Review of “Coherence of Global Hydroclimate Classification Systems” by Pisarello and Jawitz

Summary

The authors collect four existing climate classifications and create four new ones. The four new classification schemes are all some variety of dividing regions into groups with low internal variability of actual evaporation (ET) or precipitation and potential evapotranspiration (P & PET) rates. The eight classification schemes are then evaluated on their ability to produce coherent (little variability within each climatic group) and spatially non-complex hydroclimatic groups. Hydroclimatic coherence is assessed on ET, P and PET data (which were also used to create the four new classification methods), Δθ values (a measure of temporal alignment of P and PET seasonality), and modeled Q data. Spatial complexity is defined as the number of groups in the classification, if these groups are of equal size and if they are connected in space. The authors find that one of their own classification schemes (called WEC) performs best on these criteria and recommend it to others.

I have read this paper with interest and I think (hydrologically-informed) climate classifications can and should be further developed. I have however various serious concerns about the experimental design the authors present in this manuscript and about the presentation of the work in general. Briefly, I think the authors can do a much better job in explaining how their preferred climate classification scheme was developed. More importantly, I believe that their experimental design makes it a foregone conclusion that the proposed WEC scheme beats the other classification approaches, because the WEC setup uses the same data as are used for WEC evaluation. The manuscript also lacks in clarity. Especially the setup of WEC is difficult to understand. I have outlined these concerns in more detail below.

Major

I here summarize my main concerns about the experimental design. Line-by-line comments are provided below. There is a certain amount of duplication between this section and the line-by-line comments.

1. <u>It is unclear how ETA, ETV and ETC are different in concept</u>
   The authors propose three mono-variable clustering approaches, based on ET. To me, these seem like the same approach with only very minor differences in underlying details. I recommended that the authors make clearer why these minor differences are enough to treat these three schemes as completely independent. More details in line-by-line comments below.

2. <u>The WEC description is too brief</u>
   I struggled to understand exactly how WEC is set up and am left with many questions after reading the Supporting Information to this paper. I strongly encourage the authors to provide more specifics about their methods, both to let the reader better understand the results presented in this paper and to allow the reader to reproduce the WEC results if they wish. More details in line-by-line comments below.

3. <u>Evaluation data is not independent from the data used to set up the four new schemes</u>
   The authors use data from the CRU TS data set (P and PET) and simulations from the TerraClimate data set (ET) to setup the four new classification schemes. This same data (P, PET

and ET) is also used for 3 out of the 5 climate coherence evaluation criteria, where all eight schemes are compared on their ability to group this data into groups with little within-group variability. Unsurprisingly, the four new schemes that are all explicitly conditioned to create such groups with that particular data do very well in this assessment. This comparison is meaningless and should be removed from the manuscript. More details in line-by-line comments below.

TerraClimate simulations of streamflow (Q) are also used to evaluate the eight classification schemes. These Q simulations are the result of forcing a very simply hydrologic model with P and PET data which, critically, are in part obtained from the same CRU TS dataset that is used to create the four new classification schemes. The TerraClimate ET and Q data can not be seen as independent of the data used to setup the WEC scheme and this undermines the conclusion that WEC is the superior climate classification scheme. If the authors are seeking to thoroughly evaluate the capability of their cluster-based classification, independent evaluation data is needed. This might be obtained from observations of streamflow (instead of model results, the global GSIM database might be of use) or actual evaporation rates generated by modern land-surface schemes. The simple model that underlies the TerraClimate ET and Q values has been first used in 1948 and hardly modified since, and I expect that by now something better may be available. More details in line-by-line comments below.

4. <u>The spatial complexity criteria are not well justified and possibly need changing or removal</u>
   The authors define criteria that reward classification schemes that return large, connected zones with a single internal climate and uniform sizes. However, the main aim of climate classification schemes is to find locations that have similar climates, not necessarily locations that are spatially connected to one another. Neither do I see much reason to assume that all climate groups must be of equal size. These must be better justified or removed. More details in line-by-line comments below.

