Manuscript title: Coherence of Global Hydroclimate Classification Systems

Manuscript number: hess-2020-522

Reviewer 2

I find the provided responses to review comments insufficient in several places. I had outlined six major comments in my previous review. Below are brief descriptions of which ones I think deserve further consideration.

---- The lack of independent evaluation data is a critical flaw of this study.

The discussion now includes several sentences that mention that the proposed new classification systems are evaluated on data that is either the same or dependent on the calibration data, which in my opinion is insufficient to address this concern. I consider the lack of independent evaluation data not a limitation but a methodological flaw that needs to be addressed before this paper can be published. Without independent evaluation data, no honest comparison between the established and new schemes is possible.

> Response: We thank the reviewer for the emphasis to incorporate independent validation data. Please see lines 111-119 for these updates, also shown below:
>
> "Additional ET and Q datasets were used for independent validation purposes. Observation-based monthly Q from 1980-2014 were obtained at 0.5° x 0.5° resolution from monthly global gridded runoff data (GRUN, Ghiggi et al., 2019). The Global Lobal Evaporation Amsterdam Model (GLEAM) produced terrestrial daily ET for 1980-2020 at 0.25° x 0.25° resolution, which was also resampled to 0.5° x 0.5° resolution. Here, we used the updated GLEAM version 3.5a, which is based on ERA5 net radiation (satellite) and air temperature (reanalysis) datasets, downloadable at a monthly timestep (Martens et al., 2017). The GLEAM ET and GRUN Q datasets were independent from TerraClimate ET and Q datasets both temporally (Figures S1 and S3) and spatially (Figures S2 and S4). The two ET datasets were more similar than the two Q datasets, based on monthly linear models (R^2 ranging from 0.78 to 0.87 for ET and 0.47 and 0.84 for Q), and both ET and Q datasets showed spatially consistent seasonal differences."
>
> Results from these data can be seen in Figures 2, S1-S4, and S11.

---- The complexity criteria are still insufficiently supported and clear:

1a. Hydroclimatic zones do not have equal sizes in reality. The justification that zones must have equal sizes for visibility of the small zones in my opinion simply means that zones get redefined to be less hydroclimatically distinct. In other words, regions that could be distinct hydroclimates get merged to conform to the minimum zone size criterion, meaning we lose hydroclimatic insight as a consequence of the zone size requirement. As far as I can tell, this argument is not addressed in either the response or the updated manuscript.

> Response: We agree with the reviewer that the purpose of delineating hydroclimate boundaries is to identify distinct regions that share hydroclimatic similarities, which is not necessarily size-dependent. The justification for ensuring that zone sizes are not meaninglessly small is rooted in disproportionality (line 95). For example, two of the KPG zones disappeared when resampled to

0.5° x 0.5° resolution (line 158), which results in highly variable spatial relevance of zones within the context of regional analyses (line 172).

1b. In addition, pixels are an inappropriate unit of measurement for area, because the area a single pixel represents is not constant within the regular lat/lon projection used in this manuscript. This also means that any pixel-based criterion is conditional on the geographical projection used for the source data and thus not on the data itself. This is avoidable by using actual area values.

Response: We agree that zone area is a much better metric, and our methodology has been rectified to include an assessment of zone areas instead of number of pixels per zone. The results still showed high spatial variability for KPG compared to the novel frameworks (Figure 4B).

2. The authors rely on Meybeck et al. (2013) for the statement that zones should be delineated in one piece. To my knowledge the authors address neither in their replies, nor in the main manuscript my concern that this criterion rewards schemes that are not very good at what they are supposed to do, namely define regions with similar hydroclimates regardless of spatial proximity. Meybeck et al. (2013) have clear reasons for wanting to do so (mostly to not separate a river's headwaters and its lowlands) and an experimental design that supports their reasoning (i.e. using basin shapes in addition to climatic data) which do not seem to apply in this paper. Given that the streamflow data used in this manuscript is simply local P – EP and does not consider catchment aggregation at all, I would say that Meybeck cannot be used to support the use of this criterion.

Response: Meybeck et al., 2013 state that "ideally" zones would be "delineated in one piece," though they recognize this is not physically possible. A clause to reflect this limitation was added to line 177, shown below:

"This type of spatial condition is similar to the prioritizations of the MHR framework that state zones should ideally be 'delineated in one piece,' although this is not a physical reality (Meybeck et al., 2013)."

3. I now understand that the authors use the number of zones as part of a trade-off (given equal hydroclimatic coherence, the system with fewer zones is preferable) but I think this is insufficiently clear in the main manuscript. Most importantly, this tradeoff is neither mentioned in the description of the results (for example Table 1 simply lists ETA as having the best – lowest – number of zones despite it not having very high coherence), nor in the discussion.

Response: We agree that this tradeoff assessment was not made abundantly clear. The methodology has been updated to reflect a more systematic determination of what is considered "good" with respect to number of zones. Please see section 2.6 and Figure 2, where a systematic approach for choosing the optimal number of zones is outlined. Information previously gleaned from the table aforementioned by the reviewer has been transformed into Figures 3 and 4 for enhanced clarity.

---- The introduction and discussion are somewhat limited.

Given the focus on annual or longer time scales, the Budyko framework probably needs to be part of the introduction. Although not a climate classification system in the traditional sense, it is a well-established way of organizing locations based on long-term aridity (P / EP) and long-term ET and Q. A discussion of the Budyko framework can replace the current sentence on line 68 ("given the major gap regarding the inclusion of ET in climate classification systems, …") and may provide a justification for assessing the existing climate classification schemes at annual or longer scales. This could introduce a research question along the lines of "do existing hydroclimate classification schemes correspond with catchment organization as predicted by the Budyko curve?" or something similar.

> Response: A statement regarding the Budyko framework was added to the Discussion in lines 334-337, shown below:
>
> "To delineate the landscape based on ET dynamics, the Budyko framework is a longstanding, well-vetted mechanism for estimating the evaporative index (ET/P) using the primary drivers of the water budget, PET and P, as represented by the aridity index (Budyko, 1974; Milly, 1994; Reaver et al., 2020a; Reaver et al., 2020b; Zhang et al., 2004)."

I will also repeat one comment directly from the first review: 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?" I don't see where this is addressed in section 5.

> Response: The reviewer highlights important points regarding impact. Please see lines 363-377, shown below:
>
> "It is widely accepted that water and energy, chiefly in the form of precipitation and solar radiation, govern long term socioecological water availability at large spatiotemporal scales (Budyko, 1974; Berghuijs and Woods, 2016; Knoben et al., 2018; Sanford and Selnick, 2013). Several previous climate classification systems aimed to represent this water-energy interaction within bounded zones that encompass similar hydroclimatic sensitivities (Knoben et al., 2018; Meybeck et al., 2013). It was concluded here that WEC, using water and energy in the form of P and PET rates, was the best overall system for building zones that encompass similar Q rates. This suggests that the WEC scheme is valuable for assessing and predicting water availability changes given changes in water and energy. Therefore, WEC is the most relevant system for direct management understanding and application as it relates to hydroclimate dynamics.
>
> This work is a promising pathway to regionalization within many different biophysical and socioeconomic contexts, clustering drivers to form zones of similar response variable sensitivities in order to more accurately extrapolate locally derived results and regional impacts of local management practices. The WEC framework can thus inform regional to national scale management strategies in the effort to account for potential hydroclimate zone-dependent responses to climate and land cover changes."