Response to Reviewers' Comments

We appreciate the efforts of the reviewers and we thank them for their insightful and constructive comments. We have provided information points and clarifications as to how we will modify the manuscript accordingly below. We provide detailed responses to each of the reviewers' comments. For convenience, we put the reviewer comments in black font, and author responses in blue.

RC1- Anonymous Referee #1 comments:

This paper is overall well-written and aims to bridge an important gap in hydroclimate impact studies. The methodology used however currently suffers from a few major limitations which should be addressed by the authors.

We wish to thank you for your time reviewing our paper. We are confident that a modified version of this paper will address all of your comments in a satisfactory manner.

Key comments

- I wonder about the conservation of the important energy and water balance when combining P and T datasets from different sources - e.g. MSWEP precipitation and ERA5 temperature. Could you please comment on this and potential impacts in the paper?

  This is a very interesting comment, but a complex issue with many different angles. In most cases when doing hydrological modeling, energy and water balance is not taken into account by the driving datasets, so we don't believe that this is a problem specific to our work. Most data used in hydrological modeling comes from gridded information and the gridding process is almost entirely done independently for precipitation and temperature, therefore not even taking into account basic temporal correlations between both variables. Even in hydrological models that use station data, there is usually some level of interpolation/extrapolation to extend the information at the catchment scale. Most precipitation products are now developed independently of temperature, as is the case for most dataset used in this study. Reanalysis are the most consistent dataset with respect to energy budget and water balance. However, even though the weather model of the reanalysis is entirely physically coherent, the data assimilation does not preserve this physical coherency and therefore reanalysis do not fully conserve water balance. And as shown in this study (and others), reanalysis precipitation, although much improved, is not as good as other precipitation datasets. We will definitely discuss these issues in a revised discussion since they are definitely relevant.

- On P10 L280-281 you mention the impacts of the reduction in the ensemble size of the precipitation dataset on the variance analysis: "Unsurprisingly, it shows that reducing the size of

the precipitation ensemble results in a consistent decrease in the variance attributed to precipitation". This leads me to question what the impacts may be of the ensemble sizes from the contributors you investigate on your conclusions: 10 GCMs, 2 hydrological models, 2 temperature and 9 precipitation datasets. Wouldn't it be more adequate to have the same number of ensemble members coming from these various components of the chain? For example, temperature appears to play a minor role in the analysis, however, only 2 members were used here, which could impact this conclusion somewhat. Please reflect on this in your paper.

This is a fair question, and it goes to the heart of the uncertainty issue. In an ideal world, it would be best to have a similar number of ensemble members for each uncertainty component. However, the contribution to variance is related to how dissimilar the ensemble members are and not strictly to its numbers. As such, the ensemble with the fewest members can still provide the largest contribution to variance. If you start with a single member (a single hydrological model for example), each additional hydrological model added to the ensemble will add uncertainty. However, there is a point of diminishing return where adding more hydrology models will not change anything because the added hydrological response will be within an envelope defined by the existing models. So ultimately, we don't need all ensembles to have the same number of members, we need all ensemble to have enough credible members to cover the uncertainty. For example, based on the two temperature datasets used here, adding more temperature datasets is unlikely to change the results considering how little uncertainty is present in the two datasets when compared to other sources. Adding more hydrological models is likely to have a much more important impact. Based on previous published work, 10 GCMs is more than enough to frame the uncertainty contribution from this source. I think expanding this discussion in the revised version of the manuscript would be useful.

- The calibration strategy may be problematic for this study given the climate timescales explored. As mentioned by Arsenault et al. (2018): "In this study, the effect of calibration and validation is investigated on three catchments that did not show signs of non-stationarity, i.e. the mean annual streamflow did not contain a trend over a 25-year period. This allowed randomly sampling from the database to generate calibration and validation sets. This raises the question as to how the method would fare on a catchment that is subject to non-stationarity. Obviously, in this scenario, the independent test period would need to be in the most recent years and those years could not be randomly selected from the entire time series." Arsenault et al. go on to suggest alternative methods which could be used in such catchments. Could you please reflect on the adequacy of this calibration strategy for your study?

This is a good question, which has been debated previously so we completely understand the idea behind the comment. In the methods, we describe that the catchments were calibrated on the full length of the available data. The first reason is to ensure as much information as possible is contained in the parameter set, and the second is to maximize the chances of regionalization methods to perform well for the ungauged catchments. These two objectives can be seen as mostly in the same direction, but there is a twist that makes them contradictory in one sense, and the approach that we implemented is somewhat of a compromise on these issues.

