Dear Editor,

We have considered all the comments in detail and we have extensively modified the manuscript accordingly. The major changes introduced in this version of the manuscript are:

- We have emphasized in the introduction and especially in the discussion and conclusion sections how data specification and classification procedure affect the classifications.

- According to many referees' comments, we have realized that some parts of the text, especially regarding the methods section, were difficult to understand. Therefore, the argumentation of our results could result unclear. In this sense, we have deeply modified the methods section to clarify how the classifications were developed and compared.

- We modified extensively all the sections regarding the "analysis of distinctive gauges". We have include a brief explanation about the effect that these gauges are expected to produce in the classifications, we have modified the analysis regarding the class heterogeneity and finally, we have changed the argumentation of our results according to the referee's comments and advices.

- We eliminated from the results section those data that were already included and represented in tables and figures.

- Figures 5 and 6 were changed to increase their readability.

- The spatial correspondence between classifications was re-analysed using a larger number of segments and repeating the analysis 10 times to increase the representativeness of the results.

- A language revision has been done thorough the text and many sentences have been rewritten to clarify its actual meaning.

Following, we include a point-by-point response to the referee's comments and the specific changes included in the manuscript:

**Responses: to Referee N#1 comments**

1- The paper lacks a discussion of the consequences of the data processing responsible for the different results. E.g. a classification with data recognizing catchment size has to be different of a classification with the same data and considerably less influence of the catchment size (as you get with normalization by average runoff). The meaning of a specification for the data, e.g. elimination of outliers, noise, unexpected runoff behaviour, range of dispersion or loss of information, is an important task for interpreting the results.

Understanding the consequences of data specification is essential! A general recommendation for one specification, regardless the issue of the classification, cannot be made for a not completely understood Specification (PredF). Therefore, please consider the meaning of the specifications in the discussion and conclusion.

In this paper we analysed two types of specification that can affect classification characteristics: 1) The data processing (raw versus normalized flow series) and 2) the classification procedure (ClasF vs PredF). In addition, we analysed how these two specifications affect four classification characteristics: 1) classification performance, 2) hydrological interpretation of classification, 3) ability to deal with underrepresented parts of the hydrological space and 4) The spatial correspondence.

According to Referee's comment we have emphasized in the introduction and especially in the discussion and conclusions how each type of specification can affect the different classification characteristics. We included and changed many sentences in the new version of the manuscript where we addressed specifically the effect of those specifications constrains the classifications outcomes.

Within this comment, Referee said that "a general recommendation cannot be made for a not completely understood specification (PredF). After reanalysed several issues according to the comments of Referee #1 and #2 results showed that PredF presented higher classification performance and a greater ability to estimates the hydrological character of the underrepresented parts of the hydrological space. In addition, after comparing our classification with others that covered part of the study area (Bejarano et al., 2010; Solans and Poff, 2013), we could concluded that PredF generated classes that are more similar to the actual distribution of river types in the study zone than ClasF. Then, we believe that according to our analyses and results we now understand how the PredF strategy works and why it outperformed ClasF. Therefore, we change the sentence in the conclusion.

2- The PredF strategy should be explained more in depth to understand what is done with the data and what the consequences for the data are. It seems that the PredF strategy leads to a loss of information with respect to the variability of data. The resulting data may not cover the whole real data space. The advantage of PredF is the possibility to construct data for underrepresented conditions to obtain classes of equal size. However, to find classes of equal size is not a priority objective of clustering and classification. Please comment on this.

In regard to the loss of information mentioned by the referee, we have modified the manuscript to clarify what we have done with the Synthetic River Network (SRN). The misunderstanding of the procedure was probably due to the position of the paragraph in the original version. We have put it before explaining any of the two strategies. We believe that this change clarifies what has been done with the data. The reduction of the SRN from 667406 to 178296 has been done in both ClasF and PredF. This has been

done because there are certain types of rivers that are not represented in our initial data set due to the absence of gauges in these rivers. Therefore we limit the potential of our predictions to rivers represented in our initial data set which was defined according to the range of the different predictor variables. However, as stated, it has been done for ClasF and PredF, so results extracted from each strategy are comparable.

Referee is right in saying that finding classes of equal size is not a priority objective of classifications. In figure 7 we included a line representing the “theoretical” most even distributed class (i.e. if all the classes of a classification have equal size). This line was only included as a reference benchmark to see the frequency of the classes that incorporated the distinctive gauges. However, it does not indicate that all the classes should have this frequency. We have changed this issue through the text to clarify that this is not an objective of the classifications.

