Reply to Wouter Knoben’s comments

We thank Wouter Knoben for his comments on our manuscript, which will help to improve its overall quality. Below we give our replies (in blue) to his comments (in black) on how we intend to modify the paper to account for his suggestions and recommendations.

A main point of improvement I see is that the manuscript can (should?) account for sampling uncertainty in the KGE scores. For comparative analysis such as this (where differences in KGE scores between various options are used to decide which of the options is better), it is important to get some idea of the uncertainty associated with the scores themselves. An R tool exists that makes estimating these uncertainties for NSE and KGE straightforward from existing time series of observations and simulations (Clark et al., 2021).

More generally, I think the manuscript can be clarified here and there, and possibly streamlined a bit (though I am unsure what to scrap - it just seems to have a large number of figures).

Thank you for this feedback, we will clarify the manuscript where it is needed and test the methodology to account for sampling uncertainty in the KGE scores.

L21: the most **accurate** lumped model

L50: data are used to **derive** model structures

L56: multi-model approach consists **of** using several models

Terminology and syntax will be checked

L85: It's unclear if this sentence refers to Georgakakos et al., Ajami et al., or the DMIP experiment in general.

This sentence refers to Georgakakos *et al.* and Ajami *et al.*; it will be clarified in the manuscript.

L86: An interesting theoretical paper on this subject is Winter & Nychka, 2010: https://link.springer.com/article/10.1007/s00477-009-0350-y

Indeed, thank you for sharing this paper; we will include it in our manuscript.

L88-91: These two sentences seemed oddly placed here. The goal of FUSE, SUPERFLEX and SUMMA is not to create model-averages, but to create modular models where components can be easily swapped. Such modular modeling frameworks are useful, but not required, for the model averaging the rest of this paragraph is about. It may be helpful to rephrase or remove these sentences.

We wanted to extend here to other existing multi-model techniques different from the one used here (i.e. by output combination). We agree that these last sentences of the paragraph are not necessary and can be removed to avoid confusion.

L91: I think this sentence and the next one need to be clarified. In this sentence, does this approach refer to data assimilation (i.e. updating/correcting model simulations with observations) or model coupling (i.e. making multiple models exchange information, such as coupling a hydrologic model to a hydraulic model)? In the next, what exactly was done in these two studies?

The "Super Model" is a dynamic multi-model method coupling several models by making them exchange information throughout the simulation process. The method is based on the modification of the internal states of each model according to the values of corresponding internal states of the other models (one could make here the parallel with data assimilation). This interaction between the models is made through the adding in the differential equations of one model of a term coming from the other model, and vice-versa, making it a somehow hybrid method between data assimilation (of simulated data, not of observations) and coupling. This "Super Model" was tested in climatology first and then applied in hydrology. During the evaluation period, the hydrological "Super Model" performed better than a multi-model based on streamflow combinations (based on the Oudin *et al.*, 2006, approach). However, this approach does not seem to be adapted to all catchments and its difficulty of implementation is a limiting factor. Considering the previous comment, this sentence will also be removed.

L117: Can it be clarified why this choice was made?

We decided to use a deterministic modelling framework because we focus on improving streamflow simulation and not on quantifying its uncertainty. We acknowledge that a probabilistic framework could also be used, but we plan to do this in a second stage in a forecasting context. We think that it also avoids making our paper too complex here. We will shortly justify our choice in the revised manuscript.

L146: I don't understand this. Can this be rewritten?

The catchment selection criterion used is based on streamflow availability. Here, we set a threshold of 10% maximum gaps for each year over the whole period considered (1999-2018). However, this criterion may be slightly too restrictive (e.g., removal of a station installed in 2000 and presenting continuous data since its installation). In order to overcome this problem, we have decided to allow this threshold to be exceeded for a maximum of 3 years over the whole period considered. It is therefore a compromise to obtain a large number of catchments for the study without affecting negatively the models calibration by having a long enough period. We will develop the selection method in the manuscript.

L147: How should "limited human influences" be interpreted here? Some of these catchments are quite large and I find it difficult to believe that these would give (mostly) natural flows.

We agree with you, that is why we did not use the word "natural". In France, the vast majority of catchments have human influences (e.g., dams, dikes, irrigation, urbanization). What we mean by streamflow with limited human influences corresponds to stations where the streamflow records have a hydrological behaviour close to a natural streamflow. We agree that the largest basins are not natural. On the other hand, the large influences (e.g., dams) are implemented mainly on upstream sectors of the river, which means that this influence will tend to smooth out towards the downstream. We will better explain this point in the revised manuscript.

