Reviewer 2

General comments:
The authors present a very good work on assessing the reliance of model regionalization approaches on the gauged data content. The methodology is clear and easy to follow. Their results indicate that transferring the entire parameter set, thus keeping the correlation among model parameters, outperforms the parameter-averaging kriging method. The output-averaging methods using more than one donor basins is more robust than the one-donor method. The most similar method based on geomorphological and climatic descriptors tends to show higher transferability in sparsely gauged areas than the nearest neighbor method. The findings provide important reference for the community when conducting hydrological modeling in sparsely or ungauged basins.

A) My major concern is that the structure of the manuscript is a bit scattered. The abstract and results sections are lengthy, from my taste. The authors could consider potential rephrase or reorganization. In addition, some Paragraphs in the results section should move to the methodology section.
Reply: As suggested, we will substantially cut and revise the abstract and we plan to move some paragraphs from the results to the methodology section (see also reply to minor comments).

B) My second concern is that the applied two hydrological models have different calibrated/regionalized parameter spaces. In particular, the TUW model implies 11 parameters for calibration, while the GR6J model has less 8 parameters for calibration/regionalization. Some discussions on the effects of the size of parameter space on the results are needed.
Reply: The issue of the complexity/parsimony of the models is a very important topic: in our case, we assume that the complexity of the two models is similar, since the number of parameters is not so different (8 vs 11 parameters).
Previous studies in the literature applied parameter regionalisation to different conceptual rainfall-runoff models with different number of parameters: Chiew et al. (2010) and Petheram et al. (2012) implemented various models with parameter number varying between 6 and 14 without reporting a clear dependence of model performances on the dimension of parameter space; Viney et al. (2009) used models with 6 to 13 parameters and reported an increasing calibration performance for larger parameter space, but results did not show such a clear behaviour in regionalisation, where results were more dependent on the approach.
For what concern our study, in both results (l. 382-383, 443-444, 538-540) and discussion (l. 649-651, 665-666, 697-699) sessions we argued that GR6J model performances "at site" are slightly better than TUW despite the lower number of model parameters, while in general, for regionalisation, the TUW parameters seems to be easier to transfer from gauged to ungauged catchments (and this is even more accentuated when excluding nested donor catchments).
We think that this effect is probably not so much related to the number of parameters, but rather due the higher conceptualisation level of the GR6J model, that may lead to have some parameters that are very case-dependent, and are therefore related to physical basin characteristic (and thus to catchment similarity): for example, the groundwater module of GR6J includes an exchange function (not complying with the volume balance) based on two parameters X2 and X5 which are very sensitive to the specific local conditions and therefore likely difficult to transfer.
In the revised version we will add, as suggested, that the results may depend on the different model structure and on the different transferability of model parameters (depending also on their meaning and their relation with the available catchment attributes). This aspect would deserve more attention and investigation but it would need a separate ad-hoc analysis, since the comparison of the structures and of the physical meaning of the parameters of the two models is not the specific objective of our work.

C) Third, the studied 209 catchments vary across a large range of area from 13 to 6000 km^2. But, the models didn't use distributed parameter values. That is, the spatial variability of parameters in large basins wasn't considered. The author should also add discussions on the performance of the regionalization approaches in catchments with distinct basin areas (small, moderate and large basin).

Reply: This is indeed a very interesting question, that we had not explicitly analysed yet. In general, evidence from the literature tends to show that regionalisation performance as well as "at site" performances (see, e.g. Merz et al., 2009; 2011; and Nester et al., 2011) improve for increasing catchment size. For instance, in the review of Parajka et al. (2013) the authors gather and compare the outcomes of different studies predicting runoff hydrographs through rainfall runoff model regionalisation, finding a clear pattern of an increase of model accuracy with catchment scale. As they say, this is generally due to two reasons. The first is a trend for an increasing number of raingauges within a catchment as the catchment size increases, which increase accuracy of rainfall input to models. The second may be related to the aggregation effect of runoff: as the catchment size increases, some of the hydrological variability is averaged out due to an interplay of space–time scale processes, which will improve hydrological simulation.