3- Type and necessity of normalisation depends on the type of data and the purpose of analysis or classification. If you compare runoff behaviour, normalization of data is necessary for each comparison of indices depending on catchment size. To compare runoff values of different catchments, normalization can be counterproductive. Therefore the aim of the classification determines normalization or not. For other indices like the timing of extreme flow events or numbers of days with increasing flow a normalization is meaningless. Please comment on this.

Referee was right on this comment and it is something we said in the introduction “normalization can be viewed as a completely subjective choice that depends on the purpose of the classification”. Nonetheless, we have rewritten the paragraph in the introduction to include some of ideas that were pointed out by the referee in order to clarify how normalization can affect classifications.

We also introduced several changes in section 4.2 of the discussion and in the conclusion to state clearly the main implications of normalize flow series for further uses of the classification.

4- Language: Frequently the text is difficult to read and imprecise. Many things remain unclear and should be revised.

We agree with the reviewer in that some of the sentences and paragraphs of the manuscript can be difficult to read. Two of the co-authors of the paper are English native and have made a rigorous language correction of the first version of the manuscript. Even tough, all the manuscript has been thoroughly revised and many sentences and paragraphs have been changed in order to clarify its actual meaning.

5- Why do you compare 19 classifications? Are there no optimal sizes of classification?

We compared 19 classifications because there was not an optimal number of classes that can be defined a priori. So in the paper we explored if according to the classification strength and the ANOVA analysis, the optimal number of classes for each classification procedure could be define. We observed that beyond 6-8 classes, the differences between classifications are not significant, i.e. classifications with different number of classes presented similar statistical performance. Hence, other criteria, different to the statistical performance, should be employed for the definition of the number of classes. We included an extra comment in the conclusion.

6- Why do you use a different number of hydrological indices (101 for the raw data against 103 for the normalized data)? May this affect the result?

When daily flow series were normalized by dividing each mean daily flow by the annual flow regime, l1 became equal to 1 in all the gauges. In addition, lcv became equal to l2 (as lcv = lca/l1). Hence, it makes no sense to include these variables in the analysis. For that reason when series were normalized the number of hydrological indices was reduced from 103 to 101. A brief explanation was now included in the new version of the manuscript.

7- Page 953, line 21: average rock hardness: which rock characteristic is the basis of the calculation and what is the meaning of the hardness to hydrological processes? Please explain.

A sentence explaining the base for the calculation of “rock hardness” and “permeability” has been included in the manuscript. Rock hardness affects significantly river morphology. River morphology interacts with the hydrological regime and influence, in part how the water flows through the reach (for instance, it can affect the duration of a high flow event). So that, we considered that it could be an interesting variable to include in the predictors (together with permeability), as there is little information regarding geology. However, it was less important than expected and in fact was one of the least important variables.

8- Page 957, line 5: acronym OBB unknown, or should this be OOB?

OBB was changed by OOB

9- Page 965: 4.3 Analyses of distinctive gauges belongs to results Fig. 5 and 6: unreadable small figures - perhaps better in another arrangement.

Reviewer was right with this comment. Figures 5 and 6 were included as graphical examples of the results of the ANOVA analysis obtained for all the hydrological indices (the results of each hydrological index were included in the supplementary material). We reduced the number of variables presented in figures 5 and 6 and also enlarge all the

symbols to increase its clarity and readability. We think that the reduction of the number of indices in the figures is still useful to understand the main results of the ANOVA.

**Responses: to Referee N#2 comments**

1- The paper is very difficult to read: part 2, 3 and 4 need a serious overhaul.

We are not sure whether referee's comment refers to the paper structure or the language use. In terms on the structure we have organized it according to the work logic we have followed. We believe that the division of each section in separated parts, which account with the different analysis done to compare classifications, brings clarity to the paper. Regarding the language, two of the co-authors of the paper are English native and they have made a rigorous language correction of the manuscript.

Nonetheless, all the manuscript has been thoroughly revised and several sentences and paragraphs have been changed in order to clarify their actual meaning.

2- Page 952, line 3: Why do you select 103 indices for the raw flow series and 101 for the normalized flow series, or are they different series? Since you are going to apply statistical models, you will start already with a small bias.

When daily flow series were normalized by dividing each mean daily flow by the annual flow regime, l1 became equal to 1 in all the gauges. In addition, lcv became equal to lca (as lcv = lca/l1). Hence, it makes no sense to include these variables in the analysis. For this reason the number of hydrological indices extracted from the normalized series was reduced from 103 to 101. A brief explanation was now included in the new version of the manuscript.