L152: Perhaps this table can be updated with some characteristics on human influence (% urban area, number of dams/diversions, artificial storage volumes, etc)

Though we agree that this information would be useful to evaluate the actual level of influence, we think that it is a bit out of the scope of the study to link the performance of our modelling approach to the fact that catchments are influenced. Therefore, we would prefer not to include this information in the revised manuscript.

L177: It took me a while to understand this approach. It might be good to clarify that in the case of Figure 2, the main catchment (2nd red outline) is used to create 5 different experiments (itself, and those shown in the other four red outlines).

We agree, we will clarify the description of this approach in the manuscript, as well as the figure.

L184: This seems critical, so I'd suggest to explain this procedure in more detail here.

A limitation of semi-distributed hydrological modelling is the computation time required to find the optimal parameters. Strictly speaking, it is necessary to calibrate a routing scheme for each upstream catchment. In order to optimize the computation time, we defined a set of global transfer parameters at the catchment scale, which were then made more specific for each of the upstream sub-catchments according to a physical descriptor. As an example, the simple lag (T) can be determined as the ratio between the hydraulic length (d) and an average velocity parameter optimized (C0): T = d/C0.
Note that more complex routing schemes (e.g., lag&route) were also tested but did not provide significant benefit compared to the simple lag approach (Henrotin, 2022). This procedure will be better explained in the manuscript.

L187: The method, or the outcome of the method (i.e., the estimated set of parameter values)?

We will rephrase it into "a model is defined as a configuration composed of a model structure and an associated set of parameters".

L192-194: Can these changes be found somewhere? Were any tests performed to ensure these changes are correctly implemented?

Several tests have been done to ensure that the changes have been correctly implemented (e.g., consistency of parameters, comparison with daily models). The changes made have been documented but are not yet public (they will be when the airGR+ package will be available on the CRAN). Summary sheets containing the structural scheme of the different models, the pseudo-code and the table of free parameters are however available on request. A few words will be added in the manuscript on these consistency tests.

L194: Where these 13 selected for reasons of convenience (e.g., because they were already implemented) or were there other concerns (e.g., these models are known to perform well in France)?

Several criteria led to the selection of these 13 models. First, for convenience reasons since these models were already implemented at daily time step in our modelling tools. Then for reasons of confidence because these models have already been tested on many catchments in France. Finally, the possibility of adapting them easily to the hourly time step was also a selection criterion. This will be more clearly stated in the manuscript.

L203-205: Is there some sort of reference that can support these statements?

Indeed, we will refer to Thirel et al. (2023) in the manuscript to support these statements.

L206-207: This section is missing details. Things that come to mind:

1. What were the calibration settings (algorithm, computational budget, stopping criteria, etc)?

The algorithm used for calibration comes from Michel (1991) and aims to combine a global and a local optimization approach. First, a gross screening of the parameters space is performed using either a

rough predefined grid or a list of parameter sets. Then a steepest descent local search algorithm is performed, starting from the result of the screening procedure.
A lumped calibration (over 10 years of hourly data) takes about 0.5 to 6 min (depending mainly of the catchment considered and the number of free parameters). The calibration of the 78 modelling options with a split-sample test on a single catchment takes on average half a day.
A single combination takes between 0.1 and 0.2 s. Considering more than one million possible combinations of 4 modelling options among the 78 available, the multi-model takes about 5 days to be determined on a single catchment.
A computing cluster was therefore used in order to process all the catchments simultaneously.
We will add some information on calibration and computation time in the manuscript.

2. Are the models calibrated for each basin individually (i.e. one parameter set per catchment per model per objective function) or over the sample of 121 catchments (i.e. one parameter set per model per objective function)?

The models are calibrated for each catchment individually (i.e. one parameter set per catchment per model per objective function). This will be more clearly stated.

3. Are models calibrated to return a single optimal parameter set or does the manuscript account for parameter uncertainty by looking at multiple parameter sets per calibration exercise?

We did not focus explicitly here on parameter uncertainty, i.e. we did not use multiple parameter sets based on Monte Carlo simulations for example. We agree that this would be interesting to consider, but we thought this would make the article too complex. We will better explain our choice in the article.

