Referee #2

We first would like to thank the anonymous **referee 2** for the kind words in support of our manuscript and for the time spent reviewing our text. We appreciated the insightful comments that enabled us to improve the quality of our manuscript. Please, note that the original referee's comments are written in **bold**, and the author's responses are right below.

The manuscript is well written and structured, sound and pleasant to read. Little is said about the hydrological model (section 2.3) and the GCM output downscaling procedure (section 2.4). This simplifies the presentation and makes the manuscript easier to read. The readers are referred to previously published papers for more details. The results of this projection work, mainly presented in figures 4 and 5, are surprising and insufficiently discussed and commented. In fact, all projected average monthly streamflows appear very similar for all periods and the two scenarios (RCP 4.5 and RCP 8.5) and differ significantly from the actual situation (fig. 4). Such a little contrast between RCP 4.5 and RCP 8.5 is difficult to understand, especially for the second half of the 21st century where both projections differ greatly for the evolution of temperatures, which have a direct impact on potential evapotranspiration. This extremely strange outcome is acknowledged by the authors (P8, L3-5) but not explained nor discussed. As the rest of the manuscript and the conclusion are based on these results, a critical analysis appears to me as essential.

To keep our manuscript clear and easier to read, we decided to insert additional information about sections 2.3 in the supplement material following the Reviewer suggestion. Also, about section 2.4, the downscaling procedure is detailed described in Jones and Thornton (2000, 2013) as mentioned in P5, L22-25.

To clarify the Reviewer doubt regarding the similarity between monthly streamflow scenarios across the studied periods, we added the datasets in the supplementary material. In addition, it is important to remark that we use variables at a daily time step but assessed the results in terms of monthly averages on three 30-year time slices: near future (2010-2040); middle future (2041-2070); and far future (2071-2095). Therefore, the long-term monthly averages reveal the similar values of monthly streamflow scenarios. We will improve the discussion about it in the text. Thanks!!

1) The projected evolutions of temperatures and potential evapotranspiration for all periods and scenarios to be added in figure 4. This will certainly reveal clearer contrasts between scenarios.

We appreciate the reviewer's comments and suggestions. We hope to solve the problems of contrasts between scenarios by adding the projected evolutions of temperatures and potential evapotranspiration for all periods and scenarios, and, improve the discussion along with the text. We will add these full information in the supplementary material.

2) The GCM simulations for average monthly temperatures and rainfall for the actual period should also be presented. A major concern in climate change studies, especially when rainfall is considered, are the intrinsic biases of GCM models. A large amount of works have been devoted to the treatment of these biases to provide reasonable trends. Nothing is said about this problem in the manuscript and I highly suspect that the major differences between actual and projected situations, that draw the attention of the authors and on which their conclusions are focused, may be mainly due to these biases. If this is confirmed, the conclusions of the manuscript should be considered as invalid. This missing discussion and treatment of climate projection biases is a real major flaw and made me hesitate very much between suggesting "major revisions" or "rejection". It should absolutely be solved in a revised version of the manuscript.

We appreciate your concern about the treatment of climate projection. In this study, we assess the issue by using an ensemble of 17 stochastically downscaled GCM models. We chose to use an ensemble, instead of any single model projection, to reflect the range of uncertainties inherent to the current suite of GCMs, and also because reports have indicated that the ensembles, as a whole, provide superior performance to that of any individual model, as shown by Dhakal et al (2018) and Gleckler et al (2008). Mentioned in P5, L14-20.

Further, in the downscaled procedure by the MakSiM GCM statistical relationship with existing meteorological data from a met station was taken into account. Two aspects were considered in the downscaling: one was to interpolate the results of the GCM spatially; and the other was to ensure that the results are relevant to the local climate (using of 720 classes of weather, worldwide, to calculate the coefficients of a third order Markov rainfall generator). This constitutes 'stochastic downscaling' as it fits a Markov model to the GCM output and uses it to generate weather data for the site indicated (Jones and Thornton, 2013).

We will also add the GCM simulations for average monthly temperatures and rainfall data in the supplementary material.

Minor comments:

1) Since the manuscript is mainly focused on low flows, criterions specifically focused on the lower flow values should also be used to assess the hydrological model. R2, MSE and KGE are predominantly controlled by the larger discharge values.

