My general responses to the numbered comments, followed by the analysis performed to support those responses:

1. It has been common practice to make classifications of seasonal climate regimes, or of seasonal flow regimes, but not of both at once. By doing both simultaneously, the authors have taken a new approach; I think it is a legitimate choice which may have some advantages, but they need to point this out more clearly, because it explains some of the choices they have made. Why did the authors choose to classify the seasonality of climate and flow regime? Why did they not just classify the flow regime?

The advantage of classifying seasonal climate regimes (precipitation vs. E_p/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 Washington state display a very similar pattern of seasonal precipitation, but will produce very different FDCs due to their different runoff patterns. Since the timing of these peaks are not always dramatically different, the climate indicator E_p/P helps distinguish wetter and dryer catchments with similar seasonal timing of rainfall and runoff. 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.

2. In my experience, successful (in the sense of widely adopted) classifications depend on well-informed subjective decisions. In this case, the decisions include the selection of the 4 indices, the particular objective function chosen, and method for splitting catchments into groups. I think the authors should note some of the major alternatives which they did not pursue (e.g. indices which quantify the amplitude and/or phase of P-Ep or which quantify proportion of precipitation which falls as snow, an objective function which scaled the variances differently, a different splitting/grouping method), and why not.

Some of these choices are more directly defended than others. For instance, the phase of E_p/P is not addressed because within the continental U.S., every catchment's E_p curve peaks within a couple of weeks during the summer. The amplitude of that curve is implicitly included in the aridity index variable E_p/P, as a higher amplitude curve of E_p would yield a higher value for E_p/P. For most catchments in the USA, E_p is very low during winter months – ½ of the amplitude would nearly approximate the mean. The decisions with respect to the four indices are justified in terms of understanding the interplay between wetting and drying, and the timing separating rainfall from runoff, as discussed in the previous two papers. As the groups of images appended illustrate, any three indices are insufficient to understand the nuanced behavior of the catchments we examined, but the addition of the missing fourth (at least for the vast majority of MOPEX catchments) resolves the discrepancies. As for the proportion of precipitation that falls as snow, this and other indicators are undoubtedly relevant. The question is, as with all

hydrologic modeling, when and where to "draw the line." The proportion of precipitation which falls as snow is a function of broader climatic themes such as average temperature, the timing of rainfall (more during winter or summer), etc. These concepts are included, at least in large part, in the aridity index, the seasonality index, and the timing of the maximum day of precipitation. The choice to use a greedy-splitting algorithm to minimize within-group variance with each step down the tree is a fairly standard procedure. The reason for a classification tree rather than, say, another clustering algorithm (of which there are literally hundreds), was that this structure allows for qualitative insights to emerge along the way rather than a black box that delivers groups without explanation. Neural networks, nearest-neighbor algorithms, genetic algorithms, and many others can be useful for classification as well, but obfuscate the intermediate steps, and do not allow for the same connection to the physical insights gained. With this method, as "observers" of the algorithm, we can see what splits occur on what values at what point in the process, allowing us to ask "What is the most important, most distinguishing characteristic for all U.S. catchments?", "What if we only consider the non-seasonal half?", etc. Again, other mechanisms could have been chosen, and as with any classification system, it becomes somewhat unwieldy to assert "this is the very best possible technique." It is, in our humble opinion, proper to say "this method is appropriate for the problem at hand, and performs well (see response #9)."

3. The authors refer to the idea of extending the Koeppen classification so that it applies to hydrology. If the purpose of the paper is to help develop a Koeppen-like system which classifies seasonal flow regimes, why not adapt or build on the decision tree approach shown in Figure 2 of Haines et al (1988)? That paper only uses the monthly runoff data, and does not use climate data. Why do the authors consider that a classiflcation of seasonal hydrology requires any data beyond runoff data?

While the work of Haines et al shared a similar objective (and should be included / cited); the methods employed differ, clustering monthly flow regimes empirically. 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 Q regime curves. 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 tends to provide more "what" than "why." This point will be incorporated into the lit review. See comments #1 and #2 for a more thorough justification of why seasonal runoff information is useful, but ultimately, an incomplete picture.