There are two main concepts here that need to be highlighted:

1- In Arsenault et al. (2018), the main conclusion is that the more years are used in calibration, the more the parameter set contains useful information as it can compensate and accommodate for more types of events. In other words, if one were to calibrate on a single, dry year, then the validation would probably be atrocious for a very humid year. But by calibrating on a series of wet and dry years, the parameter set can compromise and be good on the entire period. The same concept is applied here, where we want to keep as much information as possible in the parameter set so as to make any relationships between catchment descriptors and parameters (if any) as robust as possible for regionalization. The same also holds for simulation, in that in the absence of any knowledge a priori of the impacts of climate change, using the entire parameter set is prudent as it protects against highly variable changing conditions in the future. It might be possible to identify non-stationarities and target them specifically, but that would then mean that in regionalization, these parameters would also likely perform much worse due to the limited information contained within. Therefore, it was decided to maximize the information content and ensure that the widest possible "spread" of conditions was included in the calibration data. This was done by taking the entire period.

2- In regionalization, we have the advantage that the "verification" set is actually the pseudo-ungauged catchment itself. Indeed, the "validation" on the donor catchments is not really useful since the score we want to improve is the regionalization skill on an independent catchment, with it's own data coverage period. So in this case the calibration is done on the donor catchments, and the "validation" is done on the pseudo-ungauged basins. Keeping years as "validation" years on the donor catchments (by split-sample calibration or other) is counter-productive because:

   a. There are fewer years to build the parameter-descriptor relationships;
   b. The donor and ungauged catchments have different periods, which means that any difference in period between the gauged and ungauged basins could artificially increase (or decrease) the apparent effect of non-stationarities.

We understand that this was not at all detailed in the previous version of the paper, therefore we will flesh it out more in the next version to ensure that the rationale is well described.

- You mention 51 different streamflow metrics, yet all results are shown for only 6 metrics. Could you please comment on the results from the additional metrics not shows here? These could perhaps go in as supplementary material?

This has been mentioned by one other reviewer. We propose to add a Table for all metrics and adding figures for the other metrics in supplementary material is a good idea. Some of the metrics are relatively similar (distributions quantiles) so we may not need to provide all of them.

- Your figures are very rich in results. Please guide the readers a bit more by mentioning what is shown in the columns/rows, etc. when introducing each figure (especially for Fig 5, 7 and 8).

  Thank you for the comment, we will ensure that the figures are explained in more detail in the revised manuscript.

Minor comments

[Page 1, lines 9-12] Please clarify here if these datasets are deterministic or ensembles.

These datasets are deterministic. This will be specified in the revised version.

[Page 1, line 15] Please explain here what CMIP5 GCMs stands for.

The acronyms (CMIP5 – fifth Coupled Model Intercomparison Project, GCM- General Circulation Model) will be defined in the revised version.

[Page 3, line 60] Please clarify here what GHGES stands for, it only comes later on L66.

Thanks for picking this up. We'll make sure the acronym is defined once it's first introduced in the revised version.

[Page 3, line 63] Could you please give readers a brief explanation of what the "change factor approach for downscaling" is here?

We propose remove the reference to the change factor method and simply mention 'using a single downscaling method' as I don't believe a description of the change factor method would be very useful at this stage of the paper.

[Page 4, line 91] Since you are looking at hydroclimate impacts, it might make sense to also refer to the Hydrological Climate Classification by Knoben et al. (2018), more adapted to hydrological studies: Knoben, W. J., Woods, R. A., & Freer, J. E. (2018). A quantitative hydrological climate classification evaluated with independent streamflow data. Water Resources Research, 54(7), 5088-5109, https://doi.org/10.1029/2018WR022913.

That's a good suggestion that will be implemented in the revised version.

[Page 4, line 107] Please clarify here what NAC2H stands for.

Good point. This should have been done indeed. NAC2H: The North American Climate Change and Hydroclimatology Data Set.

[Page 4, line 112] Please consider replacing "(or better)" with "or higher".

Thanks, we will implement this correction.

[Page 5, line 122] While it is implicit, you do not actually explicitly mention that you have used GRDC data in this paper.

We will correct this in the revised version.

[Page 5, line 136] Could you please provide your reasoning for selecting the L5 vector layer instead of the other 3 shown on Fig. 1?

Good point. This was a subject of discussion early on in the course of this work. Ultimately, it was selected as a compromise between having a sensible number of watersheds and keeping the large computational burden of this project reasonable. This will be specified in the revised version,

[Page 5, line 137] I would find it helpful if you could briefly go over the steps of the entire hydroclimatic modelling chain from Fig. 2 in the text as well. For example, I only found out from Fig. 2 that two calibrations are performed.

We will expand the presentation of Figure 2 in the revised manuscript.

[Page 5, line 146] Please summarise what the "4 groups of components of the uncertainty modeling chain" are here for added clarity for the readers.

