Reviewer 2 : Florian Ulrich Jehn

General evaluation:
This paper proposes a new method to cluster catchments based on the temporal information in their hydrological regime and uses the found regime clusters to evaluate how climate change will change the regimes clusters. I think this is an interesting approach and yields good results, especially the changing of the regimes due to climate change. In addition, the paper is overall well written and has a good flow to it. I think it can be published after minor revisions. However, I have one larger points where I think clarification is necessary.

Main point:
- Line 98: Why those five spline basis functions? How can you be sure that those are enough to represent diverse regimes? It there a connection between using five spline basis function and finding five streamflow regimes? I think this part should be extended to make it clearer what those decisions where based upon.
Reply: *Thank you for expressing your concern regarding a suitable choice of spline basis functions. We chose five spline basis functions because our tests showed that a further increase in the number of spline bases did not further improve the clustering results. This is confirmed by the overall silhouette width, which is for more spline basis functions (6 to 10) lower or very similar to the one for five basis functions (at five clusters). There is no relation between the number of spline bases and the number of regimes chosen. A choice of four or five clusters would also be optimal if more than five spline bases were used.*
We added the following specification to the text: 'The suitability of five spline basis functions is confirmed by the overall silhouette width, which is for more spline basis functions (6 to 10) lower or very similar to the one for five basis functions.'

Minor Points:
- Line 47: "The use of catchment characteristics can be problematic because there is often no clear link between these characteristics and streamflow indices (Ali et al., 2012; Addor et al., 2018)." I think this is worded a bit too strict. For example, Addor et al. 2018 indeed showed that there are differences between the link of catchment characteristics and streamflow indices, but they also showed this connection can be relatively strong for some catchment attributes. Overall, I think this section should be less dismissive of the findings of the cited papers.
Reply: *We did by no means intend to be dismissive of findings of other papers as we appreciate their work. Furthermore, we do not think that it is a bad thing to find weak relationships between certain streamflow and catchment characteristics. We rephrased the sentence to: 'The use of catchment characteristics is not always beneficial as certain streamflow indices do not show clear links to these characteristics* [*Ali et al.*, 2012; *Addor et al.*, 2018]*'. Furthermore, based on the point raised by the other reviewer, we stress that the relation between catchment characteristics and streamflow might be more apparent for event-based signatures.*

- Figure 2 and its discussion: As you are already citing my paper, I hope it is appropriate to mention that that the final version is now published in HESS (https://www.hydrolearth-syst-sci.net/24/1081/2020/hess-24-1081-2020.html) and discusses the different flow regimes in CAMELS in more depth than before (Figure 6). I think the results are very similar to the ones in this paper, but also show that the flow regimes found here can be split in more distinct groups. However, this might be more of a question of the desired granularity.
Reply: *Thank you for pointing out the necessity to update the reference. We point out the*

similarity between the clusters resulting from your and our analysis by saying: 'The five regime clusters identified also show spatial similarities with the ten catchment clusters formed by (Jehn2019) for the same set of catchments using a small set of hydrological streamflow characteristics.' (l.225-226). For some applications it may indeed be useful to have more distinctive clusters.

- Line 85: How is satisfactory model performance defined?

Reply: *Melsen et al.* (2018) *who produced the simulated streamflow data used in this study provide simulations only for the subset of 605 catchments. They chose to run simulations for this subset only as different data sources disagreed on catchment size for the remaining catchments. A reliable estimate of catchment size is crucial to ensure accurate forcing input for the lumped model. We reformulated the sentence as follows: 'In contrast, the regime change analysis uses streamflow simulated by the hydrological Variable Infiltration Capacity (VIC) model for a subset of 605 catchments, for which reliable data on catchment area was available at the time the simulations were produced* [*Melsen et al.*, 2018]*. Kling-Gupta efficiencies obtained over these basins with VIC varied from a first quartile of 0.47, a median of 0.6 and a third quartile of 0.71, with the lowest values obtained in the Great Plains'. (l.145, p.6)*

- Line 90 and following, code availability: I did not see any link to a code repository for this paper (my apologies if I missed it). I think in a paper that does propose a new method, it is important to provide the code used. While the method section explains the idea well, it is still a non-trivial task to recreate the method of this paper without any code examples to work with.