5. <u>KHC conversion to discrete categories is too simplistic</u>
   The authors convert the continuous climate classification by Knoben et al. (2018) into a discrete climate classification for comparison purposes. Rather than using a clustering approach as is done in Knoben et al. (2018), the authors instead “round pixel values until 30 climate classes are created”. This sounds somewhat simplistic to me and I recommend instead that the authors use some form of clustering to overlay separate climate classes over the continuous hydroclimatic spectrum (KHC). I suspect that this rounding procedure is one of the main causes why KHC shows very high numbers of patches in this study and why Koppen-Geiger (KPG) and KHC appear equally coherent on the Q variable in this study, whereas KHC is substantially better than KPG for finding similar Q regimes according to the assessment in Knoben et al. (2018). More details in line-by-line comments below.

6. <u>The discussion is somewhat limited</u>
   I think the discussion would be stronger if the authors address the question “what did we learn about the world?” from their analysis. More details in line-by-line comments below.

Line-by-line comments

L52. It is equally possible that actual ET rates are hardly used in classification schemes because observations of actual ET are hard to come by at the spatial scales where climate classification is typically useful. Global products of actual ET are typically the result of model simulations or derived from model simulations across this domain. Depending on the origin of the actual ET data, one could argue that a classification that uses this kind of data is more of a modeled-climate classification and is critically reliant on the accuracy of the model simulations with respect to actual (meaning occurring in reality) ET.

L82. The provided reference seems to say that TerraClimate data are at a monthly resolution. If this is the case, how was this disaggregated into daily data?

L89. KPG has not been defined yet.

L95. Should “monthly mean” be “annual mean”?

L101. Mentioning just the mean R^2 values and their standard deviations seems quite a short description of the accuracy of these fits over 61000+ grid cells on the planet. A somewhat longer description is appropriate. Do the authors’ results for P match those of Berghuijs & Woods (2016) in terms of accuracy and spatial patterns of high/low accurate fits?

In which regions are the authors' fits more/less accurate?

L107. “veteran” seems an odd choice of words. Maybe “legacy”?

L111. It is not entirely clear to me how pixel values can be rounded into creating a distinct number of categories. Clustering of the KHC climate index values (similar to how WEC is created) would be much more appropriate and is also the approach taken in Knoben et al. (2018) to create discrete categories that overlay their hydroclimatic continuum. See Figure 3 in Knoben et al. (2018). Can the authors clarify their rounding approach and justify why they use this over a clustering approach that uses the actual data of the KHC scheme?

L122. I don’t quite follow the arguments presented here that should support the idea of a classification that has an equal number of pixels in each zone and I think these arguments need to be clarified or changed. See my concerns with each individual argument below.

#1. Koppen-Geiger (K-G) has “relatively high spatial non-uniformity […] resulting in highly variable relevance for regional analysis.” I interpret this as meaning that K-G has climate classes of non-uniform size and that this is not useful if one’s region of interest falls entirely within a single K-G class out of the ~30 or so possible K-G classes. I don’t quite understand how dividing the entire globe into only 15 classes addresses this problem, as using 15 classes compared to 30 necessarily means lower granularity in the authors’ ETA scheme than is possible with K-G. Additionally, I see no reason to assume that forcing each of the 15 ETA classes to have an equal number of pixels will necessarily mean that “regional” analyses have more useful climatic information available. If we consider 5 continents globally, this scheme roughly divides each continent into 3 classes. Does this really help regional studies?

#2. "Additionally, it is useful to have a simple baseline framework upon which to compare other systems." I agree in principle, but I'd argue that even a simple baseline needs to be somewhat plausible. Dividing a map into 15 regions of equal size, based on a single variable seems a pretty low threshold to beat. Why not use an existing classification as a benchmark?

#3. "Zones should ideally be delineated in one piece." This seems somewhat counter-intuitive to me, because the express purpose of climate classification schemes is to identify regions that are similar in terms of variables X, Y and Z. Whether such regions are spatially connected is irrelevant. This also directly c argument #1, where having large areas of a single climate class is mentioned as a negative aspect of the K-G scheme.