The MSE and NSE are the two criteria most widely used for calibration and evaluation of hydrological models. They are closely related, but the results can be generalized to MSE (and similar criteria such as RSR) (Gupta et al., 2009). Using the R^2, or similar indexes as an objective function, the simulations are prior matching the high flow and these measures are oversensitive to extreme values (Jie et al., 2015). We choose to use KGE instead of MSE or NSE, whereas the KGE criteria is a decomposition of NSE (and hence MSE), which facilitates analysis of the relative importance of different components in the context of hydrological modeling (Gupta et al., 2009). Thus, Garcia et al (2017) recommend using the mean of KGE(Q) and KGE(1/Q) as an objective function to simulate low-flow indices with continuous conceptual rainfall-runoff models. Yet, performance during the calibration and evaluation periods can be considered quite good, representing both high and low flows, with R^2 and KGE values both exceeding 0.8, indicating a relatively high degree of correspondence between the model simulations and the observations (Gupta et al., 2009).

2) The precipitation unit must be clarified in figure 4 (mm/day)

Changes made according to reviewer’s suggestion

3) The figure reference numbering does not seem to be correct at some places in the text (4 rather than 5 at op8 L4 and P9L7).

Thanks for noticing our mistake. We corrected and also double-checked all figure reference numbering.

4) At P7L25: it must be specified that the authors speak about "hydrological dryness". The rainfall amounts start to rise in October and November even in the projections.

Thank you for the suggestion. We will change the term "dry season" to "hydrological dryness" along with the text.

5) P7L24: It cannot be stated, based on the presented results that rainfall extremes increase. The authors only present monthly averages. In general, the authors should avoid presenting conclusions that are not directly related or illustrated by the presented results. In the same line of thought, plant water stress mentioned on P8L2 should be illustrated (through the simulated soil water contents for instance). By the way, how is the vegetation cover reaction to the climate change taken into account? Again, it is suggested that RCP 8.5 generates more intense rain : please illustrate this fact based on the available projections. P9L20 : There is no direct relation between the increase of extreme rainfall and the possible increase of monthly rainfall in December. As for the previous remark, if it is true that the projected rainfall amounts are linked to an increase of the frequency of extreme events, this can be illustrated based on the climatic projections.

Thank you for this important remark. We will modify the paragraph (P1,L20-24) and also other conclusions that are not directly related to the presented results (P8, L2 and P9, L20). Since we wanted to quantify the relationship between climate and streamflow, we chose to use a lumped conceptual "rainfall excess" type of catchment system model, that despite the relative simplicity of its structure facilitates computationally fast data processing, and it imposes minimal requirements for input data (precipitation and potential evapotranspiration), while maintaining a suitable level of hydrological process representation (Gong et al., 2013). As mentioned in (P4, L12-22). Therefore, the vegetation cover wasn't taken into account in our investigation.

6) Figure 5 increases dramatically the undetectable contrasts of figure 4. Why? Moreover, some inconsistencies seem to exist between the two figures. If the demand is considered as relatively constant over the year (if it is not, this should be explained and commented by the authors), discharges and scarcity and vulnerability indicators should have the same dynamics. It is not the case. The lowest simulated discharges are observed

in October for all scenarios and periods (fig 4); Why are then peak indicators values computed in November? Some explanations are clearly missing.

To assess water security, we used the approach developed by Rodrigues et al. (2014), in which water use (Abstraction and Consumption) is contrasted with probabilistic levels of Water Provision , based on the fulfillment of environmental demand represented by an Environmental Flow Requirement (EFR), as cited in (P6, L5-7). Thereby, the indices are similar, but not the same. Water Scarcity assesses the impacts of water use on median water availability for consumption (50%), while Water Vulnerability expresses the susceptibility of water withdrawal for human activities under low-flow (30%), or drought-like, conditions, as described in (P6, L10-12). That's why figure 5 shows the undetectable contrasts of figure 4, besides using 50% (median water availability for consumption) and 30% (water withdrawal under low-flow), we also had to fulfill the EFR (discounting this value from total streamflow).

