Reviewer 1

1. In general, the paper is well referenced, logically presented, and the figures support the results. The methods suggested are potentially useful for many and are described in a manner that makes it easy to see the application. Improvements can be made with more careful wording related to the statistical methods, some additional references, and some simple changes to the figures and tables.

R: We would like to thank Reviewer 1 for a thoughtful and constructive revision. In the new version of the MS we addressed all of the concerns from the reviewer and we incorporated the suggested modifications to the figures and tables. In addition, we modified the wording related to statistical methods and we completed the list of references.

2. The formula used for skewness is not simply the third standardized moment. It is the adjusted Fisher-Pearson standardized moment coefficient. The authors should state this, reference it, and perhaps tell potential users of the methods they are proposing why this version of skewness is desirable.

R: We are aware that in large samples, the differences in definition are unimportant, but for small samples very different values of skewness and kurtosis can be obtained by using the various existing definitions. We clarified this in the text as follows "Although time series of environmental data may include large datasets often they are incomplete due to missing values and errors. To account for a potential bias inherent to incomplete time series or in cases of small samples sizes, we used the sample skewness or adjusted Fisher-Pearson standardized moment coefficient and the sample excess kurtosis (Joanes and Gill 1998)."

Reference

Joanes DN, Gill CA (1998) Comparing measures of sample skewness and kurtosis. Journal of the Royal Statistical Society (Series D): The Statistician, 47, 183-189.

3. Likewise, different statistical packages compute somewhat different versions of kurtosis. This appears to be closest to that of Sheskin, D.J. (2000) Handbook of Parametric and Nonparametric Statistical Procedures, Second Edition. Boca Raton, Florida: Chapman & Hall/CRC. The authors should verify this, state the version of kurtosis used, and reference it.

R: We addressed this comment in the response above.

4. Regarding the description of the Cramer test of whether or not the skewness coefficient is different from 0, the null hypothesis is misstated. The authors say "... we could not reject the null hypothesis that the distribution was skewed ("non-significant"). The null hypothesis is that the skew is equal to 0 (symmetric; Cramer, 1998). For parallel construction, the null hypothesis for the excess kurtosis should also be stated, that excess kurtosis is 0, or the distribution is mesokurtic (Cramer, 1998).

R: Based on the comments from Reviewer 2 we decided to remove this analysis in our revised MS.

5. The manuscript is generally well reference, however, there should be a reference for non-metric multidimensional scaling (N-MDS) unconstrained ordination when it is first discussed in section 2.3.

R: We have added a proper reference to this section.

6. I could not find a reference to figure 4 in the manuscript but did find a reference to figure 4 of the supplement. Because the supplemental figure is what was discussed, fig. 4 manuscript and fig. 4 supplement should be switched.

R: In the revised version, we moved figure 4 and 5 from the Supplement to the MS.

7. Results are discussed in terms of unregulated and regulated streams. To better highlight these important differences, all figures and tables (where relevant) should distinguish between regulated and unregulated sites. This would be very helpful for the reader. For example, in table 2, a line could be added between sites 5 and 6 with spanners indicating sites 1-5 are unregulated and sites 6-10 are regulated. Figure 1 for example, should have the term unregulated placed in 1b and the term regulated in 1d. That way at a glance, the reader could see the difference without reading the extensive caption.

R: In the revised figures and tables, we have incorporated this information to improve their clarity and visualization.

8. For figure 1, the plot in position c is discussed first, and then the plot in position a - this happens on both pages 6 and 7. It flows better for the reader if the plot positions in figure 1 were switched so that the one in position c now becomes the one in position a, then the text can refer to a first. Also, currently a and c are paired and b and d are paired. Pairing a and b, then c and d is a more natural way of presentation, as is reading left to right. Organizing and labeling the figure as a in the upper left (higher kurtosis, positively skewed), b in the upper right (lower kurtosis, negatively skewed), c in the lower left (unregulated cluster), and d the lower right (regulated cluster), would be easier for the reader to follow.