We have checked, as suggested by the referee, if the performance of the regionalisation methods is influenced by basin size: as suggested, we have divided the study sample in three equally numerous groups:

- Small (area <= 99 km^2, 70 catchments)
- Medium (99 km^2 < area <=285 km^2, 69 catchments)
- Large (area > 285 km^2, 70 catchments)

and we have checked for any change in regionalisation performance. Similarly to Figure 7 in the manuscript, Figure R.2.1 below shows KGE regionalisation performances for the three groups and the two models.

It may be seen that there are no strong differences between the three groups but, as expected, worse regionalisation accuracies occur in general for smaller catchments.

More interestingly, the ranking of the performances obtained by the different methods does not significantly change for different catchment size.

We do not think that adding a new, large figure in the revised manuscript is needed (as highlighted by the referee we have already too many figures) but we will specify despite the different extension of the catchments in the dataset, even if worse performances tend to occur in general for smaller catchments, consistently with previous evidence from the literature (see, e.g. Parajka et al 2013), the efficiencies of the regionalisation methods (and their ranking) do not show a clear relation with the size of the watershed. And we will also specify that in our study region catchment size ranges mainly between 13 and 1000 km^2 (90% of the basins) and just 3 watersheds extend over more than 3000 km^2.

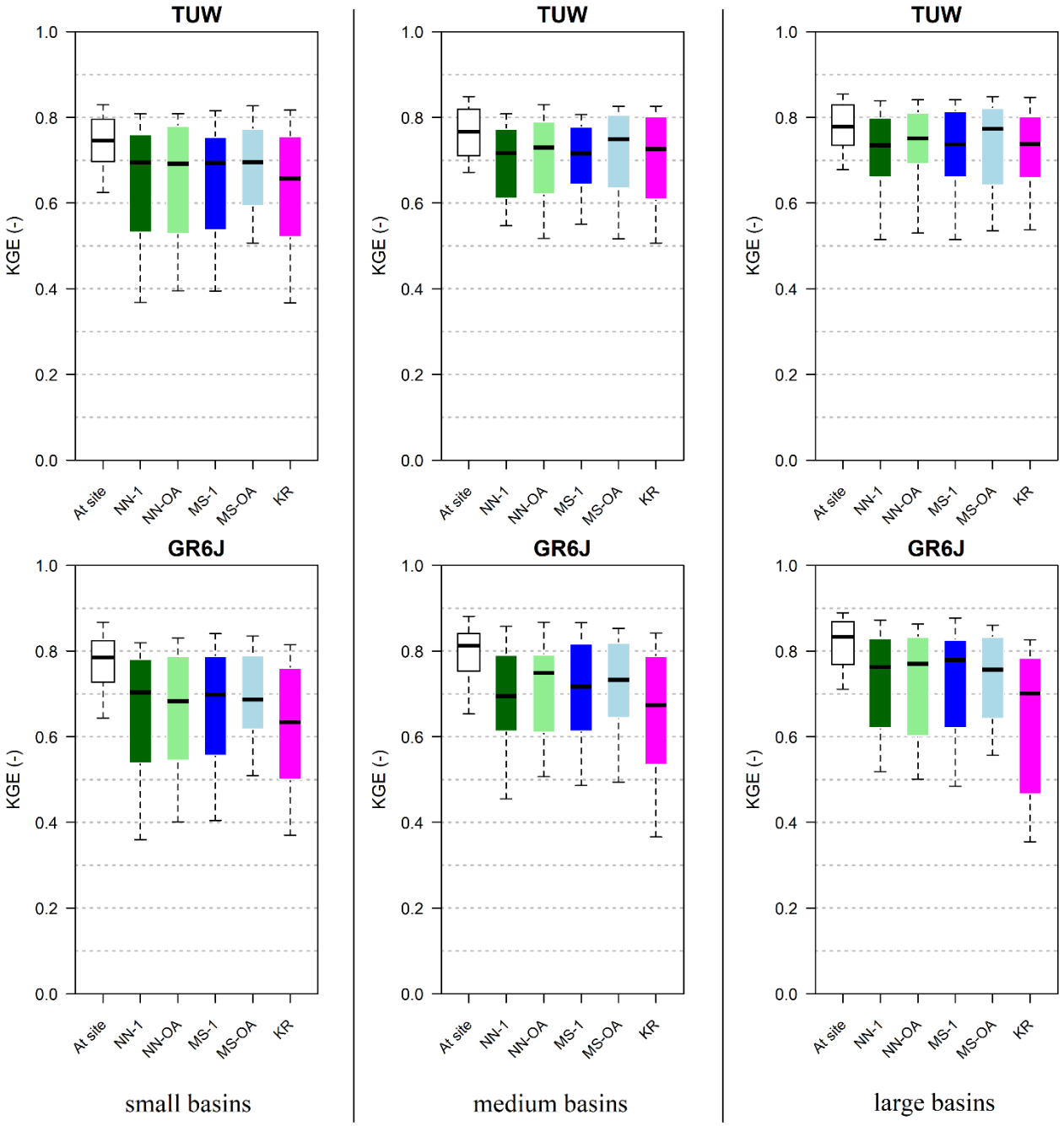
TUW
GR6J
KGE (-)
1.0
0.8
0.6
0.4
0.2
0.0
At site
NN-1
NN-OA
MS-1
MS-OA
KR
small basins
medium basins
large basins

