Review of the manuscript „A framework to regionalize conceptual model parameters for global hydrological modeling“

This manuscript examines a variety of regionalization approaches applied to regionalize parameters of four catchment-scale conceptual models to global grid cells. The performance of standard regionalization techniques based on spatial proximity, physical similarity and the combination of both is examined for several thousand catchments world-wide and is compared to the performance with at-site calibrated parameters. The combination of best-performing regionalization approaches are used to interpolate parameters from gauged locations to the grid cells word-wide and global water balance components are estimated using four different conceptual hydrological models.

The comparison of regionalization methods for global scale hydrological modeling has an immense importance for reliable estimation of global water resources. However, it is not clear how the framework proposed in this study advances the fidelity of global hydrological models. The components of the proposed framework are not clear defined making it rather difficult to understand the novelty and the advantages of this work compared with previous studies. Little insights and discussion is provided on the effect of parameter uncertainty on the estimates of global water balance components. Moreover, the introduction of more recent works on model parameter regionalization, especially the work tackling parameter discontinuity for regional and global studies is missing. Some critical assumptions (e.g., on independence among catchment descriptors or that catchments with similar catchment descriptors have similar model parameters) were neither tested nor critically discussed. The reported differences in performance of different regionalization methods and models is minimal. Finally, several missing methodological aspects regarding distance calculation between the catchments, unclear distinction between calibration and evaluation catchments and interpolation to the global grid cells makes it difficult to evaluate the credibility of this study. Therefore, I think a substantial revision of the manuscript is required. Below I present my detailed comments.

General comments

1. Introduction should clearly define the gaps that currently exist in parameter regionalization for regional and global hydrological models and should clearly state how this study tackles these problems. Currently, the Introduction is rather structured as a listing of performed studies without assessment of their advantages and disadvantages for global scale hydrological modeling making it difficult to understand the novelty of this study. In my opinion this study rather uses well-established techniques with known flaws and merely compares their performance at global scale in terms of single performance metric. A more clear novelty statement should be presented to make it clear how this study solves or advances current issues in global scale modeling and regionalization of model parameters. I miss also the discussion on the issue of model parameter discontinuity for regional and global hydrological modeling (e.g., Samaniego et al., 2017) and how the proposed framework can deal with it.

2. The description and the components of the proposed framework are not clear. It is not clear from the manuscript if the framework is actually a combination of all steps (i.e., different regionalization methods + different catchment scale conceptual models + interpolation to grid cells) or if it is just the selection procedure for regionalization methods. It is not clear which features does make it a framework and not a simple sequence of methodological steps. Moreover, the workflow itself has to be clarified too. Specifically, it is not clear if the grids

that correspond to donor catchments were preserved or regionalized as well. It is not clear which portion of catchments was left for evaluation of global regionalization. Please, also see my specific comments regarding these issued below.

3. Regionalization methods examined here refer to so called two step similarity approaches (Wallner et al., 2013) where donor catchments are independently calibrated and then their parameters are regionalized based on various similarity metrics (e.g., spatial proximity or physical similarity or both). Such approaches were reported to suffer from equifinality problem (Bárdossy, 2007; Götzinger and Bárdossy, 2007). Alternative one step approaches (e.g., Hundecha and Bárdossy, 2004; Samaniego et al., 2010; Mizukami et al., 2017) that were developed to tackle this problem are not discussed in this study. Only a very simplistic uncertainty analysis was performed in this study, but nevertheless exposed large equifinality problem for all 4 models. It was not shown how the detected parameter uncertainty has affected regionalization performance and has propagated to the global water balance estimates. Little discussion was provided on the reliability of final global water balance estimates.

4. Several crucial assumptions were neither tested not critically discussed (e.g., independency of catchment descriptors used in this study; link between physical and hydrological similarity). Moreover, several methodological details are missing or ignored (e.g., catchments affected by human activities are not rigorously filtered out; no indication on spatial discretization used in the models; no indication if the distance between catchments was calculated based on outlets or centroids; no clear distinction between donor and evaluation catchments), making it difficulty to the judge the credibility of this study. Please see my specific comments for details.