R: This is a good point. In the new version of the MS, we have changed the layout of the plots for figure 1 to facilitate this for readers.

9. The authors are careful to document the stress in figures 4, S3, and S4. However, they do not provide enough information for readers to be able to interpret that value. For example, looking at the HDR boxplots with stress values of 0.17 and 0.16, they can have rather different shapes. Therefore, a sentence or two describing how stress was calculated would help.

R: In the new version of the MS we added a more comprehensive explanation about the stress. We added the following wording “The Kruskal’s stress value is estimated as the square root of the ratio of the squared differences between the calculated distances and the plotted distances,

and the sum of the plotted distances squared (Kruskal 1964). A rule of thumb (Clarke 1993) suggests the following benchmarks: stress <0.05 – excellent ordination; stress <0.1 - good ordination; stress <0.2 acceptable ordination; stress >0.2 – poor ordination. The resulting coordinates 1 and 2 from the resulted optimized 2-D plot provided a collective index of how unique a given year was (Fig. 1c,d)".

10. There are a few grammatical corrections that need to be made: line 3, page 7, change "may indicates" to "may indicate"; line 4, page 7, change "extremes" to "extreme", or change sentence to something like " … both extremes (cold and warm values);" line 12, page 10, change "this" to "these"

R: We modified the text accordingly.

Reviewer 2

1. The work presented by Arismendi et al. represents a significant contribution to the discussion of environmental statistics, but I do not feel that it is ready for full publication yet. The authors present two techniques for assessing important shifts in the distributional properties of environmental variables. They first argue that higher-order moments, beyond the mean and variance, are better at capturing distributional shifts. They then present a technique for outlier detection, which is useful for the identification of potentially anomalous years. In my review, I am concerned with a number of statistical questions that I feel must be mentioned and addressed in the manuscript.

R: We would like to thank Reviewer 2 for a thoughtful, very detailed and constructive review. We believe that our proposed revisions in response to the reviewer's comments will significantly improve the manuscript. As recommended by this reviewer, we reworded statements and re-structured the MS by combining results and discussion into one section as well as the addition of a summary and conclusions section. In our revised MS, we addressed the statistical questions from this reviewer and added a new paragraph in the results/discussion section about potential caveats using these techniques. We consider the last paragraph particularly useful to future users which will adopt these approaches to visualize and analyze large environmental datasets. Lastly, we provided a revised introduction based on the suggestions from this reviewer.

2. In general, the paper is well-written and the material is presented in a logical fashion, but I feel that a more rigorous justification and discussion of the results in necessary. Furthermore, I feel that there are a few statistical questions that must be considered. By more thoroughly discussing the points presented below, I feel that the authors will improve the impact and presentation of their findings. This is a valuable discussion, and with improvements, I feel it should be considered for publication.

R: In the revised version of the MS, we reorganized our discussion section (see response above). However, we feel that a deeper discussion about the implications of our finding was outside the scope of our study. We used stream temperature only as an illustrative example. In our revised

MS, we have carefully considered the concerns of this reviewer by revising and modifying statements in the discussion that were not justified enough by our results.

3. Firstly, I would request that the authors revise the manuscript to reflect a more precise use of statistical language. I do not intend this to be a question of nit-picking, but I feel that more precise language will more clearly reflect the authors' intent and further substantiate their findings. To begin with, the authors state their intent to capture the 'variability' of the data. This term is not a concrete statistical term. As stated later in the manuscript, the authors appear to be more concerned with 'changes in [the] empirical distributions' (P:4731, L:7-8) and the 'shift in the shape of the … distribution' (P:4732, L:23). The term 'variability' does not describe the behavior of the entire distribution in a statistical sense; typically, this term is most closely related only to the second moment. In a similar vein, the authors note (P:4731, L:2-3) that metrics of central tendency do not capture this variability. I would argue that this is, of course, true, as metrics of central tendency are intentionally designed to capture just that, central tendency, not variability. By adjusting the language to reflect an interest in higher-order shifts in the distributional properties beyond the location, I believe the authors will make a better case for the limited utility of first and second-order moments.