Figure R2.1.

D) Finally, the authors should also specify how they sampled catchment subsets from the total 209 catchments when investigating the effects of station density. Did they sample manually or using automatic scripts? The thing here is how to guarantee that the sampled catchments are evenly distributed across the country. The sampled catchments concentrating in a small region could result in a same station density with the evenly distributed catchments.

Reply: Thank you for expressing your concern about this issue: we agree that it deserves additional clarification and further analysis.

We performed a simple automatic non-supervised sampling (and this issue will be much better specified in the revised version).

The chance that a cluster of catchments results to be sampled in a small region can not be excluded, even if it is unlikely. For this reason, the samplings are repeated 100 times for each value of station density, thus considering several groups of basins with the same density in the study region.

For addressing the point suggested by the referee, we have verified the distribution of our samples across the country, and we plan to add such new analysis to the revised version.

In order to verify if the catchment samples are sufficiently evenly distributed across the country, let's consider a group of catchments and let's measure the distance of each catchment from its closest potential donor as shown the following figure:

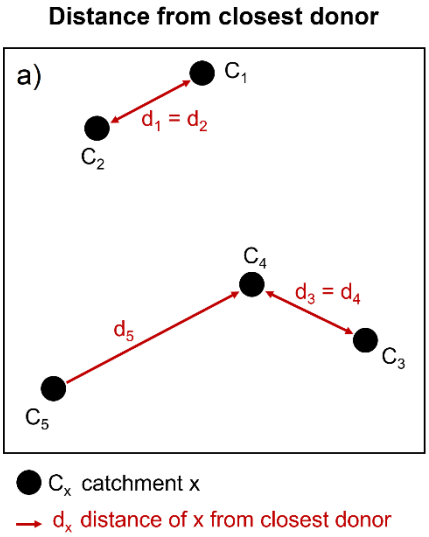

Figure R2.2. Example of distance from closest donor.

The average of the distances (d_1, d_2, d_3, d_4, d_5) of each catchment from the closest catchment (i.e. a potential donor) in a sample can be considered as a measure of the sample spatial distribution: the higher the distance the less dense the sample. As above said, for each density, 100 different samples are generated, so that for each density, we have 100 different values for such averages.