L215: This may be more confusing than helpful
We agree, it will be corrected in the manuscript.

L217-222: I'm not sure if this is the right section to bring this up, but it is unclear to me how river flow routing is handled. Judging from Figure 2, there are 2 distinct cases:

- Case 1: a catchment has exactly 1 upstream sub-catchment somewhere along the main river stem (e.g.,, the 3 right-most red outlines in Fig 2). Question: is flow from the upper catchment routed through the lower catchment somehow, or are these flows assumed to magically appear at the outlet of the lower catchment and there combined with simulations from the lower catchment?

- Case 2: a catchment has multiple upstream sub-catchments on different branches of the river (e.g.,, the two left-most red outlines in Fig. 2). In these cases a large part of the white catchment (the lower one) will have travel times for its runoff comparable to travel times in the gray subcatchments. How is routing handled in these cases?

Indeed, in connection with the commentary to L184, the routing procedure deserves more detail in the manuscript. We will improve this.

L245: This practice consists **of** separating

Syntax will be checked.

L251: Please clarify if this means the whole time series (calibration + evaluation), or the whole evaluation part of the time series

Indeed, it is not clear. We mean here by "whole time series" the whole evaluation part of the time series (evaluation over P1 when calibrated on P2 and evaluation over P2 when calibrated on P1). This was in contrast to "event based". We will clarify it in the manuscript.

L265: This figure is very helpful in understanding the methodology w.r.t. the semi-distributed approach as well as the merging. It may be worthwhile to introduce it earlier, in section 2.2 or 2.4 for example.

Thank you for the comment. We will introduce this figure earlier in the text and refer to it more frequently.

L270: It might be easier on the reader to (throughout the manuscript) call these "objective functions" instead of "calibration options", which (I think) stays more in line with existing literature

We agree, we will change it in the manuscript.

L271-272: Over the whole time series or evaluation period only (this is what the y-label says)?

Over the evaluation period, you are right. This will be corrected.

L274: I would strongly suggest to rewrite this part of the text. Using qualitative words such as "(very) good" suggests a level of confidence in these model results that must have a certain objective support. It is known that scores such as NSE cannot easily be compared between catchments (see Schaefli & Gupta, 2007), so what is "good" in one place may not be "good" somewhere else. Instead, to interpret these values some sort of benchmark is needed that gives an idea of what kind of scores other modeling approaches might obtain in these catchments. That should make it obvious how the calibration scores in this manuscript compare. Options could be:

- typical calibration KGE scores in these catchments;
- some form of predictability of the flows in these catchments (see e.g., Schaefli & Gupta, 2007; Seibert et al., 2018; Knoben et al., 2020)

- Schaefli: https://onlinelibrary.wiley.com/doi/10.1002/hyp.6825
- Seibert: https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.11476
- Knoben: https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2019WR025975

Crochemore et al. (2015) compared expert judgement and numerical criteria in order to define the quality of a simulation over several French catchments. On the other hand, we agree that these qualitative words do not indicate an improvement of the simulations and may not be important for the purpose of our study.
Indeed a benchmark is needed in order to test the relevance of a score like the KGE. Here we decided to compare our different multi-models to the best lumped single model on each catchment but the choice could have been different (e.g., taking the lumped single model most often selected as the best model for all catchments). So we will change this to refer to relative level of performance.

L275: The methodology section does not explain how the models were calibrated (which algorithm, which settings, how many iterations, etc.) - can we be certain that the differences in Figure 4 are due to the choice of transformation only, or may the other calibration settings have played a role too?

More information will be given on the calibration settings. It is difficult to be certain that the differences come only from the choice of transformations, given the different sources of uncertainty in the modelling chain. However, we remain confident that these sources provide the main information on the differences.

L277-278: I don't understand this sentence. If the intent is to say that this models struggles to simulate low flows when it is calibrated on the KGE Q+0.5 criterion, then I would suggest to remove the mention of "KGE applied to Q-0.5" - this only adds confusion. Also, what specifically in this structure makes it struggle to represent low flows?