3- Page 952, line 11: Please elaborate a bit more on the used procedure (outlined in Olden and Poff, 2003) to the reduction of the original sets of indices.

We include a sentence in the new version of the manuscript to explain briefly how this has done.

4- Page 953, lines 1 to 3: Please transfer the number of variables (n) to line 8. Since you are going to reduce them they are inappropriate when mentioned here.

Done in the manuscript.

5- Page 953, line 21: What relationship does the average rock hardness of a catchment have with a hydrological regime? Please explain.

A sentence explaining the base for the calculation of "rock hardness" and "permeability" has been included in the manuscript. Rock hardness affects significantly river morphology. River morphology interacts with the hydrological regime and influence, in part how the water flows through the reach (for instance, it can affect the duration of a high flow event). So that, we considered that it could be an interesting variable to include in the predictors (together with permeability), as there is little information regarding

geology. However, it was less important than expected and in fact was one of the least important variables.

6- Page 954, line 8: The use of synthetic indices biases the approach of the paper towards the PredF method. Moreover, since the models do not use any physically meaningful parameters, the model results are dependent on the gauges on which the models are trained. Could the model results have a different outcome in another region?

We think the referee pointed out a very important question in this comment. The classifications analysed in this paper depend upon the gauges used to develop them. For instance, based on our models we can predict class membership or synthetic hydrological indices to ungauged sites. However, given that our models are based on empirical relationships and not on any catchment physical parameter, predictions are restricted to the range of hydrological conditions present in our initial data set. This is the main reason to reduce the SRN from 667406 to 178296 segments.

Therefore, if the same procedures are applied to regions with different hydrological behaviour, classifications outcomes may be different. For instance, the generated classes would present other patterns and characteristics attending to the specific hydrology of the target regions. Nonetheless, the results arose from the comparisons between different classifications procedures do not have to differ between regions. For example, Snelder and Booker (2013) compared different classifications procedures to classify New Zealand Rivers, including the comparison between PredF and ClasF. Results were similar to ours as it was discussed in this paper.

We introduce a comment on the new version on the manuscript to address this issue.

7- Page 954, line 13: For the ClasF the entire Synthetic River Network (SNR) is used and for the PredF only 1/3 of the SRN. Could this cause a bias in performance? Explanation is needed here.

We used the same River Network (reduced to 1/3) in both cases. The misunderstanding of the procedure was probably due to the position of the paragraph in the original version. We have changed and put it before explaining any of the two strategies to clarify what has been done with the data. The reduction of the SRN from 667406 to 178296 has been done in both ClasF and PredF. This has been done because there are certain types of rivers that are not represented in our initial data set due to the absence of gauges in these rivers. Therefore we limit the potential of our predictions to rivers represented in our initial data set which was defined according to the range of the different predictor variables. However, as stated, it has been done for ClasF and PredF, so results extracted from each strategy are comparable.

8- **Section 2.7** needs to be rewritten. The following points should be addressed:

### 8.1- Why is the bias of a distinctive gauge important?

Distinctive gauges (DG) are very important elements within classification issue. They can be considered as rare elements or outliers within the gauge data set but this does not imply that they are outliers in the study region. For example, it is probable that unmodified gauges representing large rivers are scarce in catchments with an intense water management, given that much of these rivers would be modified. Therefore, gauges in this kind of river may be distinctive gauges in the gauge network. On the other hand, the installation of gauges in rivers of first and second is quite strange, at least in our study region. However, first and second order rivers can occupy more than 70% of the whole river network.

Therefore, the way the classification procedure deal with these distinctive gauges is very important. For instance, these distinctive gauges can be grouped to other ones that are completely dissimilar, which may mask the hydrological characteristics of the distinctive gauges within the classes where there are included. On the other hand, distinctive gauges can be included in very exclusive classes that present a low number of gauges but present lower dissimilarity. In both cases, their distinctive hydrologic character may be loss or underrepresented when classes are predicted to the whole river network.

The prediction of the “rare” hydrological characteristics to the whole river network previously to the segregation of classes has emerged as an adequate way to avoid these problems.

We included this explanation in the introduction and method section of the revised manuscript.

### 8.2- Please explain distinct hydrological character: is this a character within a class or is it a charter compared to all gauges? On what is this distinction based?

The distinct hydrological character refers to the gauges that present the most dissimilar values of the synthetic indices within our gauge data set. We include a sentence in the manuscript to clarify this concept.

This character was analysed as a function of the dissimilarity between each pair of gauges included in the data set. Please read the section 8.3 of the present comment where the process followed to select the distinctive gauges is explained in detail. We have also modified several parts of the section 2.7 to clarify how distinctive gauges were selected.