We developed seven demand scenarios for future periods, approached as "threshold levels", defined based on non-stationary demand as a hypothesis representative of the population growth in the Sao Paulo Metropolitan Region, as mentioned in (P6, L26-28 and P7, L1-2). The demand scenarios are represented in figure 5 by the layers, as described in figure label (P23, L 2-6). The lowest simulated monthly streamflow is observed in October for all scenarios and periods, and the highest value of Water Scarcity and Vulnerability indicators for RCP 8.5. On the other hand, in scenario RCP 4.5 we observed the highest values of Scarcity and Vulnerability indicators in November. This happens because, in the early days of November, the streamflow was so low, that couldn't even fulfill the EFR, raising the monthly index to high values, including the maximum in some cases.

We will include this explanation into the discussion section in the revised manuscript to be submitted, and also add the Water Scarcity and Vulnerability indicators data sets and calculation in the supplementary material.

7) Figure 6 is not needed since the same results as in figure 4 are presented, except that error bars have been added. It is by the way not explained how these boxplots have been build. Do they represent the inter-annual variability (this is what I suspect)? Or do they represent the variability of the projections of the 17 tested GCMs. By the way, these 17

simulations and the information provided by the variability of their outcomes are never presented nor used in the manuscript. This should be added somewhere.

Thank you for the suggestions. We want to clarify that figure 4 shows the monthly average on three periods: near future (2010-2040); middle future (2041-2070); and far future (2071-2095), and figure 6, as asked, the inter-annual monthly variability. The explanation of its construction is in the figure caption (P24, L2-5). We decided to maintain the figure 6, since presents a different result than figure 4, and we will improve the discussion of the results in the revised manuscript to be submitted. Additionally, we will add the ensemble of 17 stochastically downscaled GCM models data in the supplementary material.

8) Section 3.4 is not really related to the rest of the manuscript. These thoughts about public policies are not totally uninteresting, but not supported by the presented results. In would suggest to remove this part, or to summarize it in the conclusion of the manuscript.

We appreciate the reviewer's comment and suggestion. In the revised version, we will add substantial information and clarifications discussed by referee 2. In addition, the data set in supplementary materials. We believe in this way, there will be no doubt regarding the data sets and therefore we consider essential to maintain section 3.4. Whereas section 3.4 improve the comprehension of our scientific contributions, discusses the applicability of the results and it is critical to guide future studies.

References

Dhakal, K., Kakani, V. G. and Linde, E.: Climate Change Impact on Wheat Production in the Southern Great Plains of the US Using Downscaled Climate Data, Atmos. Clim. Sci., 08, 143, doi:10.4236/acs.2018.82011, 2018.

Garcia, F., Folton, N. and Oudin, L.: Which objective function to calibrate rainfall–runoff models for low-flow index simulations?, Hydrol. Sci. J., 62(7), 1149–1166, doi:10.1080/02626667.2017.1308511, 2017.

Gleckler, P. J., Taylor, K. E. and Doutriaux, C.: Performance metrics for climate models, J. Geophys. Res. Atmos., 113(D6), doi:10.1029/2007JD008972, 2008.

Gong, W., Gupta, H. V, Yang, D., Sricharan, K. and Hero, A. O.: Estimating epistemic and aleatory uncertainties during hydrologic modeling: An information theoretic approach, Water Resour. Res., 49(4), 2253–2273, doi:10.1002/wrcr.20161, 2013.

Gupta, H. V, Kling, H., Yilmaz, K. K. and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377(1), 80–91, doi:10.1016/j.jhydrol.2009.08.003, 2009.

Jie, M.-X., Chen, H., Xu, C.-Y., Zeng, Q. and Tao, X.: A comparative study of different objective functions to improve the flood forecasting accuracy, Hydrol. Res., nh2015078, doi:10.2166/nh.2015.078, 2015.

Jones, P. G. and Thornton, P. K.: MarkSim: Software to Generate Daily Weather Data for Latin America and Africa, Agron. J., 92(3), 445–453, doi:10.2134/agronj2000.923445x, 2000.

Jones, P. G. and Thornton, P. K.: Generating downscaled weather data from a suite of climate models for agricultural modelling applications, Agric. Syst., 114, 1–5, doi:10.1016/j.agsy.2012.08.002, 2013.

Rodrigues, D. B. B., Gupta, H. V and Mendiondo, E. M.: A blue/green water-based accounting framework for assessment of water security, Water Resour. Res., 50(9), 7187–7205, doi:10.1002/2013WR014274, 2014.