L124. It is unclear to me how the proposed ETA scheme should be interpreted in a temporal sense. Given that the number of pixels in each class should be the same, this means that the resulting 15 classes are only valid for a given snapshot in time. If the chosen time period for this classification changes, the underlying ET data would change, and wouldn't therefore the number of pixels in each class change too? If we do not want to violate the "equal number of pixels in each class" concept, this means that classes need to be redefined when the underlying time period changes and thus the classes do not have a consistent meaning for different time periods. Imagine a theoretical case where PET uniformly increases over the planet with a constant value. The lowest ETA class now corresponds to a very different real-world climate than it did before.

L124. Using number of pixels is not necessarily a way to guarantee a relatively even distribution of zones in terms of area (which the name "ET Area-optimizing hints at). Based on the CRU data, each pixel represents a certain area on a regular latitude/longitude grid. Pixels in such a grid do not translate easily into real-world areas. A single 0.5x0.5 degree pixel at the equator might be approximately 50 km^2, while the same pixel size near the poles would represent a fraction of that area.

Can the authors clarify why having even distributions in the number of pixels is desirable, even if this could potentially lead to a very uneven distributions of zone size in km^2?

L133. I don't quite understand what makes ETV different from ETA. ETA imposes groups on the empirical ET CDF. As a result, each ETA group consists of regions with similar ET values (i.e. low CV within the group).

ETV seems to minimize within-group CV of ET values by imposing groups on a normal distribution fitted to the empirical CDF. It seems to me that the only difference between ETA and ETV is that ETA uses the simulated ET values (from TerraClimate) directly, whereas ETV uses an approximation of these ET values.

ETA with 10 groups has CV = 0.2 (Fig S1B); ETV with 10 groups has CV = 0.2 (Fig S2). What does the fitted normal distribution add to this analysis that makes ETV with 29 groups (as determined on line 144) substantially different from ETA with 29 groups?

L136. Is "S1" the correct cross reference? I don't see a fitted cont. uniform distribution in Fig S1.

L144. Why did the authors choose to use 29 zones? To me it currently sounds that to maximize within-zone ET coherence, one would simply keep imposing more groups until each zone contains a single pixel and within-zone ET CV equals 0.

L145. I'm again a bit confused about the difference between this approach and the preceding ones. Constructing an (empirical) CDF already puts locations with similar mean ET values close to one another, which is also what the clustering in ETC tries to achieve.

Additionally, because global mean ET values are approximately continuous (as evident from Fig S1B, S2 and S3), K-means will be trying to impose distinct boundaries on continuous data and therefore tend to gravitate towards clusters of approximately equal size. The authors have already defined ET zones of equal size with minimal within-zone ET variability in their ETA approach. So what does using a clustering algorithm add?

Equally, comparing Figure S2 and S3 seems to show that ETV and ETC generate approximately the same CV for the same numbers of clusters/groups, but with some scatter in the ETC values (potentially caused by the initial guesses for cluster centroids, see comment below).

L146. K-means clustering is quite dependent on the initial guess of cluster centroids for the location of the final clusters. A multi-start framework shows to what extent this vulnerability influences the final clusters. Was the K-means algorithm used in a multi-start framework? If not, why not?

L150. Is there any particular reason why CV = 0.1 makes a good threshold?

L151. Should “systems” be “system”?

L162. I find the description of this new classification scheme in the SI too brief to understand in detail what's going on. I gather that the authors used K-means clustering on various combinations of data but the rest of the method escapes me. This must be addressed, because it (1) makes the authors' claims that WEC is the best out of the 8 classification schemes difficult to assess; and (2) makes the classification impenetrable to others who might wish to reproduce or use this work.

Some of the questions I currently have upon reading the SI:

1. How were the combinations of P, PET, ET, Q and delta theta determined? I notice that not all possible combinations are present in Table S1.

2. What were the K-means settings? Was the algorithm restarted multiple times to test cluster stability?

3. How were the thresholds for coherence chosen (SI, page 3)?

4. What does (1.50) refer to in “KPG(1.50)”? Is this the 50% deviation mentioned in the main text? If so, this should read “KPV value * 1.5”. Also, the use of “=” signs is extremely confusing in the lists on this page and should be removed.

