On behalf of myself and my co-authors, I would like to thank this reviewer for his/her commentary. Many of these comments were raised by other reviewers as well – which does tend to solidify their importance. We will attempt, in these responses, to better place this work in its proper context in terms of the existing literature and the other three papers.

This manuscript addresses the important issues of catchment classification and the need to understand controls on catchment similarity.

I have several comments on this manuscript:

I have read one of the other manuscripts (Part IV) in this four-part series and I am uncertain of the innovative contributions of this manuscript, particularly in relation to the other manuscripts. The manuscript does not use the flow-duration curve in the analysis and the context of this work in relation to the other three papers is not clear. I am also not convinced a classification on the four seasonality variables moves forward the science of catchment classification beyond previously published work. To address these concerns, the literature review needs to be strengthened to make the case for why this analysis is innovative and unique relative to other published classifications. The introduction should also include more detail about how this paper relates to the others in the series.

This paper, in general, attempts to classify seasonal runoff behavior (more specifically, the seasonal flow regime for daily streamflow) using four fairly basic, easily-accessible features. These features are chosen, in part, from insights gained from the distribution-modeling in Cheng et al (2012) and the process-modeling in Ye et al (2012). These four variables then form a classification system that clusters catchments that are similar with respect to these indices. Encouragingly, these groups of catchments display similar runoff regimes (as hypothesized), and also form regions that are largely geographically continuous. The classification system constructed herein is similar in many ways to the classification system developed by Koppen a century ago, though in this case, hydrologic information is included in the classification methodology where Koppen excludes such features. Koppen's work has been largely viewed as an acceptable mechanism for climate classification – we contend that our work, analogously, is appropriate for classifying seasonal runoff behavior.

There is certainly other work within the literature with similar goals. For instance, Haines et al (1988) shared similar objectives (and will be cited in a revision), but the methods differ from our classification system, as Haines et al clusters monthly flow regimes from many catchments empirically. While this analysis does obtain clusters of catchments with similar flow regimes, two catchments with similar flow regimes may not necessarily produce similar FDCs due to other climatic factors (Cheng et 2012, Ye et al, 2012). Like our analysis, Haines et al does create groups of similar catchments where variance within is lower than variance without. However, Haines et al does not address WHY these groups appear as they do, simply clustering streamflows, then discussing the characteristics of the streamflows. The analog would be were our tree to cluster only the FDCs – a task at which we might excel given a 25-year advantage in computing power over Haines et al, but would omit the hydrologic and climatic traits that cause a catchment to behave as it does.

Another key difference between our work and its predecessors is the geographic scope of classification. For example, Moseley (1981) also attempts to classify hydrologic responses – focusing upon the mean annual flood and the coefficient of variation with respect to flood discharges. The distinction lies in the fact that Moseley (1981) is constructed to delimit ~175 relatively small catchments in New Zealand. This requires numerous narrowly-defined characteristics and finely split classes.

Ogunkoya (1988) also searches for fine understanding of a specific location, choosing 15 catchments in Nigeria to analyze. These locations are assessed in terms of eleven parameters that describe hydrologic responses, then partitioned into five groups. Lithographic features are considered along with other details that might be less applicable to a broadly-defined classification system that aims to be minimalist in its information requirements.

Burn (1997) overlaps with our analysis insofar as it applies seasonality metrics to better understand flood frequency. Once again however, by focusing upon 59 prairie catchments in central/western Canada, specifically chosen because of their comparable climates, the goal becomes a much narrower region of interest. Their database does not contain deserts, mountains, swamps, forests, etc – only regimes driven by snowmelt-driven spring floods. Thus, feature selection can be tailored in a manner that would not suffice within our database of 428 climatically-diverse catchments.

Burn & Goel (2000) does model a broader and more diverse group of catchments, covering the whole of India. However, using Koppen-Geiger's classifications, there are six climates that cover the entirety of India, three of which cover the a significant majority of the land area. The United States, in contrast, contains two to three times the diversity of climates. Burn and Goel deploy the k-means tool to extract clusters, though their choices for features (catchment area, length of main stream of river, slope of main stream) are a challenge to obtain in certain locations. As discussed, clustering algorithms of this nature yield groups that are similar, but do not specify the physical drivers of that similarity.