5. Writing and structure of the paper needs to be improved. Moreover, please check consistency of tense and plurals/singulars of nouns and verbs.

Specific comments

Line 36-43: According to this sentence in Line 36-38 land surface scheme models are the most commonly used global hydrological models. In the next two sentences it is stated that due to large biases of these models global hydrological models are developed. These statements are confusing, please clarify.

Line 44-60: I miss here an overview of the approaches that tune model parameters of global hydrological models (e.g., WaterGAP by Döll et al., 2003).

Line 71: I do not think that regionalization technique proposed in Samaniego et al. 2010 (referred here is Luis et al., 2010) can be attributed to clustering or hydrological classification.

Line 74: I would welcome here more insights on why despite numerous studies on comparison of regionalization approaches it is still not clear which of them perform better and how this study contributes to identification an appropriate method.

Line 76-77: Accounting for uncertainties in global hydrological modeling is an important issue that indeed was not tackled sufficiently in the previous studies. However, it is not clear how this study accounts for uncertainties? From what I have seen later in the paper, the authors indicated presence of considerable parameter uncertainty in each model and showed that global runoff estimates vary considerable among tested models. But does it really mean that the uncertainty is accounted for in this study?

Line 96: Please specify the spatial resolution of the precipitation product.

Line 101-102 and Table 1: I urge the authors to carefully examine the assumption about independency among catchment descriptors. I am afraid some of them have strong dependencies (e.g., slope and elevation or aridity index and mean annual potential evaporation). Instead of rather uninformative Table 1 that presents mean, max and min values of catchments descriptors globally (clearly a wide variety is to be expected here due to the global focus of the study), a correlation matrix would instead confirm or oppose the assumption of independency. The second assumption on the existence of well-behaved relationship between catchment descriptors and model parameters is certainly more difficult to prove, but at least a note on possible issues with this assumption has to be stated (see e.g., Odin et al., 2010 or Merz et al., 2020).

Line 111-119: Indeed, presence of regulated catchments creates a considerable obstacle for hydrological modeling. Does the selection based only on catchment area is able to filter out the regulated catchments? I suggest authors to include additional criteria that are customary in large-scale hydrological modeling (e.g., remove catchments with large dams, Lehner et al., 2011, Grill et al., 2019; test the closure of water balance, Beck et al., 2016).

Section 2.3: The rationale on selection of these four models is not clear to me. It is noted (Line 76-77) that by considering models of different structure and concepts one might account for the model uncertainty. However, from the description provided for each model in this section, it is not clear which differences in their concepts and structures apart from the number of parameters exist. Moreover, I miss here the indication if the models were applied in lumped or in a distributed fashion.

Line 156-157: According to the catchment selection rules listed in Line 115-116 catchments with at least 5 years of observations were selected. This would mean around 3.5 years for calibration in these catchments. How many catchments have such short calibration period? Do you think such short calibration period might increase parameter uncertainty?

Line 167-177: I would welcome here the rationale on selection of these regionalization methods. Most of the approaches used here are two step similarity approaches, meaning that in the first step the model parameters are identified at gauged locations independently from similarity or spatial proximity of the catchments. In the second step calibrated parameters are regionalized assuming that closer catchments or more physically similar catchments have similar model parameters. Such approaches are not able to account for equifinality of model parameters (Bardossy, 2007; Götzinger and Bardossy, 2007). Therefore, neighboring catchments and even physically similar catchments not necessarily obtain similar parameter sets during calibration (Oudin et al., 2010). The methods used in this study are likely to suffer from a similar problem. This issue has to be addressed here and put into perspective of one step similarity approaches that were specifically developed to target this issue (e.g., Samaniego et al., 2010; Wallner et al., 2013; Mizukami et al., 2017).