5. Why does the reader need to know the KPG coherence scores at this point in the analysis?

6. Is "number of parameters" equivalent to number of K-means clusters? If not, which K-means parameters are meant?

7. If number of parameters is not the same as number of clusters, how was the appropriate number of clusters chosen?

8. What does "number of patches" refer to? How was its threshold determined?

9. I don't understand how/why the P,PET clustering system was chosen out of all possible options, nor why 22 zones are considered appropriate (SI, page 3).

10. The caption of Table S1 states that cells with a "1" in them indicate a combination of variables and number of clusters that meet the criteria specified on page 3 of the SI. From Table S2 however, it seems that neither ET,PET not P,PET meet the "number of patches" criterion. Why are these then shown in table 1 as if they do meet all criteria?

11. Similarly, neither ET,PET nor P,PET meet the CV(ET) criterion (all values are > 0.33), and all but one fail the CV(Q) criterion (only P,PET with 22 zones has CV(Q) < 1.31). Why does table S1 show these results as meeting the criteria? Why define criteria at all if they are not used?

12. I don't understand why there are gaps in Table S1 between certain rows with "1" in them. For example, column P,PET. If 22 clusters are sufficient to meet all criteria (indicated by a "1"), and all criteria are aimed at minimizing differences within a single cluster, it is logically impossible that using 23-27 clusters gives worse results than using 22 clusters, especially considering that 28 and 29 clusters suddenly do meet all criteria again. The only explanation I can think of is that certain settings in the K-means algorithm prevent it from finding the most optimal cluster configuration when 23-27 clusters are used. The multi-start issue I have mentioned before is a possible culprit.

13. Why were ET,PET and P,ET chosen for further analysis in Table S2 and not others?

L158. "Zone complexity" is undefined up until this point and the reader can only guess at what this means by reading the SI. I suggest to clarify what is meant by this in the main text and to also explain whether it is low or high zone complexity that is desirable.

L160. The SI could use a header to indicate where this section starts.

L164. This section seems to be the justification for many of the authors' methodological choices, in particular about their selection of classification schemes. I suggest to move section 2.6 to the beginning of section 2, so that reader already has access to this information before it is needed to understand the authors' methodology.

L168. The authors argue that "classification systems should consist of a relatively even distribution of pixels across zones, avoiding disproportionally large or small zones" (similar to an earlier argument in section 2.4). I don't understand this argument for two separate reasons:

1. I don't think there is much reason to assume that climatic zones should follow an even distribution across the planet, either in terms of pixels or in terms of area. Globally, deserts are big and alpine regions are relatively small. A classification scheme that tries to create climatic zones of equal (pixel) size will not capture either climate properly, and thus offer little in terms of hydroclimatically relevant information.

2. Like I argued before, I don't know if number of pixels is an appropriate unit here. KPG polar zone ET might take up a fair number of pixels on a regular grid, but in terms of total area the arid classes dominate (compare Sahara size is ~9.2 million km^2, whereas Greenland is ~ 2.2 km^2). I don't think pixels are a particularly helpful unit for this analysis.

I recommend to remove this criterion from the analysis.

L169. The authors argue that "Zones should be as hydrologically continuous as possible (Meybeck et al., 2013), minimizing patchiness or fragmentation". Like I argued earlier, this is counter intuitive to me. Climate classification systems are intended to find places that are similar climatically, regardless of physical distance. By penalizing systems for patchiness, the authors effectively favor classification schemes that generate large areas of single climate zones, without providing any justification that such schemes are more representative of the real world. In effect, the less data a scheme uses, the more likely it is to generate large connected areas of climate zones and thus, according to this criterion, the better this scheme is. This seems extremely counterintuitive to me and I recommend to remove this criterion and to re-do the analysis.

L185. Like before, I'm really not sure why different numbers of pixels contained in different climate classes is a bad thing. If different climate types cover differently sized areas on a map, than that's simply how the global hydroclimate is. Re-drawing the climate class boundaries to create equally large zones is not adding any new hydro-climatic insight to the problem (in fact, one might argue that such an approach uses *less* hydro-climatic insight).