### 8.3- Why do you select four dissimilar gauges? Please explain.

We selected four distinctive gauges attending to the dissimilarity between gauges. As it was stated in the section 2.7 of the manuscript we first calculated, based on the synthetic indices scores, the dissimilarity between each pair of gauges and then, the

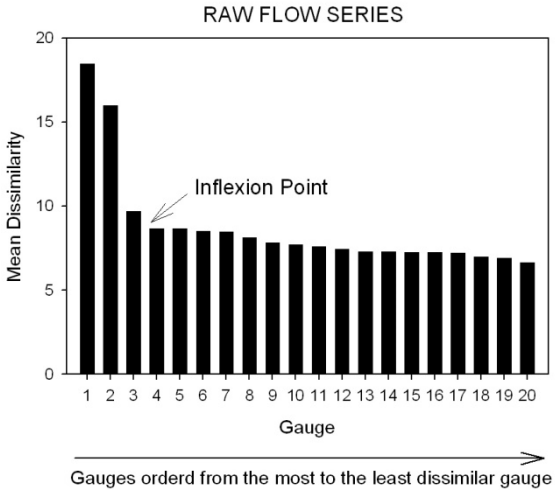
corresponding mean dissimilarity for each gauge with all the others. This analysis was done separately for the synthetic indices extracted from the raw and the normalized series, so the distinctive gauges were different for each type of series as it was explained in the manuscript.

We then ordered the gauges from the most to the less dissimilar gauge and analysed how the dissimilarity values decayed. We select a number of gauges corresponding to the first important inflexion point in this decay trend. We found that this inflexion occurred in the fourth gauge both if dissimilarity was calculated from the raw or the normalized flow series, as you can observed in the Figure 1 included in this comment (we only included the top 20 dissimilar gauges). Therefore, DG1 is the most dissimilar gauge while DG4 is the less dissimilar gauge within the 4 selected DGs. We included this explanation in the manuscript.

In addition, we believe that including four gauges in enough to observe how the different strategies deal with the distinctive gauges. We think that the inclusion of more gauges would have not provided clues about this issue but it could reduce the clarity of the analysis.

We include a sentence in the manuscript to clarify the selection of distinctive gauges process but we think that it is not necessary to include the figure as the paper already includes many figures and tables. Nonetheless, we could do it if referee and editor believe that it is going to be favourable for understanding the methodology.

A)



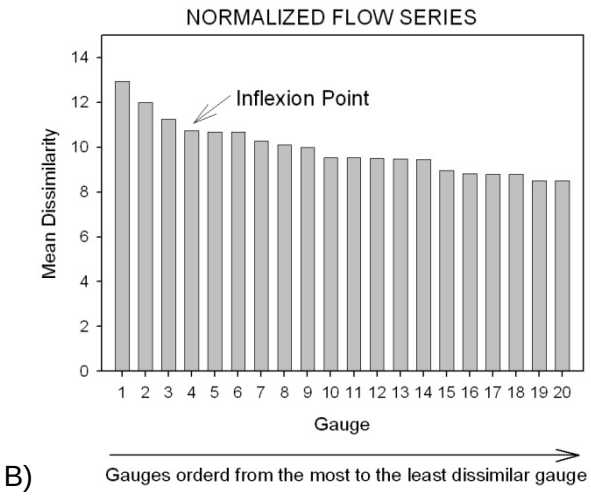



Figure 1. Mean dissimilarity for the top 20 dissimilar gauges based on the synthetic indices calculated form (A) raw flow series and (B) normalized flow series

8.4- Large distances do not necessarily imply a heterogeneous class. This is not true when the distinctive gauge forms an outlier. Please reconsider.

Referee was right in the comment. Large distances do not imply larger class heterogeneity. It actually implies that the DG is more dissimilar to the other gauges in the class than these gauges between them. In the new version of the manuscript we have weighted the distance between the DG and the medoid of its class according to the mean distance between the medoid and all the other gauges belonging to the class. This value indicated how much different is the DG relative to the other gauges included in the class. We have explained this in the text and we have changed table 4 accordingly. We have also changed the results, discussion and the conclusions extracted from this analysis.

8.5. A low frequency of a class does not automatically imply that this class cannot represent certain characteristics of the hydrological space properly. Please reconsider.

