Thank you very much for your very thorough review. We provide below some answers to the reviewer's remarks.

Major comments

1. Methods: I think that showing the impact of transformations on ranks, rather than on the actual absolute errors that were used to generate those ranks, may hide the real effect that the choice of mathematical functions has on streamflow simulations and, more importantly, may distort a lot the differences in performance (and their perception) among the various types of transformation. For example, what is the difference in mean absolute error between rank 1 and 10? I encourage the authors to show the effects of transformations more directly; for example, they could use a normalized mean absolute error for different streamflow categories, to make the results comparable among catchments.

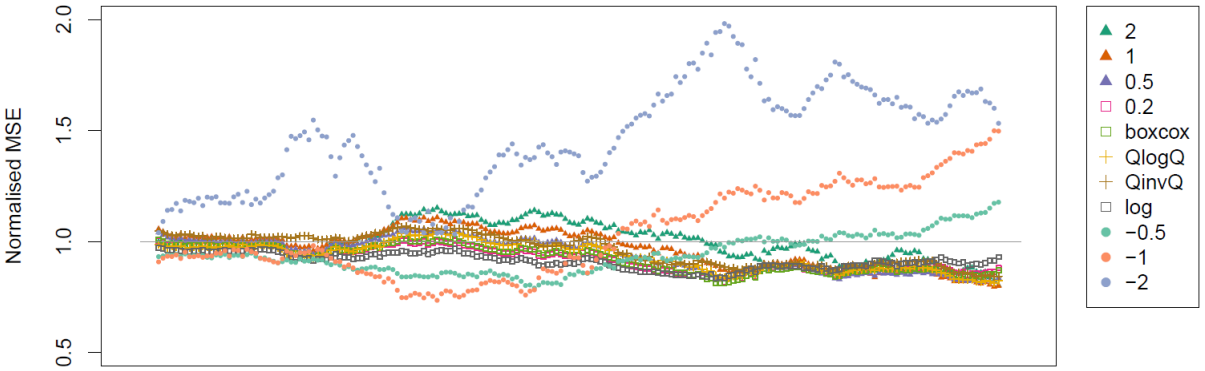
A1: We thank the reviewer for this interesting comment. The idea behind choosing to work with ranks, instead of direct (normalised) errors was to i) be less impacted by different orders of errors magnitudes between catchments or ranges of streamflows, and to ii) answer the question of what are the best transformations, rather than how good transformations are. We recognize that assigning ranks can have the effect that a rank difference of 1 can both signify a small error or a larger one. We however want to stress out that ranks are accounted for time step by time step, meaning that if differences between two simulations are very low, there can be changes in the order quite easily, which then results in similar average ranks over the intervals.

In order to provide some food for thoughts, we processed as suggested by the reviewer:

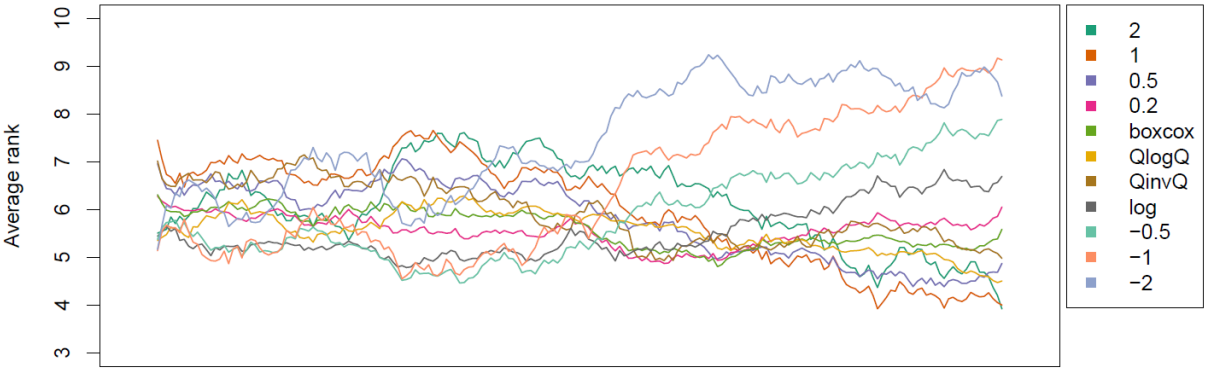
1. the hydrological model is calibrated against observed streamflows for a catchment and with a given objective function, successively with different transformations,
2. for each time step, the absolute error is calculated for the simulations obtained with the nine (or 11) transformations,
3. the time series of daily errors are sorted according to the sorted observed streamflow time series,
4. the sorted errors are aggregated over 200 sequential intervals of an equal number of time steps to smooth the results and facilitate the visual analysis.
5. the aggregated sorted errors are normalised by the average of errors over the nine (or 11) transformations, interval by interval.

