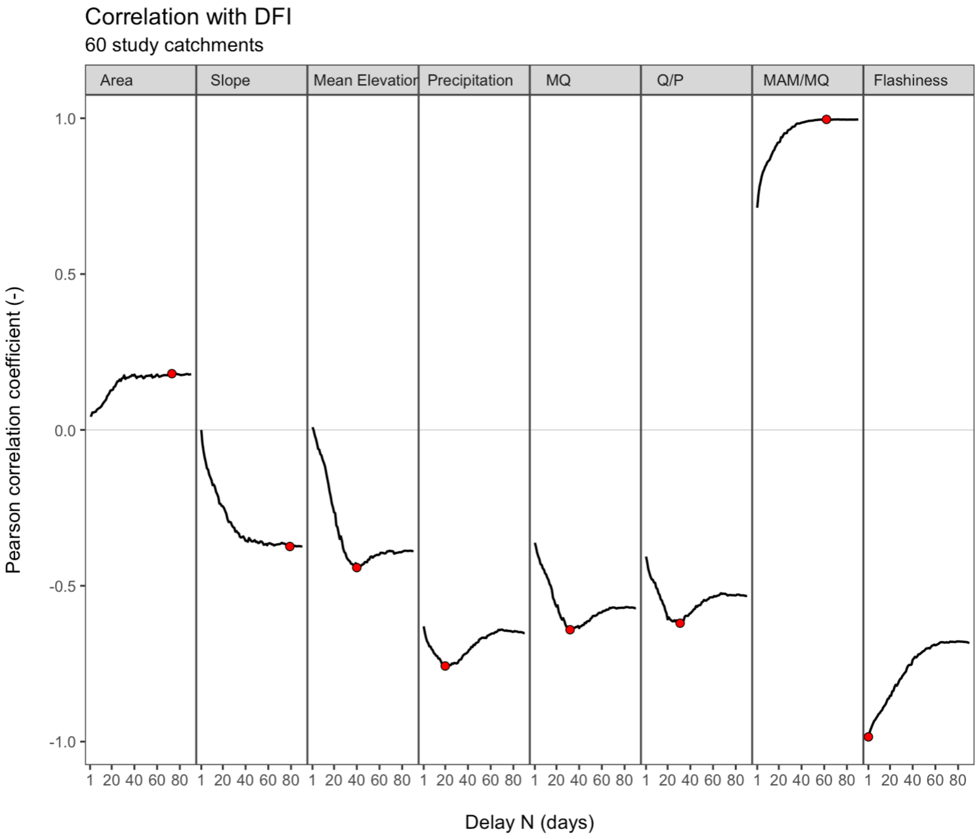

Figure 1: (suggestion for additional figure): Correlation strength between the DFI and various flow and catchment characteristics. DFI is calculated for block size N = 1 – 90, the red dot indicates the highest absolute correlation coefficient. With this figure we can argue that a low flow sensitivity measure like MAM/MQ gives better correlation over 60 study catchments than the other 7 variables (e.g. area, slope, elevation etc.). Differences between independent characteristics and flow-derived characteristics should be discussed.

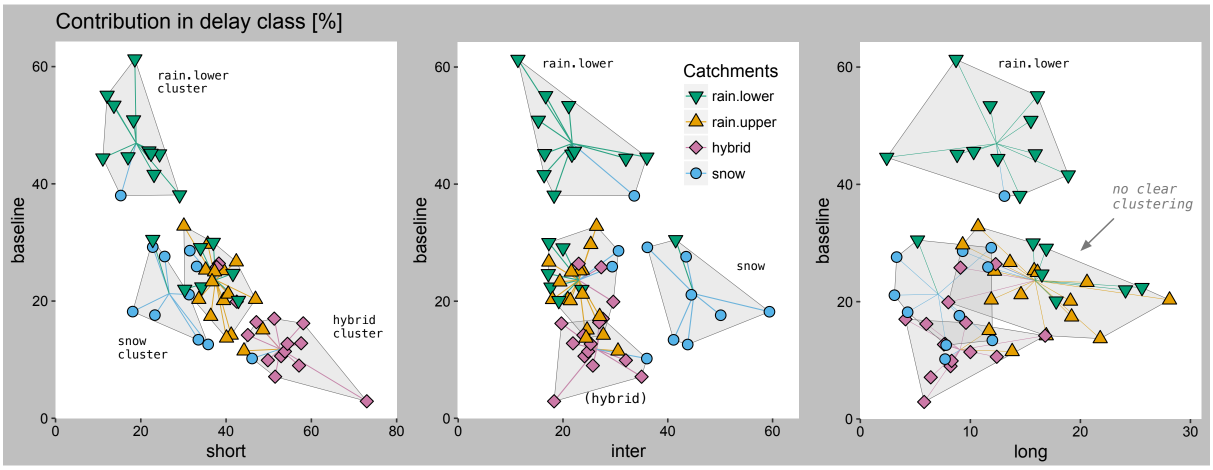

Figure 2: A k-means clustering with the 4 relative contributions (short, inter, long, baseline) for all catchments. Hypothesis was that we should find homogenous clusters regarding our classification approach (i.e. homogeneous green, orange, magenta, blue catchment dots in each cluster). Outcomes of cluster analysis should be discussed.