The figure below shows the average "distance within sample" of the closest available donor catchment across the 100 generated sub-sets for the different values of station density (each boxplot refers to the 100 values of average distance calculated for each sub-set). The average distance from the closest donor in the original, full density dataset (grey point in the figure) is around 8.5 km. As expected, the median target/donor distance (middle black solid line in each box) increases with decreasing density: it is true that also the variability of the distance, as shown by box size and whiskers, gradually increases with the reduction of station density, but such increase is overall modest: even for the lowest density, it is limited to +/- 18% of the median for the 80% of the samples. The fact that, on average, the distance between a target catchment and the closest gauged catchment consistently increases for decreasing density proves that the samples with lower density do not tend to cluster/concentrate the catchments in a small region, but there is an even distribution over the country.

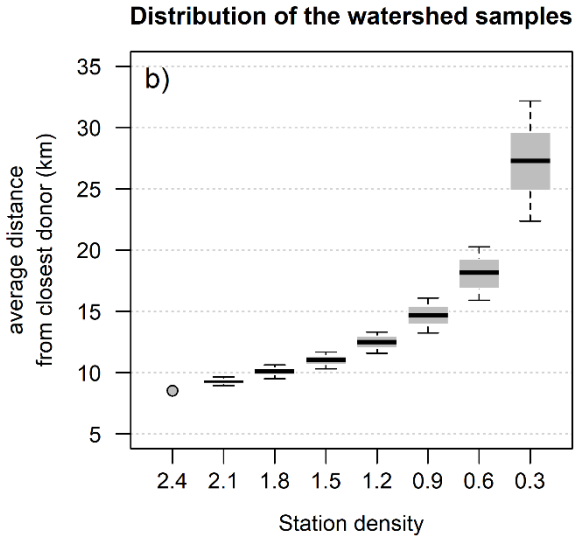

Figure R.2.3. Boxplots of the average distance within sample from the closest available potential donor catchment across the 100 generated sub-sets, for different values of station density (gauges/1000km^2). Whiskers extend to 10th and 90th percentiles. The grey point indicates the average distance from the closest donor in the original dataset."

Specific Comments:

1. Please, keep "HBV" and "TUW" consistent. Otherwise, the readers would find three hydrological models in this study. Actually, there are only two models.

Reply: Thank you for the correction. We will change "HBV" to "TUW" when referring to the model used in this study.

2. Lines 8-10, consider to move to introduction section.

Reply: We think that it may be useful for introducing the topic to readers that may not be acquainted with the regionalisation of rainfall-runoff models, and we would prefer keeping this sentence in the abstract.

3. Lines 21-22, may consider to remove.

Reply: We agree, it will be removed in the revised abstract.

4. Lines 33-34, you already specified "how" above. Consider to remove.

Reply: We agree and will modify the abstract avoiding the repletion.

5. Line 67, "the best the best".

Reply: We will correct this typo.

6. Line 96, what do you mean by "continuous simulating daily models"?

Reply: We mean they are not event-based models. We will change it to "continuous-simulation daily rainfall-runoff models" as used more frequently in hydrology (e.g., by Crooks and Naden 2007).

7. Line 116, "the more".

Reply: We will change "to the more than" to "to more than".

8. Line 259, how to calculate "stream network density", please specify this

Reply: This attribute was computed in previous studies over the same data set (Section 2, l. 133-135), but we will try to explain it better in the revised manuscript.

9. Lines 305-322, consider to move that to introduction. Here, you could add description on your method to choose the number of donor catchments.
Reply: We would prefer to keep this section here since it is specific for the choice of the number of donors, which is not the main issue of the manuscript. We are afraid it could divert the readers' attention from the main issue if moved to introduction.