Reply: *It is correct that we have not yet provided a link to the repository with the catchment clusters. This is because we wanted to wait for the DOI of this manuscript to be available before we created a DOI for the dataset. The dataset can now be accessed via HydroShare: The link to the dataset was added to the manuscript (*https://doi.org/10.4211/hs.069f552f96ef4e638f4bec281c5016ad*).*
To facilitate the reproduction of the functional data clustering approach, we added details on the R-packages and functions we used: 'The analysis is performed in R using the packages fda.usc [*Febrero-Bande and Oviedo de la Fuente*, 2012] *and fda* [*Ramsay et al.*, 2014] *and the following functions: (1) conversion of regimes to functional data objects: fdata, (2) creating of B-spline basis functions: create.bspline.basis, (3) computation of spline coefficients for all regimes: Data2fd. The clustering into regime classes is performed using the R-package stats (R Core Team, 2019). A Euclidean distance matrix is computed using the matrix of n= 671×5 spline coefficients (Figure 1.2 a–b) (dist). We use a hierarchical clustering algorithm (hclust) allowing for non-elliptical clusters (Gordon, 1999) with Ward's minimum variance criterion, which minimizes the total within-cluster variance (Ward, 1963). To identify an optimal number of clusters, we cut the tree a tk= 2,...,30 clusters (cutree) and compute the mean silhouette width (Rousseeuw, 1987), which provides a measure of clustering validity, for the different numbers of clusters'*

- Line 226: I am unsure if avoiding small clusters is a sign of a good clustering. As river behavior is a natural process, I would expect it to follow some kind of normal distribution, which would result in some bigger clusters and some smaller, more extreme clusters. For future research, it might be a good idea to explore a continuous classification as done by Knoben et al (https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR022913) or a fuzzy clustering approach to avoid the arbitrary cut-off points of clusters.

Reply: *One can always form new clusters of 'special cases' to decrease the within cluster variability. However, the formation of more clusters in our case resulted in a decrease in average silhouette width, which is not desirable. Furthermore, the formation of more clusters with smaller* *in-between cluster differences, would have diverted the focus from the detection of major regime changes to the detection of minor regime changes. We agree that the use of probabilistic instead of deterministic class memberships may be desirable for some applications. For this particular application, however, we found deterministic clusters to be more appropriate.*

- Line 256: The difference here might again be a question of granularity. Especially the catchments in Florida behave very uniquely.
Reply: *The finding that the correct regime class is not simulated in certain regions when forcing the model with GCM output is likely related to the fact that certain processes in these areas are not well represented by GCMs.*

- Figure 5: It is very difficult to distinguish the lines from each other here. I think it might be a good idea to increase the size of this figure. Also, I would recommend to use more easily distinguishable colors.
Reply: *Thank you for pointing out the need to increase the legibility of this figure. We rearranged the subplots of this figure in order to increase the size of the individual subplots and darkened the color of the control regime to increase contrast with respect to the regimes simulated using the GCM output.*
Modification: Figure 5

References used in this response to the reviewers

Addor, N., G. Nearing, C. Prieto, A. J. Newman, N. Le Vine, and M. P. Clark (2018), A ranking of hydrological signatures based on their predictability in space, *Water Resour. Res.*, *54*, 8792–8812, doi:10.1029/2018WR022606.

Ali, G., D. Tetzlaff, C. Soulsby, J. J. McDonnell, and R. Capell (2012), A comparison of similarity indices for catchment classification using a cross-regional dataset, *Adv. Water Resour.*, *40*, 11–22, doi:10.1016/j.advwatres.2012.01.008.

Febrero-Bande, M., and M. Oviedo de la Fuente (2012), Statistical Computing in Functional Data Analysis: The R Package fda.usc, *J. Stat. Softw.*, *51*(4), 1-3-, doi:10.18637/jss.v051.i04.

Melsen, L., N. Addor, N. Mizukami, A. Newman, P. Torfs, M. Clark, R. Uijlenhoet, and R. Teuling (2018), Mapping (dis) agreement in hydrologic projections, *Hydrol. Earth Syst. Sci.*, *22*, 1775–1791, doi:10.5194/hess-22-1775-2018.

Ramsay, J. O., H. Wickham, S. Graves, and G. Hooker (2014), Package “fda”: Functional data analysis, *CRAN*.