Applying that to the same example station as in the manuscript, using GR4J calibrated with NSE and results shown over the calibration period, leads to the following plot (which is similar to Figure 6 of the article; please note we use here the new representation as introduced in answer A20 to the reviewer's remarks):

Figure 6 with calculations as suggested by the reviewer:



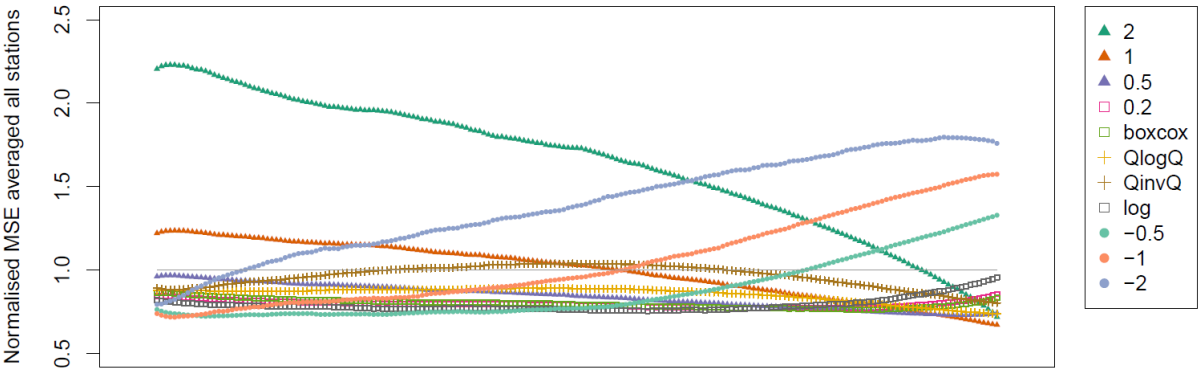
Actual Figure 6 from the submitted manuscript:



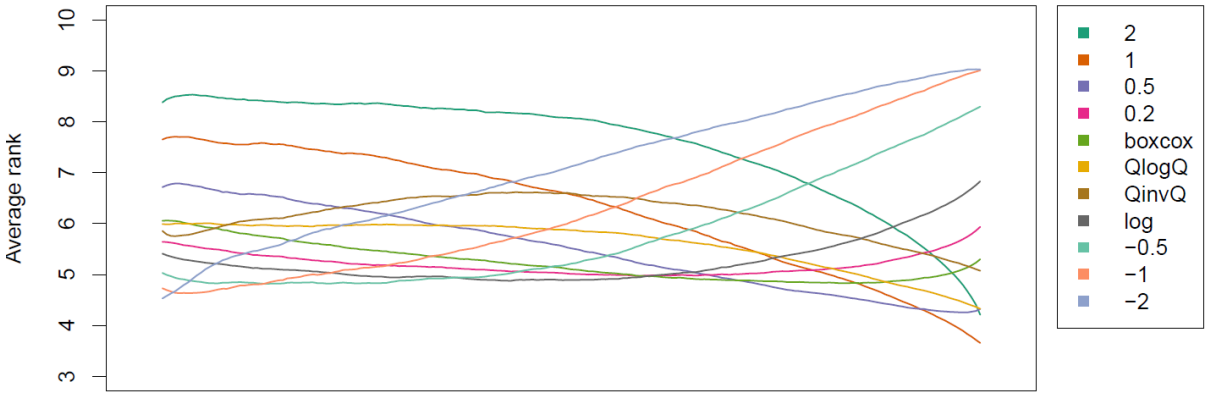
While the general shape of these curves are similar to the one prepared with ranks, it is clear in this new figure that some discrepancies are visible: transformation -2 seems rather far from other transformations most of the time. Another notable difference is that many transformations seem to remain close together for most of streamflow ranges. While this could lead to the conclusion that these transformations can be used interchangeably because they seem to lead to very similar errors, we must note that the very high normalised MSE of one or two transformations leads to smaller differences for the other transformations. In other words, this procedure is impacted by the large error of some transformations and leads to less informative results.

We then automatized this procedure to all 325 stations, with GR4J calibrated with NSE and results shown over the calibration period, and averaged the normalized errors, which gives the following plot (which is similar to Figure 9 of the article):

Figure 9 with calculations as suggested by the reviewer:



Actual Figure 9 from the submitted manuscript:



Here again, we find some similar results as when working with ranks: the general conclusions are not changed. The same groups of transformations still are the best over the same ranges of streamflows.