10. Line 289, by running.
Reply: We will correct it, thank you.

11. Lines 366-373, move to methodology section.
Reply: We agree: we will move it to a new methodology section.

12. Line 392, "as anticipated" → "as introduced".
Reply: We will correct it, thank you.

13. Line 402, "so relevant" → "so obvious".
Reply: We will rephrase the sentence.

14. Lines 425-426, 460-471, move to methodology.
Reply: The first lines will be moved to a new methodology section (along with lines 366-373). For what concerns lines 460-471, we consider the identification of the nested catchments and the definition of the percentage threshold as part of the results, since they depend on the case study. Thus, we would prefer to keep them in section 4.3.1.

15. Line 507, KGE and NSE.
Reply: We will correct it, thank you.

16. Lines 528-533, consider to rephrase that. Hard to follow.
Reply: We will simplify and divide the sentences, thank you.

17. Lines 538-540, could you add some discussions on that?
Reply: We believe it is related to the different structure and number/sensitivity of model parameters. As said above, it is of course a matter of strong interest and it would deserve more attention but we prefer to not include further discussion here since it is not the focus of the study. However, we will address again the issue in this section as well.

18. Lines 570-577, consider to move to methodology.
Reply: As prompted also by comment D, we agree that we need to better clarify the sampling procedure already in the methodology section.

19. Lines 685-687, please rephrase that, cannot follow.
Reply: The sentence will be rephrased.

20. Please, consider to reduce the number of figures. Try to aggregate Figures 1 and 2. One of the options is try to present the performance of the two hydrological models in one figure instead of showing separately, such as figures 10-11, 12-13, 14-15. Or you may consider to provide the figures in a supplementary file.
Reply: As suggested, we will aggregate figures 1 and 2, as well as figures 10-11, 12-13 and 14-15. We would prefer to keep them in the main text, since they are needed to interpret the discussion and conclusions.

References used in this response to the reviewer

Chiew, F. H. S.: Lumped Conceptual Rainfall-Runoff Models and Simple Water Balance Methods: Overview and Applications in Ungauged and Data Limited Regions, Geogr. Compass, 4/3, 206–225, doi:10.1111/j.1749-8198.2009.00318.x, 2010.

Crooks, S. M. and Naden, P. S.: CLASSIC: a semi-distributed rainfall-runoff modelling system, Hydrol. Earth Syst. Sci., 11, 516–531, https://doi.org/10.5194/hess-11-516-2007, 2007.

Merz, R., Parajka, J., and Blöschl, G.: Scale effects in conceptual hydrological modeling, Water Resour. Res., 45, W09405, doi:10.1029/2009WR007872, 2009.

Merz, R., Parajka, J., and Blöschl, G.: Time stability of catchment model parameters: Implications for climate impact analyses, Water Resour. Res., 47, W02531, doi:10.1029/2010WR009505, 2011.

Nester, T., Kirnbauer, R., Gutknecht, D., and Blöschl, G.: Climate and catchment controls on the performance of regional flood simulations, J. Hydrol., 402, 340–356, 2011.

Parajka, J., Viglione, A., Rogger, M., Salinas, J. L., Sivapalan, M., and Blöschl, G.: Comparative assessment of predictions in ungauged basins – Part 1: Runoff-hydrograph studies, Hydrol. Earth Syst. Sci., 17, 1783–1795, https://doi.org/10.5194/hess-17-1783-2013, 2013.

Petheram, C., Rustomji, P., Chiew, F. H. S., and Vleeshouwer, J.: Rainfall-runoff modelling in northern Australia: a guide to modelling strategies in the tropics, J. Hydrol., 462–463, 28–41, https://doi.org/10.1016/j.jhydrol.2011.12.046, 2012.

Viney, N. R., Perraud, J. Vaze, J., Chiew, F. H. S., Post, D. A., and Yang, A.: The usefulness of bias constraints in model calibration for regionalisation to ungauged catchments, in: Proceed, 18th World IMACS/MODSIM Congress, Cairns, Australia 13–17 July, 2009.