Indeed, the intent is to say that this model struggles to simulate low flows when it has been optimized to represent high flows; we will clarify the sentence in this sense to avoid confusion in the manuscript. The reverse is also true since a model optimized on low flows will have more difficulty representing high flows (e.g., NAM0 or GARD).
TAN0 has stores in series, where each store is a conceptual representation of a type of flow (e.g., surface, sub-surface, intermediate, sub-base and base). The problem encountered with this model during a calibration focused on high flows is the difficulty (or even the impossibility) to store water in the deep stores for the summer period. In case of floods, the flow mainly comes from the surface store, the model tends to generate the total runoff only with the store representing surface/sub-surface flows. Consequently the filling of the deep store becomes low (or non-existent) and the associated parameters insensitive. The presence of null streamflow on an inverse evaluation criterion thus takes enormous weight and drastically decreases the value of the KGE (and a fortiori the composite criterion), despite the adding of a small epsilon value (Pushpalatha *et al.*, 2012).

L279-280: if I understand the methodology section correctly, the "mean KGE" is the mean of the 3 different calibration approaches (line 251), but this section seems to refer to something else that only involves the Q+0.10 results. Suggest to clarify.

The “mean KGE” refer to the composite evaluation criterion. We will find a more explicitly name to avoid confusion in the manuscript. Here, the “Q+0.1” refers to the transformation applied during the calibration.

Figure 4: Why do we not see the best and worst results in each case?

The best and worst results have not been shown for the sake of readability of the figure. As mentioned above, for some catchments, some models show difficulties to represent a range of streamflow different from the one on which they were calibrated (which induces KGEs that can be negative and strongly impacts our composite criterion). We thus wanted to highlight here the overall capacity of the models to reproduce the observed streamflow rather than giving a strong weight in the figure to the outliers (which are however interesting!).

L287: Here also I would recommend to simply drop the word "high". This has an implied meaning of "good", which would need some sort of objective justification. Simply stating facts (e.g.,, "The benchmark models have a median KGE of 0.91, with low variation [..]") should be enough.

We agree; we will make this correction in the manuscript.

L288: What exactly is shown in these histograms? Just eyeballing it seems to suggest that the bars shown for Q-0.5 sum to a lower total than the bars shown for Q+0.5. Shouldn't there be the same total number of models in each of the three categories?

No, the total number of models in each of the three categories should not be the same. In the case of single lumped models, the total number must be equal to the number of catchments (1 model selected on 1 catchment). The histogram answers the following question: How many times each model is selected as the best lumped single model over the 121 catchments? We will clarify this point in the manuscript.

L290: Interesting. I'm not sure if this helps but similar findings can be found in earlier literature and our own recent paper:

Perrin, C., Michel, C., & Andréassian, V. (2001). Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. Journal of Hydrology, 242(3-4), 275–301. https://doi.org/10.1016/S0022-1694(00)00393-0

Knoben, W. J. M., Freer, J. E., Peel, M. C., Fowler, K. J. A., & Woods, R. A. (2020). A brief analysis of conceptual model structure uncertainty using 36 models and 559 catchments. Water Resources Research, 56, e2019WR025975. https://doi.org/10.1029/2019WR025975

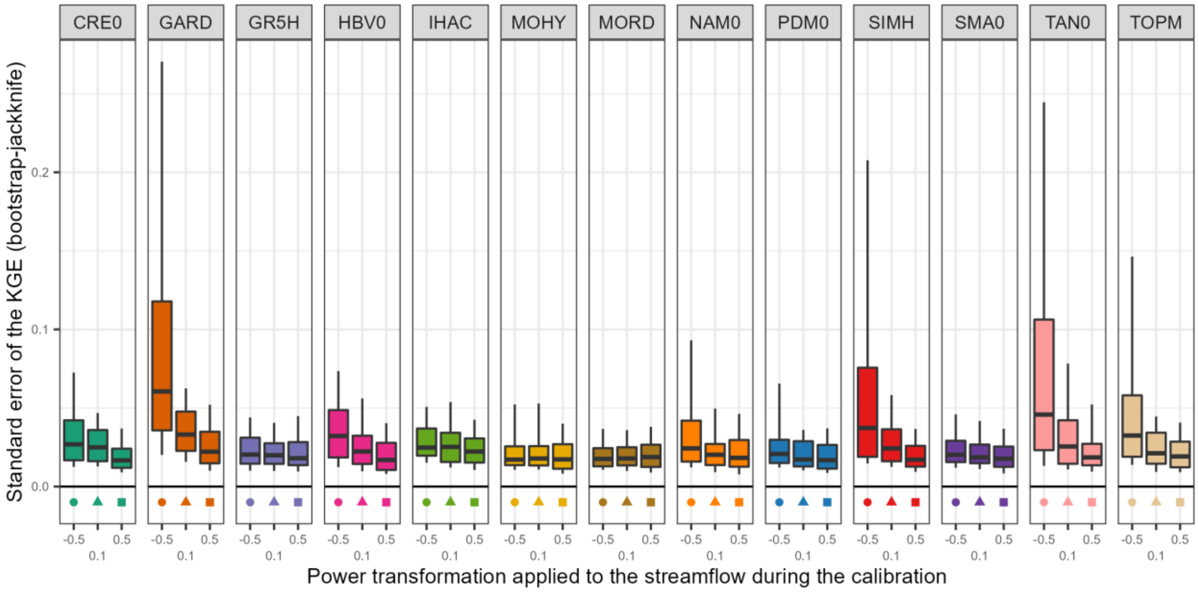
Indeed it was a result we were expecting. We will acknowledge that by referring to these papers.

