On behalf of myself and my co-authors, I would like to thank this review for these thoughtful and thorough critiques. It is our hope that these responses will help clarify and explain our choices as well as provide some additions to this work that will improve the overall product. As some of these comments do overlap with the commentary offered by Dr. Woods, some of these responses make reference to our previous replies. Thank you again.

1) I found the rationale for performing classification to be rather weak. For motivation, the authors state that (P7087, Line 6) ": : :a catchment's regime curve (ensemble mean of the within-year variation of runoff) has a major impact on the shape of the FDC". This rationale might be good enough to motivate why some of your four metrics were chosen, but has nothing to do with why classification is required in first place. Moreover, if the authors consider the shape of FDC to be an important hydrologic property of catchments, why not use FDC directly to perform classification?

Much of this question is well-addressed by the responses to Dr. Woods. We agree that further justification for classification itself is necessary and will ultimately be included in the final manuscript. Hopefully the response below, taken, in part, from responses to comments #1, #2, and #3 from Dr. Woods, sufficiently address the comment above .

The advantage of classifying seasonal climate regimes (precipitation vs. Ep/P) as well as seasonal flow regimes is, in addition to general, holistic understanding, the ability to gain insights into the nature of the FDC. The analysis in the first two installments of this four-part paper suggest that both seasonal climate and seasonal runoff play a role in understanding the FDC. An example can be added to the introduction to elaborate upon this point. For instance, two catchments, one in Missouri, and another in Georgia (Ye et al) might display very similar seasonal streamflow, but the subtler differences in seasonal precipitation result in a different FDC. Similarly, the two catchments in the figures 6a and 6b (from the response to Dr.Woods) from Washington state display a very similar pattern of seasonal precipitation, but will produce very different FDCs due to their different runoff patterns. Cheng et al (2012) illustrates that the parameters that define the FDC display a spatial pattern across the continent that is not explained by seasonality of streamflow or runoff exclusively. For this reason, it is only with both perspectives included that the full picture emerges. The introduction will be revised to include these points.

It is possible to classify a given signature using the observations of that signature exclusively, this is often mathematically successful, but less qualitatively insightful. For instance, the work of Haines et al (1988) shared a similar objective (and should be included / cited), but the methods employed differ from our work, as Haines et al clusters monthly flow regimes empirically (using the monthly flow regimes from many catchments). Haines et al does achieve the goal of obtaining clusters of catchments with similar flow regimes. However, as discussed in previous sections, similar flow regimes may not necessarily produce similar FDCs, as other climatic factors play a role (Cheng et al and Ye et al, 2012). Like our work, the algorithm aspires to create groups of similar behavior where the variance within is less than the variance without. However, Haines et al does not address WHY those groups appear as they do. As I interpret this work, the algorithm simply clusters streamflows, then discusses the different characteristics. The analog would be if we were to construct a tree to cluster only the FDCs. Presumably, we would be very good at it with a 25-year advantage in computing power, but what would be missing are the climatic and hydrologic characteristics that cause a catchment to behave the way it does. The work contained in Haines (and others taking a similar approach) tends to provide more "what" than "why." This point will be incorporated into the lit review.

2) In the current format, the introduction misses an opportunity to provide an overview of how previous studies have approached classification, and how the particular approach taken by the authors is going to be different/better than what has been done already. The references provided in the manuscript were

either opinion commentaries [Dooge, 1986; McDonnell and Woods, 2004] or reviews [Blöschl and Sivapalan, 1995; Olden et al., 2011], and without any critique or thoughts on previous approaches. There is a rich history of hydrologic classification studies that are very relevant to the approach used by the authors (see Mosley [1981], Ogunkoya [1988], Burn [1997], Burn and Goel [2000], Sawicz et al. [2011], and several studies referred in Olden et al.[2011]). I believe that setting up the current study in the context of previous classification efforts will greatly strengthen the paper.