Section 2.5: Several important details are missing in this section. Was the best performing parameter set regionalized or a mean parameters from 10 best sets mentioned in Lines 310-315? Was spatial proximity defined based on geographical position of the outlet or of the catchment centroid? Was the whole parameter set transferred or each single model parameter was regionalized independently? No details on regression method is provided. It should be specified what kind of regression model was used, which parameter estimation method was used and if all catchment descriptors were used to build the regression model. It should be also reported if the regression model was built for each individual model parameter or for the whole set.

Line 182-184: How accurate is distance calculation on these projected coordinates globally? Why not to simply calculate geographical distances? Does "catchment position" refers to catchment outlet or catchment centroid?

Line 187: I think Luis et al. 2010 and Samaniego et al. 2010 is the same study.

Line 210-211: What does this sentence mean? Please clarify which catchments were used for the evaluation of regionalization techniques.

Line 213-218: It is not clear how exactly the parameters were regionalized to grid cells. Were the calibrated parameters of donor catchment preserved during regionalization? If they were preserved, were they assigned to catchment centroid or catchment outlet? How other grids within donor catchments were assigned?

Section 2.7: It is not clear what GSRS states for. Till Line 219 part I had an impression that GSRS is the proposed framework for the selection of an appropriate regionalization technique to transfer parameters of catchment scale conceptual models to the global grid cells. Therefore, the statement in this sentence comes surprising. Please clarify if GSRS is actually the proposed regionalization framework mentioned elsewhere in the manuscript. Please clarify its description and components.

Line 221-225: It is not clear to me how the Network Response Routing converges grid cell runoff to catchment streamflow. Please clarify it and indicate what are the two parameters mentioned here. Moreover, please clarify how these two parameters exactly were calibrated. If these parameters were not regionalized how is it possible to obtain catchment streamflow at ungauged locations (i.e., grid cells)? Furthermore, I find merging of the conceptual model parameters with Network Response Routing parameters rather confusing. Please explain why this step was necessary.

Line 226: Does this mean that Network Response Routing is the part of framework? So does it mean that the framework is actually GSRS+NRF?

Line 241-242: Is precipitation gauge density is the only reason for good performance of hydrological models in these regions? Could these spatial variability in model performance result from inability of selected models to simulate discharge in drier areas compared to wetter areas?

Line 266-268: Where these results can be seen?

Line 273-276: Why should poorly calibrated donors be considered in further regionalization? Is likely that by using the poorly performing parameter set additional uncertainty will be added to the regionalization. What was the criteria to choose the threshold value?

Line 285-289 and Figure 5: Are the difference in the performance of different regionalization methods considerable? I see considerable differences among calibration, global mean and regression method compared to all other regionalization approaches. Apart from that the differences are minimal. Taking into account that it seems that only one best performing parameter set was used for regionalization, I am wondering if these differences would still be visible after accounting for the parametric uncertainties.

Section 3.2: I am missing here the discussion on such poor performance of the regression method. Could it be linked to the multicollinearity of catchment descriptors? Was the whole parameter set regionalized with this method or were regression models built for each model parameter individually? Was a global regression model built or was it region-specific?

Figure 6: I am not sure that I see in this Figure that 1500 km is an optimal threshold. For GR4J 1300 km and for HMETS and XAJ 1200 km seems to be more appropriate. How exactly was threshold of 1500 km identified? Is this threshold sufficiently robust given that there are large fluctuations in the performance of the methods around this distance (e.g., for GR4J for 1400 km the methods based on both physical similarity and distance outperform all other methods for all catchments, while for 1300 km and 1500 km spatial proximity is the best performing method). Why the x axis in the subplots differ? For some it reaches till $18*10^5$ m for others only till $15*10^5$ m. For all subplots the last bar seems to be cut. Please clarify what "Physical Similarity DIS" in the legend refers to?

Line 301-304: The difference in performance between the models is of the same magnitude as between the regionalization methods (apart from global mean and regression method). Therefore, I find the conclusion that model structure has an insignificant effect while the choice of regionalization method is important rather inconsequent.

Line 310-320: This is a description of methods and should be moved to the methods Section.

Line 311-312: Please clarify which criteria was used to select sufficiently different parameter sets.