L295: I believe that this and later sections are incomplete without accounting for the sampling uncertainty in KGE scores. Recent work using several hundreds of catchments in the USA shows that individual timesteps can have disproportionate impacts on KGE (and NSE) scores, and that as a result KGE sampling uncertainty (i.e., uncertainty in the KGE score based on which timesteps are used to compute the value) can be really large: uncertainty varies from ~0.05 KGE points to over 0.50 KGE points, depending on locations (Clark et al., 2021).

Given that this paper uses comparison of KGE scores, quantification of the sampling uncertainty is needed to determine if these KGE differences for different model setups fall within or outside the uncertainty in the scores themselves. The Clark paper also includes an R library that can be used to obtain KGE uncertainty estimates from existing timeseries.

https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2020WR029001

We agree that there is uncertainty in KGE scores. We have already worked on this issue in previous studies that showed the sensitivity of the criteria values to a few time steps in the series (see e.g., Berthet *et al.*, 2010a, b).
As proposed, we applied the bootstrap-jackknife in order to quantify this uncertainty. This methodology was used over our sample of 121 catchments and over our 39 lumped single models (structure/calibration option pairs). The results show that for 90% of the cases, the KGEs have uncertainties lower than 0.06 with a median of 0.02. However, differences can be noted according to the catchments, the structures, the calibration period, the period used to apply the booststrap-jackknife and especially according to the transformations used during the model calibration. Indeed, KGE values from simulations optimized for low flows are more uncertain than those optimized for medium or high flows. We will add few sentences about this topic.



L303: The **right** part of Figure 6

Corrected.

L304-305: This is really hard to follow. Perhaps a step-by-step explanation of what is being done here (and why that is a useful thing to do) can be added.

We will clarify this part in the manuscript.

L305: The **semi-distributed approach** seems

We agree with this change.

L309: I don't think this is necessarily very surprising. KGE scores cap out at 1, so if a lumped model obtains scores already close to 1 there's little room for improvement.

True

L310: Not necessarily helpful, but is it really a limitation if a structure needs some form of semi-distributed approaches to provide accurate simulations? One could argue that structures that don't need that may be too flexible, to the extent that they (somehow) don't need to represent known physical properties of the system to provide streamflow simulations that look similar to observed ones.

This is a possible vision of things indeed. We will try to rephrase this point.

Figure 6: This information is really hard to make out. If this all needs to be one figure, a heatmap may be more appropriate. If it is important to keep the three different calibration approaches visible, three different figures may be more appropriate.

Indeed, the figure can be difficult to understand, we will improve it, as also suggested by the other reviewer.

L331: Maybe I missed it, but why is this expected?

Numerous articles have shown an improvement linked to the multi-model, which is what we expected. On the other hand, the value of the increase was less expected indeed.

L335-337: A reference to the Winter & Nychka paper (here or later) that I mentioned before may be appropriate, as these findings are in line with the argument presented in that paper

We agree.

L397: Out of curiosity, at what epsilon does LMM become acceptable in all catchments? Perhaps a small inset in Figure 13 can be added that shows epsilon on X, and percentage of catchments where LMM can be used on Y

The LMM become acceptable in all catchments at 0.027.

L403: It might help the reader to specifically mention that this means the epsilon = 0 case

We will clarify it.