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

With respect to our goals of extending Koppen-Geiger, the Koppen-Geiger system focuses largely on temperature and aridity, averaging precipitation and temperature values on monthly and annual scales. The boundaries between regions are generally specified by observing shifts in native vegetation. Our work aims to incorporate this information via inclusion of seasonality of precipitation and the aridity index. One of the appeals of Koppen-Geiger is the simplicity of its five primary groups (A – E, then subset into more detailed groups) along with, like our classification algorithm, the relative ease of accessing necessary information at a variety of locations. However, in excluding hydrologic information, Koppen-Geiger does not distinguish certain catchments that clearly display different weather, climates, and filtering behavior. For instance, Koppen-Geiger labels the entire southern half of the U.S., from Texas to Florida (through the Bayous), northward through the great plains, Appalachians, and mid-Atlantic coasts as one single climate class. Some of these catchments are arid (Oklahoma & Texas, E_p/P ~1.9) while others are very humid (North Carolina, $E_p/P < \sim 0.4$). Some of these catchments receive steady monthly rainfall (mid-Atlantic) and others are notably seasonal (Midwest). Likewise, Koppen-Geiger classified catchments in Washington state and Oregon (some of the *most humid locations in the U.S.) identically to certain locations in Southern California (some of the most arid). Understanding the distinctions in rainfall/runoff timing does facilitate more nuanced understanding – our classification system will never classify a warm/dry location identically with cool/humid locations. We will integrate this idea into the lit review as well.*

Finally, though admittedly this work focuses on the regime curve, it is because of our hypothesis that this classification will ultimately serve to enhance our understanding of runoff behavior. Figure 2, from our response to reviewer #2, presents a table to illustrate the decreasing variability with respect to flow duration curves with each step down the tree. Figure 14 from the original manuscript illustrates that annual regime curves also cluster by class effectively.

My other major comment relates to the use of the four similarity metrics. Two of these metrics incorporate streamflow: 1) the aridity index and 2) the timing of maximum runoff. Therefore, I would not characterize the conclusions as having power "for predicting regime behavior across the continental US." (p. 7111; line 2) In my mind, regime dynamics go beyond only timing of the maximum runoff and mean annual streamflow (which is part of the aridity index) and I do not think extrapolation is justified. This again makes me question the unique contribution of this manuscript.

While the timing of maximum runoff, incorporates streamflow, the aridity index does not. The aridity index is calculated simply by summing the total potential evapotranspiration (PET) and dividing by the sum of all precipitation received. Runoff is absent from this calculation. While there is no question that regime dynamics can include factors that are not directly considered within the classification system developed, after reviewing hundreds of regime curves (like those in figures 1a and 1b in the discussion paper), these four indicators were sufficient to cluster those that appear similar, while any three could be insufficient in certain cases (see the paired images in figures 3-6 from the response to reviewer #2).

The clustering algorithm description is detailed but does not mention why this particular algorithm was selected over others. Is there something more desirable about this type of clustering algorithm than other algorithms?

The rationale for a classification tree rather than another classification algorithm of which there are hundreds, was that this particular structure enables analysts to gain qualitative insights. This insights emerge naturally throughout the tree's development process rather than a black box that delivers groups, but not the necessary explanation. Neural networks, nearest-neighbor algorithms, k-means, genetic algorithms, and numerous others can be useful for classification, but obscure the intermediate steps. This prevents making meaningful connections to the physical features involved. With this approach, as 'observers' of the algorithm, we view which splits occur on which values at which point in the bifurcation process. This allows the questions to be asked, "what is the most important, most distinguishing characteristic for all U.S. catchments?", "what if we only consider the non-seasonal half?", etc. While clearly other algorithms for clustering could have been chosen, it becomes unwieldy to assert "this is the very best possible technique." However, it is, in our humble opinion, fair and proper to state that this method is appropriate for the problem at hand, and performs quite well.

Author's note: In response #9, to reviewer #2, we enumerate the variability of the four indices within the biggest groups as compared to their variability over the dataset as a whole (428 catchments). This helps substantiate the claim "performs quite well."

I am not sure this manuscript contains enough new and unique contributions to be considered a stand-alone paper. I do commend the authors for their analysis of the stability of the classes by using a smaller dataset of catchments and in how they assigned class names to the resulting classification. It is my hope that the authors' responses to these comments will serve to underscore the innovations in the paper and that the paper will ultimately be accepted.

We hope the commentary above, the other classification systems used to place our work in its appropriate context, and our explanation to this and other reviewers will substantiate this analysis.

Editorial comments:
p. 7100, 8-16: This paragraphs reads differently from the rest of the text. Should there be a reference here that is used to describe the algorithm?

The algorithm described is an adaptation of Brieman et al, 1984 & 1993, both of which are cited in the previous paragraph.

There are a few places where the text is quite informal and should be edited: p. 7100, line 3: Remove "from scratch."

We will remove those two words.

Section 4.4: Use of "Big 6." Can you say "the largest six" classes in place of this?

We will make this change as well.

p. 7112, line 9: Should be a question mark at the end of the sentence and not a period.

Agreed. We will make the change.