This is a valid suggestion, though we do refer to Sawicz 2011. One of the pivotal differences between our work, and its predecessors is the scope of our classification attempts. For instance, Moseley (1981), like our work, attempts to classify hydrologic responses; in their case the mean annual flood and the coefficient of variation with respect to flood discharges. However, this system is highly specified for classifying ~175 comparatively small catchments in New Zealand, and thus includes numerous narrowly-defined characteristics and finely split classes. In fact, Moseley even suggests that broad regionalization would be more suitable in the American Midwest than such a small, physically complex area.

Ogunkoya (1988), like Moseley before, looks for nuances in a specific location – in this case, 15 catchments in Nigeria. These 15 catchments are investigated in terms of eleven parameters depicting the patterns of hydrologic response, then split into five clusters. This requires consideration of lithographic details and other nuances that might seem less appropriate in a classification that attempts to broadly applied and minimalist in its information requirements.

Burn (1997), finds some parallels with our work in its application of seasonality metrics to help understand flood frequencies. However, its geographic region of interest is an assortment of 59 prairie catchments in central/western Canada. These catchments were chosen specifically because they experience comparable climates, and thus, all present hydrologic regimes driven by flood events from spring snowmelt. As their database does not contain deserts, mountains, swamps, forests, or any number of other types of catchments, feature selection can be tailored differently than our database of 428 catchments that are tremendously diverse from a climatic standpoint.

Burn & Goel (2000) attempts to model a broader and more diverse assorted of catchments (India). However, using Koppen-Geiger as a reference, there are six climates that cover the entirety of India, three of which cover the vast majority of the land area. In contrast, the United States contains two to three times the number of climates. The k-means technique employed is certainly an effective method of extracting clusters, though the features used (catchment area, length of main stream of river, slope of main stream) are quite difficult to gather in certain locations. Moreover, as discussed in our previous comment, clustering algorithms of this ilk present groups that are similar without specifying the physical drivers that yield such similarity.

These works should be discussed and cited in our literature review – we appreciate the reviewer's statement that our work requires contextualization.

3) The authors state that they seek to develop a precursor to extending the Koppen-Geiger climate classification (basically a hydrologic equivalent). It might be helpful to provide more information about the Koppen-Geiger classification itself in the Introduction. Specifically, what variables did they use to perform classification? and what were the key reasons for their classification effort to be so successful? Was simplicity alone their strong point? This provides a strong motivation to extend this classification approach into hydrology (the primary aim of the authors).

The Koppen-Geiger classification system focuses primarily on temperature and aridity. More specifically, temperature and precipitation values averaged on monthly and annual scales. Boundaries between regions are generally defined by shifts in native vegetation. Without question, our work

attempts to incorporate this information with our inclusion of precipitation seasonality and aridity index. Koppen-Geiger's appeal lies in the simplicity of five primary groups (which are then subset into more specific groups) along with, like our classification system, the relative ease of accessing the necessary information at any location. However, by excluding hydrology from the system, it fails to distinguish certain catchments that intuitively, clearly display different weather, different climates, and different filtering behavior. Consider the fact that Koppen-Geiger considers the entire southern half of the United States, from Texas through the bayous to Florida, and northward into the great plains, Appalachian mountains, and coastal, Atlantic catchments to be the same climate class. Some of these catchments present aridity indices less than 0.4 (North Carolina) or as high as 1.9 (Oklahoma & Texas). Some of these catchments receive steady monthly rainfall (mid-Atlantic) while others are notably seasonal (Midwest). Similarly, Koppen-Geiger classified catchments in Washington state and Oregon (among the most humid catchments in the United States) identically to catchments in Southern California (often among the most arid). Understanding the distinctions in rainfall/runoff timing allows for more nuanced understanding. The classification system we have offered never classifies very warm and dry catchments identically with cool, humid locations. This point should be integrated into the manuscript in the lit review, better contextualizing our work.