4. The authors do not provide any evidence to substantiate their bold claim that the four indices "… represent the minimum amount of information that is needed to classify regime behavior within the continental US". Where is that evidence?

To help validate this statement, the figure 1 presents the decrease in variance as a function of layers down the classification tree using all four of the chosen indices as possible split criteria (the analysis that forms the crux of this paper). The paper's classification tree can be reconstructed with only three of the four indices (omitting each of the four in separate trials). In each case, the deterioration of the ability to reduce variance within groups is pronounced, and thus it can be argued that none of the four variables can be safely omitted. Figure 1, which should be included (or referenced) in section 4.4 or 4.5 demonstrates that each of the four variables substantially contributes to the power of the classification. We have determined that these four are sufficient, recognizing that future researchers might consider the proportion of variance explained to be inadequate, and add further indicators to improve specification.

5. Another way of asking the same question is to seek clarification over the authors' choice of objective function. Why is "regime similarity" defined as being composed of similarity in seasonality, aridity, timing of precipitation and timing of runoff?

The objective function was chosen after inspection of the 400+ regime curves (images like 1a. and 1b. from the discussion paper) and noting that if these four key features are very similar for any two catchments, so too will be the regime images, while if only three are very similar, the two regime images might differ notably. Please see the paired images at the end of this document for illustration of this concept.

6. I find the rationale behind the 4 indices chosen is acceptable, but incomplete. If the purpose is to identify a minimal set of variables which can be used to discriminate amongst different seasonal regimes, why are there no comparisons of competing groups of variables, nor any quantitative measure of the success of this particular set of 4 indices. Indeed how is success quantified? For example, what is the within-group and between-group heterogeneity of monthly flow regimes for the groups shown in Fig 14?

The objective function that is minimized with each recursive split of the tree is the within group variance with respect to the four key indices. Figure 1 illustrates that this value is effectively decreased with each step down the tree. This is the measure of success of the method. Moreover, figure 2 displays the variance with respect to 100 key percentiles of the FDC as we progress down the tree. In other words, not only are the four key indices being grouped, but the FDCs of the constituent groups are well-organized as well.

While there are other variables that could be used in the construction of another tree, for instance, the average temperature, the works from Cheng et al and Ye et al (2012) imply that these four processes are pivotal and the results suggest they are effective descriptors.

7. "a key objective of this research is the classification of regime behavior using an absolute minimum quantity of data" The authors have not shown that all 4 indices are essential. Would the results have been similar if one or more of the 4 indices was omitted?

See comment #4; this analysis is presented in the chart.

8. “several classes of catchments are distinguished, in which the connection between the catchments’ regime behavior and climate and catchment properties becomes selfevident” I do not think that the connections are self-evident; indeed the description of processes in the Conclusion relies on a considerable body of knowledge (e.g. roles of snow and frozen soil) not included in the classification.

The word “self-evident” should be replaced with “clearer.” Consider for instance, the class ISQJ in which a somewhat seasonal (I) catchment with a semi-arid (S) climate does not store precipitation in the form of snowpacks, and thus its runoff peaks in summer at roughly the same time as its precipitation (Q & J). In terms of freezing/thawing, as mentioned previously, while these are not explicitly contained in the classification system, the combination of the four indices contains this information implicitly.

9. How similar are the flow regimes of the groups derived? [the graphs in Fig 14 are too small to gain more than a qualitative impression – I cannot read the vertical axis to see the units or scale]

In terms of the overall variance of the full dataset, the following are the within-group variances for the six most common classes. LWC – 26.9%, LPC – 23.9%, LPM – 29.8%, LJ – 43.3% (with 140+ catchments), ITC – 28.1%, ISQJ – 46.5%. Considering that these groups comprise 77% of the database, this is quite encouraging, as these clusters contain must less than half of the variance of the original dataset using very simple indices. Moreover, we could always split on other, more detailed variables addressing snow, freezing/thawing, or other more nuanced characteristics at a later date to further partition the groups found. Figure 14 (discussion paper) should be improved to make its scale information clearer to readers.