Low frequency of a class in the observed space (i.e. in the gauges network) does not imply low frequency in the complete fluvial network. With the last sentence of the section 2.7, we referred that when very low frequency is obtained for a class after classifying the whole network, it is likely that this class is below its actual spatial distribution. For example, the classes that include the DG in the rawClasF classification presented frequencies below 1% at many class levels. Hence, we believe that the hydrologic characteristic represented by this class would be also underrepresented and even lost when frequencies are so low.

We modified this sentence to clarify what we actually analysed.

9- **Section 2.8.** If you select a subset of 500 segments out of an entire set of 667 406 segments, your selection comprises about 0.1% of the entire set. If you select a subset of 500 segments out of an entire set of 178 297 segments, your selection comprises about 0.3% of the entire set. How representative / significant is your subset. Please elaborate on this.

We based this analysis in Snelder and Booker (2013) who selected a subset of 400 segments at random from a river network containing 560 000 segments and obtained similar results. According to referee's comments we have increased the number of segments form 500 to 1000 and we have repeated the analysis 10 times to avoid the effect of the variability in the selected data set. Nonetheless, results varied vey slightly so the same conclusions were extracted from the spatial correspondence analysis.

10- **Section 3.1.** For the raw series you use the first five PCs and for the normalized flows you use the first six PCs as selected according to the broken stick method. However, as stated in section 2.6 you intend to use only the first five PC. Please make the text consistent.

We always selected the 5 hydrological indices with the highest values in the retained PCs. This was done to interpret the hydrological meaning of the classifications. However, the retained number of PCs was defined according to the broken stick method. The number of retained PCs varied depending if the PCA was done on the raw or on the normalized flow series. Thus, if the PCA was applied on the raw flow series the broken stick method indicated that the optimal number of PCs to retain was 5. In contrast, if the PCA was done on the normalized flow series the broken stick method indicated that the optimal number of PCs to retain was 6. We rewrote the manuscript to clarify the method.

11- Page 959, line 10. You state that there are no significant changes from 6-7 to 20 classes. Does this mean that approx. 60% of your classes are redundant? Please comment on this.

This sentence refers to the differences between classifications with different number of classes not between classes in the same classification. This means that a classification comprising to 6 classes present a similar ability to discriminate hydrological indices than a classification comprising 20 classes.

However the sentence was rewritten to clarify the meaning of the text.

12- Please integrate part 3 into part 4 and call it Results and Discussion. Please do not write too much of the contents of your tables in your text. Refer to the tables when you elaborate on the values and use the discussion to explain these results. This will shorten the paper and make it better to read. Please rewrite the discussion part and make it clearer. This part is very difficult to read and in some parts very difficult to understand. A more concise text would enhance the readability of this part.

We agree with the reviewer in that some of the sentences and paragraphs of the manuscript can be cryptic and difficult to read. Therefore, we have revised and changed much of the methods, results and discussion to improve the paper and clarify it. However, regarding the structure of the results and discussion, we still think that the manuscript will be more clear it they are treated as separate parts as it is generally done in many scientific papers. The main objective of the paper the comparison between different classification approaches, including the initial data treatment, the classification procedure and the number of classes. We also compared 4 different classification outcomes. Considering the large number of analysis and its complexity, we truly think that the blending of results and discussion would only increase the complexity and difficult the understanding. Moreover, we have intentionally divided each section in order to facilitate its readability.

13- Page 963, lines 16-20: please move this part to the introduction.

Reviewer was right in this comment. A very similar sentence was already included in the introduction so we delete it from the discussion.

14- Page 965, line 10: Is “predictability” as listed in this sentence, dependent on a gauge or on a method? Please comment on this.

Predictability refers to a specific hydrological index extracted for each flow series (Appendix A). Given that this index was included within the timing attribute it was erased from the manuscript to avoid possible confusions.

*15*- Page 966, line 5-10: These sentences are not clear. Please rephrase.

The sentence was changed.

Page 966, line 16-20: These sentences are not clear. Please rephrase.

The sentence was changed.

16- Page 967, line 11: Please reconsider, you cannot recommend something you do not completely understand.

After reanalysed several issues according to the comments of Referee #1 and #2 we have observed that PredF presented higher classification performance and a greater ability to estimates the hydrological character of the underrepresented parts of the hydrological space. In addition, PredF generated classes that are more similar to the actual distribution of rivers in the study zone after comparing our classification with others that covered part of the study area (Bejarano et al., 2010; Solans and Poff, 2013). Then, we believe that according to our analysis and results we now understand how the PredF strategy works and why it outperformed ClasF. Therefore, we change the sentence in the conclusion.

*17- Page 958, line 5, line 10 and line 22: should OBB not be OOB?*

OBB was changed by OOB