As a consequence, we prefer not to change the whole methodology used in the manuscript and we will stick to the one proposed so far. We will however mention this other option and discuss it in the revised version.

Additional suggestions to make their analysis more impactful:

- Since NSE is formulated as a function of the sum of squared errors, the authors could report the fractional contribution of the total squared error for the 1, 10, 100, 1000 largest error days obtained with the various transformations (see Figure 10 in Newman et al. 2015). This could provide quantitative support to some statements that the authors make (e.g., L192-193, L338) referring to the number of days where a specific transformation has more weight.

A1.2: We thank the reviewer for this suggestion.

We therefore calculated the fractional contribution to the squared error for the various experiments, to verify if the assertions we made were sound.

First, we want to stress out that the methodology proposed by the reviewer and by Newman et al. (2015) is strictly valid only when we calibrate a model with the NSE objective function, and non-composite transformations. Indeed, NSE and MSE relate to a linear function, as shown by Gupta et al. (2009), in their equation 2:

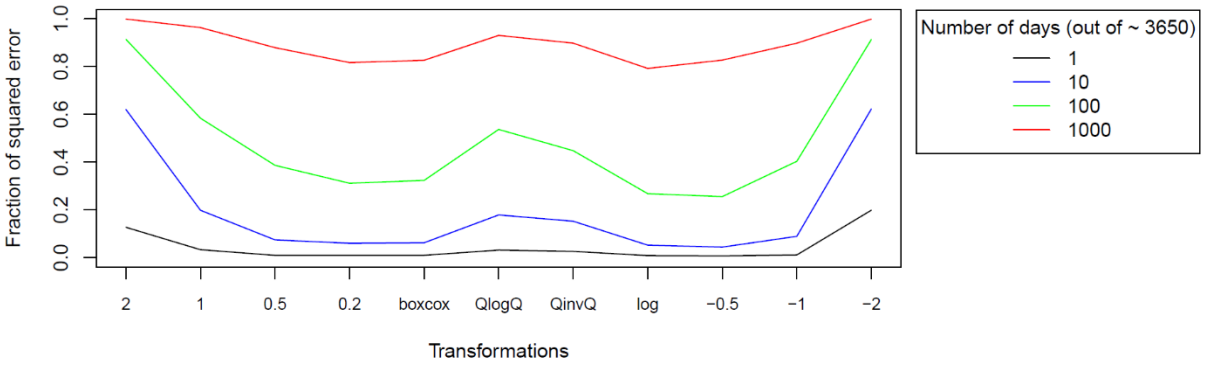
$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{n}(x_{s,t} - x_{o,t})^2}{\sum_{t=1}^{n}(x_{o,t} - \mu_o)^2} = 1 - \frac{\mathrm{MSE}}{\sigma_o^2}$$

Therefore, NSE relates to the squared error (SE) with a linear function too, and the fractional contribution to the SE, for a given time step t1, as written in the following equation, corresponds to the contribution to the NSE:

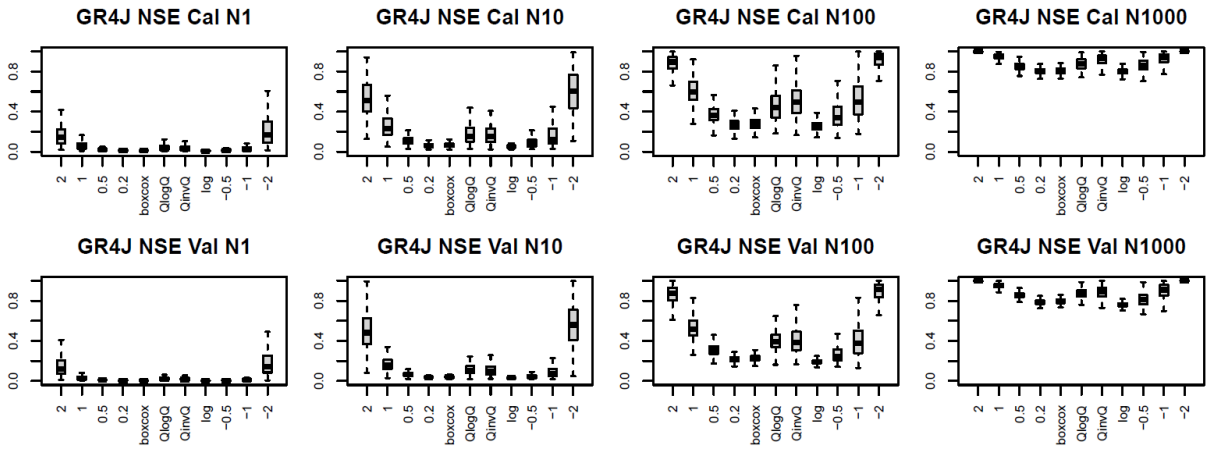
$$\text{Frac\_contr}(t1) = \frac{(\text{Qobs}(t1) - \text{Qsim}(t1))^2}{\sum_{t=1}^{n}(\text{Qobs}(t) - \text{Qsim}(t))^2}$$