Appended Images and Description (To be included, potentially, in section 2.2)

The two catchments in figures 3a. and 3b. are very similar in terms of aridity index, and their peak days for rainfall/runoff. However, the mountainous catchment in Montana (3a.) receives precipitation quite evenly throughout the year, while the Midwestern catchment in Iowa (3b.) receives more rainfall during summer months. The winter precipitation in Montana forms snowpacks, leading to peak runoff in the form of melt water in early June. The catchment in Iowa sees its runoff maximized during the same week as well – but the driver is rainfall, not melting. This distinction is nicely distinguished by the variable “seasonality,” yet thoroughly missed by the other three variables.

The catchments in figures 4a through 4c are virtually identical in terms of seasonality, as none of them display a strong seasonal signature for rainfall apart from a slightly higher quantity of winter precipitation. All three catchments display peak rainfall within essentially one week in early January and peak runoff within the same week of late-May/early-June. Without knowledge of aridity, these catchments would almost certainly fall within the same class. However, in looking at the catchment in 4c., the quantity of melt-driven runoff is nearly a full order of magnitude larger than the catchment in 4a. The difference is nicely explained by the substantial differences in aridity index.

The catchments in figures 5a and 5b display a very similar quantity of seasonality with respect to precipitation, comparable aridity, and peak runoff during the same week. However, these catchments are distinguished by the fact that the catchment in 5a receives its precipitation during winter, out-of-phase with respect to PE, and thus, accumulates snow which exits as melting snow months later. The catchment in 5b, because its precipitation pattern peak is shifted almost exactly ½ year from the catchment on the left, receives precipitation in-phase with PE, and produces, despite the similar timing, dramatically less runoff.

The catchments in figures 6a and 6b are both located in Washington. Both present significant seasonality, extremely humid climates, and seasonal precipitation that arrives out-of-phase with PE, peaking on the very same day in late November. However, these two catchments present distinctly different climates as one emits maximum runoff in December (6a) and the other peaks in June (6b). This implies a differing mechanism of runoff – 6a receives runoff from winter rainfall that exits immediately (low residence time), shown by the Q regime curve mirroring the P regime curve, while 6b produces runoff from winter rainfall and even more notably from spring melt, thus producing a Q regime curve that does not follow P.

Figure 1. Decreasing entropy down the tree –performance with each variable removed

Figure 2. Decreasing FDC variance (layer-by-layer down the tree)

Figures 3a. and 3b.
(Left) #203, Montana – E_p/P ~ 1.1, MaxDayP = 152, MaxDayQ = 165… Seasonality ~ 0.15
(Right) #91, Iowa – E_p/P ~ 1.3, MaxDayP = 151, MaxDayQ = 166…Seasonality ~ 0.39

Figures 4a. , 4b. , and 4c.
(Left) #243, Colorado – Seasonality ~ 0.19, MaxDayP = 13, MaxDayQ = 148…E_p/P ~ 1.7
(Center) #300, Montana – Seasonality ~ 0.23, MaxDayP = 5, MaxDayQ = 151… E_p/P ~ 1.2
(Right) #342, Montana – Seasonality ~ 0.23, MaxDayP = 10, MaxDayQ = 153… E_p/P ~ 0.7

Figures 5a. and 5b.
(Left) #162, Idaho – Seasonality ~ 0.48, E_p/P ~ 1.0, MaxDayQ = 149… MaxDayP = 8
(Right) #388, Iowa – Seasonality ~ 0.45, E_p/P ~ 1.2, MaxDayQ = 152… MaxDayP = 175

Figures 6a. and 6b.
(Left) #346, Washington – Seasonality ~ 0.52, E_p/P ~ 0.38, MaxDayP = 330… MaxDayQ = 344
(Left) #392, Washington – Seasonality ~ 0.49, E_p/P ~ 0.35, MaxDayP = 330… MaxDayQ = 159