4) P 7089, Line 22: I found the paragraph summarizing other studies in this series to be a distraction from the classification message of this paper.

Does this opinion intend to suggest its omission altogether or its relocation to a more suitable position?

5) P 7092, Line 8: "While this image does provide useful information about the within year (daily) variability of the chosen variables, for the purpose of catchment classification in this paper, a sliding, 30-day moving average is generated". Please provide a rationale for using a 30-day moving average filter.

Many hydrological analyses (Koppen-Geiger, Haines et al, and others) deploy monthly regime data to depict seasonal patterns of rainfall and runoff. A 30-day moving average achieves this idea of a 30-day window without creating arbitrary boundaries. In this regard, a smoothed regime curve with 365 days of sensitivity is achieved without losing a connection to the previous work done with monthly regime curves. This can be explained in the methods section.

6) The following are fairly sweeping statements without any back-up or citations from the authors: P 7093, Line 17: "These four variables are chosen not only because they are succinct descriptors of processes that underpin seasonality of runoff, but also because they represent the minimum amount of information that is needed to classify regime behavior within the continental US". P 7097, Line 25: "Our hypothesis in this paper is that a combination of the 4 similarity indices governs the regime behavior and can be the basis of their classification". First, it is not even clear if there is redundancy among the four chosen variables. For example, day of peak precipitation and day of peak runoff could be highly correlated in many, if not all, places. Therefore, minimality of the information content is a big unknown here. The study by Sawicz et al. [2011] offers a good approach into how this issue can be dealt with. Second, 3 of the 4 chosen variables are climatic metrics. This leads to an implicit assumption by the authors that climatic similarity is the primary controller of hydrologic similarity. While this has been shown to occur over large regions by previous studies (e.g., see Patil and Stieglitz [2012]), it would be helpful to explicitly state this assumption if the authors wish to limit the classification to these 4 variables.

Line 17: In the response to Dr. Woods each of the four variables was removed and the classification system reconstructed. The loss of clustering power was substantial. Pairs of catchments are also presented where four variables are sufficient to distinguish, but three are not. Inspection of over 400 regime curve images like those presented in figures 1a and 1b reveal that if the four key indices are largely similar, so too will be the overall regime behavior. However, many examples can be shown

where the three of the four variables are similar, but the catchments' regimes are not. Other variables were considered, such as Q/P – but ultimately eschewed because they are correlated with other variables (E_p/P) and failed to improve the quality of classification.

Line 25:A quick point on independence. Seasonality and aridity index are almost entirely independent (r^2~0.14). Seasonality and date of max precipitation are fully independent (r^2 < 0.01). Seasonality and date of max runoff are almost entirely independent (r^2~ 0.14). Aridity index and date of max precipitation are independent (r^2~ 0.05). The same is true for aridity index and date of max runoff (r^2~0.06). Counter-intuitive though it might be, the r^2 value connecting the date of maximum precipitation and the date of maximum runoff is only 0.21. Though there are clusters where the maximum runoff follows the maximum rainfall by a few days or weeks, there are also numerous catchments with virtually constant annual rainfall, yet still characterized by a defined runoff peak. Finally, there are catchments that receive their highest rates of precipitation during fall/winter, then store that water in snowcaps, yielding peak runoff in April, May, or June. These distinct cases explain the need for all four variables. These quantifications of independence can be discussed in the methods section where the four indices are initially described.

The reviewer comments that "three of the four chosen variables are climatic metrics." Certainly, this is factual. However, the hydrological piece is not merely the inclusion of the date of maximum streamflow, but rather, the interplay between the date of maximum precipitation and the subsequent date of maximum runoff. Furthermore, this timing integrates with the seasonality of rainfall to depict the timing of rainfall's arrival and the seasonal residence times that lead to runoff production. Finally, adding the concept of aridity determines the fraction of water that arrives in various seasons that ultimately has the capacity to form runoff (is E_p in phase with P or not?). Thus, though three metrics are climatic, all four play a hydrologic role as well. This insight probably fits in introductory discussion of what this paper aims to achieve.