We will gladly expand the presentation of these components in the revised version.

[Page 6, lines 150-152] Could you please summarise briefly in the text as well which metric(s) and time-period were used for the calibration? I read the added information later on in Section 3.1.3, please mention here that more details are given in that later section.

Good point. We will do as suggested.

[Page 6, lines 155-158] This is a repetition to an extent of P6 L147-149. Please consider merging these two paragraphs for more clarity. It is also unclear to me how the 1150 African stations (L155) became 1145 catchments?

We clearly missed this in the proofreading stage of the manuscript. We will correct this oversight in the revised manuscript.

[Page 6, lines 162-163] More importantly, have they been shown to perform well in Africa specifically?

While these models have been mostly used outside of Africa, there are some cases where the models have been used over Africa (e.g. Essou and Brissette, 2013; Gosset et al.,2013; Simonneaux et al., 2008). Nonetheless, we strongly believe that a successful application in the same climate zone over another continent is certainly a robust enough justification. We will add the above references to the revised version and expand the justification.

Essou, G. R., & Brissette, F. (2013). Climate change impacts on the Oueme river, Benin, West Africa. Journal of Earth Science & Climatic Change, 4(6), 1.

Gosset, Marielle, et al. "Evaluation of several rainfall products used for hydrological applications over West Africa using two high-resolution gauge networks." Quarterly Journal of the Royal Meteorological Society 139.673 (2013): 923-940.

Simonneaux, V., et al. "Modelling runoff in the Rheraya Catchment (High Atlas, Morocco) using the simple daily model GR4J. Trends over the last decades." 13th IWRA World Water Congress, Montpellier, France. 2008.

[Page 7, lines 201-204] Is this single simulation used as a reference against which to verify the other simulations produced as part of the analysis? Please clarify as it confused me a little bit. When you say "Based on the hydrological modeling performance on the 350 gauged catchments", do you mean the calibration performance? Please clarify here.

Yes, by hydrological model performance, we mean the calibration performance tested using the KGE objective function. This will be clarified in the revised version.

[Page 8, line 213] I would have liked to read a bit more about the variance analysis, about the methods and the aim of this analysis. E.g. What are the variance components and what do they tell you? Is such an analysis computationally expensive to run?

We will add the main relevant details in the revised version. Essentially, for each catchment we get 360 values for each metric, each value related to a unique combination of 1 GCM, 1 precipitation dataset, 1 hydrology model and 1 temperature dataset. The variance analysis attributes the percentage of the total variance of this vector of 360 values, to these components, including the interactions between these components, interactions meaning that the behaviour of one source depends on another source (for example, precipitation dataset may generate lots of variance with some GCM but not for others). Computing the main effect and first order

interactions is relatively cheap, computationally speaking, but higher orders (which typically carry much less variance) become exponentially costlier.

[Page 8, lines 214-216] Could you please provide an overview of the metrics computed, perhaps in a table?

Good idea. A table of the metrics will be added to the revised version of the paper.

[Page 8, lines 232-233] Please clarify that this observation is with regards to calibration for these 350 catchments. It could otherwise be misinterpreted taken out of context.

Will do. Your interpretation is correct.

[Page 9, line 256] The mean Summer Q also appears to show a different signal from the other metrics.

It is somewhat in the middle between the low-flow and the other metrics. This is not entirely surprising, since summer is a dry season over East Africa (between the spring and autumn monsoons) and therefore somewhat related to the low-flow metric. We will make a remark to this effect in the revised version.

[Page 9, lines 260-261] This is arguable and quite complicated to see. Perhaps putting the metrics in columns and the contributors in rows might help see these better?

This is a complex figure. We're not sure that doing so will help but we will try it and judge if one is preferable to the other.

[Page 9, lines 268-269] Which variance contributor is Fig. 6 shown for?

We will clearly have to be more explicit about this Figure in the revised version. The analysis of variance shows the percentage of total variance for each contributor, irrelevant to the total absolute variance. Trying to attribute relative variance to something that has little variance to begin with may not be that useful. So Figure 6 shows the absolute variance to try to make sense of this.

[Page 10, lines 284-285] It seems to be that most of the drop is seen between ensembles 1 and 2, rather than 2 and 3. Please also add the ensemble numbers 1-4 from table 4 to Fig. 7.

We see sizeable drops from 1 to 2 as well as from 2 to 3. We will rewrite it this way. We will add the ensemble numbers to the all relevant Tables and Figures in the revised version.

[Page 10, lines 290-292] Hydrological model uncertainty appears dominant over precipitation uncertainty for low flows.

Correct. We will state it clearly in the revised version.

[Figure 2] In order to be clearer for the readers, please consider adding clear subheading for each box in this diagram.

This is a good suggestion. We will implement it.