L204. Seeing the authors' assessment scheme indicates some serious methodological concerns, centering around the fact that no real independent evaluation data has been used.

1. Classification schemes are evaluated on their ability to create low within-zone variability of ET values (i.e. low CV(ET)). This is the exact same data that has been used to condition the authors' ETA and ETC schemes on and, unsurprisingly, when one specifically sets out to create groups with as low as possible within-group ET variation and then uses that same data to see how well that worked, these schemes are impossible to beat. In my opinion the CV(ET) comparison is meaningless because it cannot reasonably be expected that any scheme beats the ETA and ETC schemes in this comparison.

2. The same argument can be applied to the CV(P) and CV(PET) criteria. The authors' WEC scheme is specifically conditioned on creating groups with low variability in these two climate indicators and therefore the comparison with existing classification schemes is meaningless. The only question this seems to answer is "are established climate classification schemes better at clustering P and PET values then a clustering algorithm can cluster P and PET values?" Apparently and not entirely unexpectedly, they are not.

3. The ET and Q data in this study are taken from the TerraClimate dataset (Abatzoglou et al., 2018). To quote directly from Abatzoglou et al. (2018):

   "A one-dimensional modified Thornthwaite-Mather climatic water-balance model (WBM)[22,31] was used to calculate monthly water balance from 1958–2015. The WBM is a single bucket model applied consistently across global land surfaces that operates on a

monthly time step and considers the interplay between precipitation, ET_0, as well as soil and snowpack water storage. The WBM accounting scheme considers runoff as the excess of liquid water supply (precipitation and snowmelt) used by monthly ET_0 and soil moisture recharge. Soil water is extracted during months where ET_0 exceeds liquid water supply, with the extraction efficiency of soil water declining exponentially with the ratio of soil water to extractable soil water capacity. Under such conditions, actual evapotranspiration is counted as the liquid water supply plus the soil water utilized and climatic water deficit is the difference between ET_0 and actual evapotranspiration."

Due to its simple design and monthly time step, this model has little to no capacity to generate non-linear (and thus realistic) hydrologic behavior. It is thus likely that these simulated Q and ET data are strongly correlated to the forcing data used to generate them. TerraClimate uses the CRU TS v4.0 as one of its inputs, while CRU TS v4.0.4 is used by the authors to provide the P and PET data for their classification. WEC is thus conditioned on a dataset that is very similar to the dataset used to generate the ET and Q data that are used to evaluate the different classification schemes in this paper. It is therefore not entirely surprising to see WEC perform (reasonably) well on the CV(ET) and CV(Q) criteria. In a comparison such as this, independence of the evaluation data is critical to guarantee a fair comparison. I suggest to replace the evaluation data for Q with observations from for example the global streamflow attributes dataset GSIM, or those from the Global Runoff Data Centre. ET might be obtained from modern land surface models run at a global scale, instead of from a bucket model first conceived of in 1948 and hardly changed since.

4. As argued before, neither having equal numbers of pixels in each zone nor having a low number of patches seem particularly indicative of a useful climate classification scheme to me. I suggest to remove these if they cannot be better justified.

5. The fact that ETA has "the fewest number of zones and very high coherence for zone size (by far the lowest CVz)" (L200) should not come as a surprise when the entire premise that underlies ETA is dividing the available number of pixels in a small number of zones of equal sizes.

6. It is odd to see that ETA has the best score for "zones" (indicated in bold), because it has the lowest number of zones. This is counterintuitive to me, because this seems to favor climate classification schemes that are explicitly not good at their intended purpose, namely to define different climatic regions. The ideal score on this assessment would be obtained by a scheme that returns 1 climate zone, which means that all climates are classified as being the same. I recommend to remove this criterion.

7. Concluding, I believe that the evaluation criteria heavily favor the authors' classification schemes and that the fact that WEC performs well on them is no more than an artifact of the experimental design. The evaluation data cannot be considered independent from the data used to generate WEC (and ETA, ETV, and ETC) and the pixel, patches and zones criteria are not well-justified. This undermines the authors' conclusion that WEC is a better

classification scheme than Koppen-Geiger or any of the other existing classification approaches. Independent evaluation data is needed.