Line 328-329: This statement requires a citation. If this is a finding of this work consider reformulating this sentence accordingly.

Line 329-330: One way to quantify the influence of parametric uncertainty for regionalization results is to perform regionalization using all equifinal parameter sets (i.e., 10 best performing parameter sets in this study) and analyze the differences in regionalization performance and global runoff estimates.

Line 348-352: How the deterioration in performance between the regionalization at catchment scale and regionalization at grid cell can be explained? If the same donor catchments were used and the best performing regionalization methods were selected why deterioration occurs at grid cells? Is this deterioration is caused solely by averaging over the grid cell?

Line 352: Does "-G" stands for global version of respective hydrological model? Please introduce appropriately this new notation and explain how the global version is different from catchment version. Later (Line 354, 356, Table 4 and elsewhere) "-G" notation is not used anymore, although it seems like the global set up of the models is meant there as well.

Line 359-361: The values reported here and in Table 4 seems to be median KGE for all catchments (including donor catchments), while Beck et al. (2016) has reported KGE values for evaluation catchments only. Therefore, this comparison is not fair. Moreover, please consider reporting regionalization performance for evaluation catchments only. Without these results it is difficult to judge the credibility of the proposed framework.

Line 364-366: In the later part of this paper the authors refer to the 4 catchment scale conceptual models regionalized to grid cells as global hydrological models (GHMs). The argumentation provided in the review of global hydrological models by Sood and Shmakhtin

(2015) that global hydrological models were built for macroscale water resources management and not for streamflow simulation cannot be justified here, as Sood and Shmakhtin (2015) do not refer to regionalized catchment scale conceptual models but to actual global hydrological models that either are the components of general circulation models or stand-alone hydrological models with very few or no calibrated parameters that run directly at the global scale. Instead, the models derived here were specifically designed to improve streamflow simulation and calibrated to the observed streamflow.

Figure 9: Please clarify in the caption if the cumulative curves are showed for all catchments or evaluation catchments only. It seems that y axis is cut at 0, please indicate it in the caption. Clarify what CAL and GSRS stands for in the caption.

Table 4: Please clarify what “framework” stands for? Is it an equivalent of “GSRS” from Figure 9?

Line 383-384: According to Table 5 SIMHYD-G performs the best. Please clarify the statement in this sentence.

Line 389-390: Although I generally agree that the balance between model flexibility and complexity is very important, I would say that the results of this study rather show the opposite. Minor differences in KGE between the models (0.01 of KGE between the best and the worst one; this is exactly the threshold for performance difference that the authors selected to identify equifinal parameters earlier) indicates that the choice of the model played rather negligible role here. Moreover, earlier (Line 300-303) it was stated that model structure played insignificant role for regionalization performance in this study supporting rather the other conclusion of this work that potentially any model can be used in the proposed framework.

Line 422-424: A reference is needed for these estimates.

Line 426-430: This portion makes me wonder why the available global datasets on anthropogenic influence (see my comment to Line 111-119) were not considered to filter out the affected catchments. Are there any prospects on obtaining reliable global precipitation datasets that can be used in the future studies?

Line 462: Usage of “GHM” in this context is confusing, as in this study the GHM were built from catchments-scale conceptual models by combining them with various regionalization approaches. If I understood correctly these steps combined are the framework proposed in this study. Consider using “conceptual model” instead.

Editorial comments

Abbreviations: The manuscript is oversaturated with abbreviations. Some of these abbreviations are never used after their introduction (e.g., LSS, PUB, AOF), some other abbreviations are only used once or twice (e.g., IDW, AM, CDF). Consider omitting using unnecessary abbreviations in the manuscript and especially in the abstract (e.g., NRF, IDW, KGE).