7) P 7094, Line 1: "Once the classification system is established, even approximate or fuzzy answers to these questions can help towards a first-order classification of regime behavior, subject to further data collection and analysis". I do not understand the need for such a statement. There are many more questions in hydrology than the four chosen by the authors, and fuzzy answers to any of those questions will lead to a first-order classification. This is precisely the reason why there is no universally accepted catchment classification system.

We recognize that perhaps this phrase was not optimally worded. The intention was not to argue that these four indices can receive fuzzy responses and yield a first-order classification while no other set of indices could make such a claim. The point we were attempting to convey is that for many classification systems, specific information is required, without which a location's classification could shift dramatically. Consider with Koppen-Geiger, tropical rainforests are classified because all 12 months have an average rainfall greater than 60mm. This requires some significant data estimation if gauges are unavailable. However, the variables in our classification system, in terms of placing a catchment on our tree require only "fuzzy" inferences. In other words, in a hypothetical catchment where a local resident can tell us, "the heaviest rain arrives in the summer, it does not rain much during the winter, and our streams are highest in the summer when the rain is highest" and our observations reveal vegetation consistent with semi-arid climates, we can classify to "ISQJ." No further information is needed. With Koppen-Geiger, it might be very difficult to estimate how many months exceeda given threshold for precipitation or temperature and a "fuzzy" estimate might be insufficient. Going back to the "ISQJ" example, if we examine another similar catchment located farther north and a denizen tells us once again "it rains most frequently in the summer, but less often in the winter," but then tells us "our streams are highest in early spring" and we see vegetation consistent with a more temperate climate, we classify to "ITC." An approximation is fine for our tree where for other systems this type of approximation might be inadequate. Perhaps this should be added to the discussion section.

8) P 7102, Section 4.1: No rationale is provided for the class divisions of the four variables. For instance, why does aridity index have 5 classes, but seasonality index and day of precipitation peak have only 3 and 4 classes respectively? This has important implications on how the clustering results are interpreted.

The class divisions occurred after seeing the clusters that emerged. For instance, with aridity index, there were a couple classes where E_p/P fell well below 0.5, some classes with $E_p/P > 2.5$, and three notable groupings in between. For this reason, 5 classes were selected. However, with respect to seasonality, in examining groups it became evident that there were catchments with very little seasonality, catchments with extremely high rates of seasonality, and intermediate catchments. Thus, three were chosen. The intention had been to generate as few classes as possible. This could be added briefly to the results section.

Works Cited

Burn, D. H. (1997), Catchment similarity for regional flood frequency analysis using seasonality measures, Journal of Hydrology, 202(1-4), 212-230, doi: 10.1016/s0022-1694(97)00068-1.

Burn, D. H., and N. K. Goel (2000), The formation of groups for regional flood frequency analysis, Hydrological Sciences Journal, 45(1), 97-112, doi: 10.1080/02626660009492308.

Haines, A. T., Finlayson, B. L., and McMahon, T. A., (1988) A global classification of river Regimes, Applied Geography, 8, 255-272.

Mosley, M. P. (1981), Delimitation of New Zealand hydrologic regions, Journal of Hydrology, 49(1-2), 173-192, doi: 10.1016/0022-1694(81)90211-0.

Ogunkoya, O. O. (1988), Towards a delimitation of southwestern Nigeria into hydrological regions, Journal of Hydrology, 99(1-2), 165-177, doi: 10.1016/0022-1694(88)90085-6.

Sawicz, K., T. Wagener, M. Sivapalan, P. A. Troch, and G. Carrillo (2011), Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA, Hydrology and Earth System Sciences, 15(9), 2895-2911, doi: 10.5194/hess-15-2895-2011.