R: We agree. In the revised introduction, we adjusted the language as is suggested and highlighted our interest in higher-order shifts in the distributional properties beyond the location. Also, we provided some illustrative examples about the utility of higher order moments.

4. The use of statistical terms and the discussion of statistical techniques could be improved elsewhere in the manuscript as well. By doing so, I think that the authors will make a more clear statement of their findings and avoid the pitfalls of loose language. For example, on page 4730, line 22, the authors refer to 'most traditional statistics' as relying on parametric assumptions of 'variability [which I read to mean distributional, or parametric, assumptions]. I think, as written, this statement is too broad and unsubstantiated. The term 'statistic' typically refers to a particular number or metric derived from data. Based on the next sentence, it seems that the authors mean to refer to statistical methods, not particular statistics. The distinction between statistics and statistical methods is important, the latter being potentially subject to parametric assumptions while the former are typically not subject to such assumptions (e.g. the mean can be calculated for any dataset, but interpretations and testing of the mean relies of parametric assumptions). Even if referring only to statistical methods, I would argue that the statement is still too broad: there exist whole field of non-parametric statistics. The authors appear to recognize this between pages 4730 and 4731, but dismiss this field and the use of non-parametric transformations as removing 'variability'. Certainly the authors would agree that ranks and transformations still possess distributional properties that can be assessed and, in some cases, translated to statements about the original data (e.g. if the logarithms of a variable are normal, the original variable is lognormal). It might be useful for the authors to point out particular methods or classes of methods that they are trying to improve, citing examples of usage in the literature.

R: We agree. In the revised introduction, we addressed all of these concerns.

5. In addition to precise terminology, I would like the authors to reconsider their application of the statistical tests presented on pages 4735 and 4736. Though I was not able to check that exact citation in Cramer (1998), from my reading of sampling properties in other statistical texts, the estimators of the standard error of skewness and kurtosis rely on parametric assumptions of normality. That is to say that the calculations of 'SES' and 'SEK' are only valid when a sample of size n is drawn from a normal population. I would ask that the authors provide further support for the use of these tests by showing normality or presenting evidence that this parametric assumption is not necessary. As it is, this seems to contradict the authors previous decrying of parametric methods (P:4730, L:22). Furthermore, the discussion of the null hypotheses and type-I errors could be improved. The null hypotheses should be that the skewness is zero and that the kurtosis is zero, respectively. With regard to type-I errors, the authors present a two-tailed alternative at a significance level of 0.05 and then draw one-tailed conclusions. In reality they are then conducting two 0.025-level tests, meaning that their conclusions are stronger than they state. Similarly, the authors conducted these tests on concurrent decades. Typically tests are applied on independent datasets, but I wonder if temporal correlation may affect the power of these tests. It may be worthwhile for the authors to consider the true type-I error in the presence of temporal correlation.

R: This is good point and we agree with the reviewer on the concerns about the use of this statistical test. In fact, Joanes and Gill (1998) reported that sample skewness is an unbiased estimator of population skewness for normal distributions, but not for other distributions. Thus, we have decided to remove the statistical test for skewness and kurtosis from our analysis as well as we revised and rephrased our results and conclusions accordingly (we found no major changes to our findings). In the new version of the MS, we only maintained the criteria proposed by Bulmer (1979) to define the status of the skewness. To define the status of kurtosis, we adopted a similar criteria, we added "…We used similar procedures to define the status of excess kurtosis. We defined five categories that included "negative kurtosis or platykurtic" (if kurtosis was < -1), "moderately platykurtic" (if kurtosis was between -0.5 and -1), "positive kurtosis or leptokurtic" (if kurtosis was > 1), "moderately leptokurtic" (if kurtosis was between 0.5 and 1). Finally, if kurtosis was between -0.5 and 0.5, we considered the distribution as "mesokurtic"…". As suggested, we discussed the potential for serial correlation in a new paragraph in the results/discussion section.