L406: What is severity? Neither Section 2.5 nor Appendices A or B explain this. It would be good to add a brief explanation of these 5 metrics somewhere.

Severity corresponds to the largest deficit volume during the event (more details can be find in Caillouet et al., 2017). We agree to add a brief explanation on events metrics.

L412: I find it difficult to interpret these (and other) numbers. Are these large or small errors? It might be helpful to add some sort of indication of the range of (i) low flow duration, (ii) mean annual severity, (iii) mean peak duration, (iv) peak streamflow, and (v) mean high flow values so that the reader can place these errors in that context.

We propose to introduce a table to analyse this point more easily.

Figure 14: Should this be "mean annual **low** flow duration difference"?

True, we will clarify it on the figure.

L424: I believe this is the right section to refer back to the Winter & Nychka paper. The key element they highlight is that it's not the number of models in an ensemble, but how different they are that matters. I don't know how to easily quantify model differences, but this view may harmonize these findings and point interested readers into future research directions.

We agree with the relevance of this article in this section.

L523: Seeing how the possible loss in performance can be larger than the average gains we see from using multi-model approaches (0.05 compared to 0.03), it would be good to also quantify the "gain in computation time" mentioned here, to provide some quantitative support for this statement.

We will add a few sentences on computation time in the manuscript.

L535: Possibly, but an important nuance here is that these two models are selected from a much larger set (13 in total, right?). This means that two models may be sufficient to get large increases in simulation accuracy but it cannot be just any two models - they need to be carefully selected.

This is correct. This is also what we showed in Section 4.1 where the best combination of 4 fixed models (combinations of 2 to 4 among 4 models) does not reach the performance of 4 free models (combinations of 2 to 4 among 78 models). If we want to be even more critical, we could argue that a combination of 2 free models have a higher gain than a combination of 4 fixed models.

References

Berthet, L., Andréassian, V., Perrin, C. & Loumagne, C. (2010) How significant are quadratic criteria? Part 1. How many years are necessary to ensure the data-independence of a quadratic criterion value? *Hydrological Sciences Journal* **55**(6), 1051–1062. Taylor & Francis. doi:10.1080/02626667.2010.505890

Berthet, L., Andréassian, V., Perrin, C. & Loumagne, C. (2010) How significant are quadratic criteria? Part 2. On the relative contribution of large flood events to the value of a quadratic criterion. *Hydrological Sciences Journal* **55**(6), 1063–1073. Taylor & Francis. doi:10.1080/02626667.2010.505891

Caillouet, L., Vidal, J.-P., Sauquet, E., Devers, A. & Graff, B. (2017) Ensemble reconstruction of spatio-temporal extreme low-flow events in France since 1871. *Hydrology and Earth System Sciences* **21**(6), 2923–2951. Copernicus GmbH. doi:10.5194/hess-21-2923-2017

Crochemore, L., Perrin, C., Andréassian, V., Ehret, U., Seibert, S. P., Grimaldi, S., Gupta, H., et al. (2015) Comparing expert judgement and numerical criteria for hydrograph evaluation. *Hydrological Sciences Journal* **60**(3), 402–423. Taylor & Francis. doi:10.1080/02626667.2014.903331

Henrotin, E. (2022) Quelle fonction de propagation choisir pour relier les différentes mailles d’un modèle hydrologique semi-distribué ? (Mémoire de Master 2). Master Sciences de l’Eau et de l’Environnement de l’Université de Tours et INRAE Antony.

Michel, C. (1991) *Hydrologie appliquée aux petits bassins ruraux*, Vol. Hydrology handbook (in French). Cemagref, Antony, France.

Oudin, L., Andréassian, V., Mathevet, T., Perrin, C. & Michel, C. (2006) Dynamic Averaging of Rainfall-Runoff Model Simulations from Complementary Model Parameterizations. *Water Resources Research* **42**. doi:10.1029/2005WR004636

Pushpalatha, R., Perrin, C., Moine, N. L. & Andréassian, V. (2012) A review of efficiency criteria suitable for evaluating low-flow simulations. *Journal of Hydrology* **420–421**, 171–182. doi:10.1016/j.jhydrol.2011.11.055

Thirel, G., Santos, L., Delaigue, O. & Perrin, C. (2023) On the use of streamflow transformations for hydrological model calibration. *EGUsphere* 1–26. Copernicus GmbH. doi:10.5194/egusphere-2023-775