Line 15: Why Network Response Routing is abbreviated as NRF. Should it be NRR instead?
Line 54: a priori
Line 67, 71 and elsewhere: Samaniego et al., 2010 instead of Luis et al., 2010
Line 76: Moreover
Line 85: delete “widely”
Line 383: “proposed” instead of “proposal”

Line 384: I think you mean here that it was not confirmed in this study.
Line 385: **"**that" instead of "who"
Line 387: **"**is not reduced" instead of "does not reduce"
Line 389: **"**in" instead of "on"
Line 389: might be
Line 453: "highest" instead of "largest"
Figure 1: Please clarify if the Figure actually shows the location of catchment outlets or centroids.
Figure 3: Please indicate in the caption if model efficiency corresponds to calibrated model parameters here.
Figure 4: Why global mean and regression methods are not shown here? Please label the x axis. It would be useful if you will select the same colors for performance classes as in Figure 3.
Figure 5: please specify what CAL stands for. If it stands for calibration, consider moving its box to the first position. Please clarify box plot structure in the caption since whiskers and the outliers are not plotted here.
Table 2: Why only SPA and SPI-OUT are mentioned in the caption?
Table 6: Please indicate if any model was used to provide runoff estimate in the study of Korzun et al. (1978) and GRDC; please indicate which period was modelled in Widén-Nilsson et al. (2001). What does "approximately" means for GRDC time period?
Figure 11: Please indicate the scale for precipitation. Consider transforming discharge to mm/day to make it comparable with precipitation. Please remove a black line on the right and the double full stop in the caption.

References

Bárdossy, A. (2007). Calibration of hydrological model parameters for ungauged catchments. *Hydrology and Earth System Sciences*, *11*(2), 703–710. https://doi.org/10.5194/hess-11-703-2007

Götzinger, J., & Bárdossy, A. (2007). Comparison of four regionalisation methods for a distributed hydrological model. *Journal of Hydrology*, *333*(2–4), 374–384. https://doi.org/10.1016/j.jhydrol.2006.09.008

Grill, G., Lehner, B., Thieme, M., Geenen, B., Tickner, D., Antonelli, F., … Zarfl, C. (2019). Mapping the world's free-flowing rivers. *Nature*, *569*(7755), 215–221. https://doi.org/10.1038/s41586-019-1111-9

Hundecha, Y., & Bárdossy, A. (2004). Modeling of the effect of land use changes on the runoff generation of a river basin through parameter regionalization of a watershed model. *Journal of Hydrology*, *292*(1–4), 281–295. https://doi.org/10.1016/j.jhydrol.2004.01.002

Lehner, B., Liermann, C. R., Revenga, C., Vörömsmarty, C., Fekete, B., Crouzet, P., … Wisser, D. (2011). High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management. *Frontiers in Ecology and the Environment*, *9*(9), 494–502. https://doi.org/10.1890/100125

Merz, R., Tarasova, L., & Basso, S. (2020). Parameter's Controls of Distributed Catchment Models—How Much Information is in Conventional Catchment Descriptors? *Water Resources Research*, *56*(2), 1–18. https://doi.org/10.1029/2019WR026008

Mizukami, N., Clark, M. P., Newman, A. J., Wood, A. W., Gutmann, E. D., Nijssen, B., … Samaniego, L. (2017). Towards seamless large-domain parameter estimation for hydrologic models. *Water Resources Research*, *53*(9), 8020–8040. https://doi.org/10.1002/2017WR020401

Oudin, L., Kay, A., Andráassian, V., & Perrin, C. (2010). Are seemingly physically similar catchments truly hydrologically similar? *Water Resources Research*, *46*(11), 1–15. https://doi.org/10.1029/2009WR008887

Samaniego, L., Kumar, R., Thober, S., Rakovec, O., Zink, M., Wanders, N., … Attinger, S. (2017). Toward seamless hydrologic predictions across spatial scales. *Hydrology and Earth System Sciences*, *21*(9), 4323–4346. https://doi.org/10.5194/hess-21-4323-2017

Wallner, M., Haberlandt, U., & Dietrich, J. (2013). A one-step similarity approach for the regionalization of hydrological model parameters based on Self-Organizing Maps. *Journal of Hydrology*, *494*, 59–71. https://doi.org/10.1016/j.jhydrol.2013.04.022