Reference

Bulmer MG (1979) Principles of Statistics. Dover Publications Inc., New York.
Joanes DN, Gill CA (1998) Comparing measures of sample skewness and kurtosis. Journal of the Royal Statistical Society (Series D): The Statistician, 47, 183-189.

6. With the above improvements to terminology and methodology, I would like the authors to consider the broader issues of temporal trends in statistical moments. The skewness and kurtosis, in a formal sense, are moment ratios, not raw moments. As such, the skewness and kurtosis are functions of the lower-order moments. Similarly, the variance, a centralized moment, is a

function of the first moment. This functional dependence means that changes or temporal trends in higher-order moment ratios or centralized moments may be the result of changes in only the lower-order moments or higher-order moments. For example, a trend in the skewness could be attributed solely to a trend in the mean, with the second and third raw moments remaining unchanged, or attributed to a change in the raw third moment. Because skewness and kurtosis, as opposed to the third and fourth raw moments, might be more environmentally relevant, this concern may not be huge problem. Still, I think it is important to note that trends in these moment products are inter-related, making it difficult to attribute changes to any single driver; a comment to this effect might prove valuable.

R: This is an interesting point. We added a note about this in the results/discussion section.

7. The authors present a technique for identifying anomalous years in a time series. While this is indeed a useful technique, I am concerned that it only identifies years in the tails of the distribution rather than true outliers. Merely being in the tail of a distribution does not make a point an outlier. I feel that the term outlier may inadvertently connote an erroneous or otherwise concerning value. As executed, the method presented identifies points beyond the 95% region as outliers. We would, by definition of the 95% region, expect to find 5% of the observations outside of this region due to pure chance. Looking at the figures and tables in the manuscript and supplement, I do not see strong evidence that significantly more or less than 5% of the years at a given site, on average, were outside this region. For this reason, I think it is important to note on that the years identified are the most extreme, not necessarily 'outliers'. Additional tests would be needed in order to identify outliers as such. In large part, this is an issue of terminology, but I consider it an important distinction. With this distinction, I think that the authors have not shown that more or fewer points are 'outliers' in the regulated and unregulated sites. More interesting, to my mind, is the irregularity of the region at unregulated sites, when compared to the regulated sites. The authors make this last observation (P:4739, L:27-29), but I would love to read more discussion.

R: We agree that this is an issue of using the adequate terminology. In the revised MS, we changed "outlier" to "anomalous" throughout the text. We clarified that an anomalous year was classified based on existing data. This point is mentioned in a new paragraph in the results/discussion section. In addition, we discuss the potential use of the geometry of the regions and their relationship to the level of "stress". Lastly, to clarify the visual interpretation of the figures we added a table in the Supplement that identified each anomalous year at the 90% and 95% CI.

8. I think that this manuscript could benefit from a more thorough introduction and a more concrete discussion. While useful and well-written, I felt that the introduction left me with lingering questions supporting the justification and applicability of this work. I found myself struggling with some key questions: By providing clear answers, I think the manuscript would motivate the work more strongly and imperatively. What is the exact goal of this work? To identify change in the distribution not captured by trends in the first and second moment. What is

this work important? Environmental changes may affect the distribution beyond the first and second moment. Ecosystems and organisms are sensitive to such distributional changes. How is this work different from previous works? The authors made this point clearly, but did not demonstrate a marked improvement: This work looked at higher-order moments. By providing a clear, well referenced discussion of these points, I believe the manuscript will provide a much stronger case.

R: We agree. In our revised introduction we added a final paragraph that included the suggestions from this reviewer.