L205. The caption mentions that certain schemes have statistically similar results. How was this determined?

L205. I find the number of patches for KHC extremely high. Can this be a consequence of the "pixel-rounding" procedure described on line 111? This result indicates that this rounding procedure is unnecessarily crude, because the KHC groups shown in Knoben et al. (2018) are not nearly as fractured as this. Given the authors' use of actual clustering algorithms in this paper, applying those to the KHC data seems feasible.

L214. This grouping procedure needs to be explained in more detail than is currently done. Is this the result of another k-means clustering exercise? Is phi_bar the mean aridity over the group?

L216. Where are these relationships shown? How do these compare to those from Koppen-Geiger classes?

L217. "… indicating higher spatial variability (lower coherence) with increasing aridity." I'm not entirely sure how to interpret this statement. Which conclusion should the reader draw here?

L226. This discussion seems to imply that WEC provides higher granularity in certain areas. I think this argument can be reversed too: whereas KPG shows some longitudinal diversity in climate zones, WEC lumps London, Amsterdam, Berlin, Warsaw, Moscow and a substantial part of the Russian Federation into a single climate zone. I think a more balanced discussion of WEC is needed.

L235. "Discussion and conclusions"

L238. This statement is a bit inaccurate, because Knoben et al. (2018) use the monthly ratio between P and PET as one of their predictors.

L243. It needs to be clarified what is meant by "parameters" here. This seems to say that the parameters used by WEC are mean P and mean ET. What about the number of clusters used and other k-means settings?

L243. "a notable aspect of system complexity" This argument is unclear. Why is the number of parameters relevant in a climate classification scheme? With the definition of "parameters" as "number of climatic variables used" (as I gather from this section), this sentence seems to imply that the less data used, the better the classification scheme is. Can this be clarified?

L245. It should be no surprise that WEC, as the result of a k-means clustering algorithm applied to P and PET data, consists of groups that have low internal variability on that same P and PET data. Also, Table 1 seems to say that KPG scores better on CV(ET) and CV($\Delta\theta$) while WEC scores higher on CV(Q) only, albeit with a higher standard deviation than KPG gets.

L250. I find this odd. As the authors say (lines 48-49), Knoben et al. (2018) compare their classification scheme to KPG and find that KHC is better for hydrologic purposes (in this case for grouping locations with similar hydrologic regimes expressed through a variety of metrics). In the authors' words, KHC is "more coherent with respect to Q" then KPG. Why does this not show in the

authors' assessment? This could be related to "pixel-rounding" or to the use of simulated Q in this study.

L250. Additionally, if KHC is not high in Q coherence, why is it indicated as one of the three best systems in Table 2?

L252. I suggest to remove these statements because they are trivial. If one clusters ET data one gets groups with high ET coherence. If one divides all data into 15 equal-sized groups, one gets groups with equal sizes. This should not need to be presented as a benefit of these approaches.

L257. Which variables do the authors mean when they refer to "water budget dynamics"?

L265. I don't quite follow this sentence, perhaps because I don't understand which variables are meant with "water budget components".

L283. Given all the methodological concerns that need to be addressed, I do not think this conclusion is currently well supported.

L289. This seems overly optimistic to me. Which kind of environmental management decisions can be made at a scale where a handful of Koppen-Geiger classes or WEC groups provide a useful indication of local conditions?

L289. What I miss in this discussion is a critical assessment of hydroclimatic understanding and how the authors' proposed WEC adds to this. Existing classification schemes are based on hypotheses about how the world works and about which elements of the global climate are first-order controls on the resulting hydroclimate. WEC is simply a clustering method that finds regions with similar P and ET values. What does this teach us about global hydroclimatic relationships? If WEC is better than established methods that do rely on theory, then where is this theory faulty or incomplete? In other words, I think this discussion would be much stronger if the authors where to consider the question "what did we learn about the world?"