9. In the Results and Discussion section, I feel that some of the conclusions are only loosely substantiated. The authors may be served by expanding the Results section and relabeling it a Results and Discussion and relabeling the Discussion as a Summary and Conclusions. As it stands, the Discussion makes a series of claims that I am concerned with. In the first paragraph, and throughout the manuscript, the authors argue that they have showed higher moments to be an alternative to lower-moment analysis. I think this reads too much like lower-moment analysis should be rejected in favor of higher-moment analysis. Because no comparative analysis was presented, showing that higher-moment analysis identified trends were lower-moment analysis did not, I do not feel that this claim is substantiated. Unless this comparative analysis is presented, I think it important to claim instead, as the authors imply on page 4741, line 18, that higher-moment analysis provides only complementary information. In the second paragraph, the authors state that the outlier detection technique presents a 'more complete and realistic view' of the data. Against what is this comparative statement made? The authors then argue the distributional analysis is more appropriate than single-metric analysis. I agree with the latter, but do not see how it is directly related to outlier detection. Finally, on page 4741, lines 21 through 24, the authors claim to have shown that water regulation masks climate influence. I believe this, but I do not think this is substantially supported by the results. The authors showed differences in temporal trends across regulated and unregulated sites, but I do not think a definitive conclusion is justified. It might be better to include this thought only as a discussion point or conjecture.

R: We agree and in the revised MS we clarified and modified our statements accordingly. First, we highlight the importance of our approach as a complementary rather than alternative analysis that is able to capture long-term changes in empirical distributions. Second, we clarified the utility of the outlier detection technique as a complete and realistic view that contrasts environmental regimes across individual years. Third, we moved the statements about the impact of water regulation to the results/discussion section.

10. I would like to thank the authors for their work. I believe that this is a very interesting study. The call for distributional analysis of environmental variables will have a significant effect on how we, as a field, consider stationarity. I commend the authors on this excellent first effort. I hope that my thoughts are helpful in improving the project. The authors' writing was well-executed, presenting a concise treatment of their work. In addressing my comments, I hope that the authors will retain this style. I look forward to the next iteration of this manuscript.

R: We thank this reviewer for a constructive review of our MS.

11. Technical comments: (I have tried not to tread the same ground as the previous reviewers.)
P:4731, L:9: The authors introduce the concept of 'regimes'. I am not sure what is meant by these, as it seems to be equivalent to distributions of the environmental variable. I do not think it is necessary to introduce the idea of regimes. This idea does not seem relevant later in the manuscript.

R: We considered the suggestion of this reviewer. In the revised introduction, we eliminated the excessive importance of the concept of 'regimes' from the text.

12. P:4731, L:9-11: This sentence starts with the word 'Typically' and concludes that these 'typical' methods are not used in practice. This strikes me as inconsistent: if they are not used, are they typical?
P:4731, L:21: Change 'in stream temperature' to 'in the distribution of stream temperature'
P:4732, L:14-17: The sentence starting with 'For example…' is a bit awkward. The statement is not clear. I think that the readability could be improved by revising or adding something like 'by' in '…captured [by] only using…'

R: We modified the text accordingly**.**

13. P:4734, L:1-8: Why is this introduced? The authors state that this transformation is used to compare values across sites, but I do not think that this comparison is ever presented.

R: We clarify in the revised MS that this transformation was intended to visualize and scale across sites rather than to statistically compare among them.

14. P:4734, L:11: Change 'changes in environmental variables' to 'changes in the distribution of environmental variables'.
P:4736, L:20-23: This sentence could use revision. I think that the comma splice midway through makes the sentence seem too general. It reads almost like a result or discussion point, though this is in the methods section.
P:4738, L:2: Change 'Stream temperature empirical distributions' to 'Empirical distributions of stream temperature'
P:4740, L:26: Change 'here take advantage' to 'here takes advantage'
P:4741, L:15: To whom is 'their' referring? The antecedent is vague, not agreeing in quantity with the opening clause.

R: We modified the text accordingly.