In addition, for combined transformations QinvQ and QlogQ, we considered the fractional contribution of a time step as the average of the fractional contributions of the two transformations (1 and -1, or 1 and log transformations).

This leads to the Figure below for illustrating our assertion of lines 192-193. This figure gives the fractional contribution of squared error for the 1, 10, 100, 1000 days with the most error, for the 11 transformations for GR4J calibrated with NSE on the Fecht River, and over the calibration period. This illustrated very clearly that for transformations 2 and -2, there is a large weight on very few days for the objective function calculation. For instance, more than 60 % of the contribution rely on 10 days for these transformations, whereas it is lower than 20 % for other transformations. This figure will be added as a supplementary material to justify our assertion.



This analysis was extended to all 325 catchments and is presented in the figure below. In this figure, we still use the GR4J model calibrated on NSE. N1 to N1000 represent the number of time steps having the highest fractional contribution. Cal and Val mean respectively calibration period and validation period. It is here again very clear that extreme transformations rely on a more limited number of time steps than other transformations. This figure will be added as a supplementary material to justify our assertion. Similar results are obtained with GR5J and GR6J, but they will not be shown in the supplementary material.



Finally, as the KGE cannot be written as a linear function of MSE, there is no such straightforward relationship between the term above and the objective function, when the objective function is the KGE or KGE'. Still, we produced these analyses also for KGE and KGE'-based calibrations. As they lead to very similar plots, we chose not to add them in the supplementary material.

- Show the impact on some streamflow characteristics (e.g., Pool et al. 2017), also known as hydrological signatures (e.g., Addor et al. 2018; McMillan 2020).

A1.3: Thank you for your comment. In addition to the already-present mean annual streamflow and baseflow index signatures, we will add the central slope of the flow duration curve as recommended by reviewer 1, but also the aridity index and the center mass of annual runoff (see answer A25).

2. In my opinion, some figures are incredibly complex (e.g., Figures 5 and 8), making the communication of the main messages unnecessarily cumbersome. What do the numbers 1 to 11 represent? Are they related to the number of transformations? Figures 6, 9, 10, 11 and 12 are better to show inter-method differences, though these could (should?) show results of actual mean absolute errors. Additionally, Figure 10, 11 and 12 could be merged into one to facilitate the comparison (the same comment applies to Tables 3, 4 and 5).

A2: We apologize for those complex figures, which we believed would provide additional information to other figures. The numbers 1 to 11 represent the ranks, as written in the y-label and in the caption. They are therefore indeed related to the number of transformations, as the best transformation is ranked first, and the worst is ranked 11th. Figures 6, 9, 10, 11 and 12 show the average ranks. It means that they show an aggregated information compared to Figures 5 and 8, which rather show the distribution of ranks. If the

reviewer believes that Figures 6, 9, 10, 11 and 12 are sufficiently informative, we will stick to these ones and remove Figures 5 and 8.

We could merge Figures 10, 11 and 12. However, it means that we would have to reduce their size and therefore their readability would be worse. In addition, and maybe more important, the reader will have to go back and forth in its reading, which could make it uncomfortable. The same goes for the Tables 3, 4 and 5. For these reasons, we would prefer not to merge them.

Minor comments

3. L9-10: "…can sometimes be different from what could be expected…". I recommend the authors avoid including vague sentences like this throughout the manuscript, especially in the abstract.

A3: We propose the following: '… can sometimes be different from their expected behaviour".

4. L19-20: From my view, there is general consensus in the community that no universal hydrological model structure exists, since each one is an assembly of hypotheses on the functioning of a specific hydrological system (Clark et al. 2011). This has motivated a proliferation of flexible modeling platforms such as FUSE (Clark et al. 2008), SUPERFLEX (Fenicia et al. 2011), Noah-MP (Niu et al. 2011), SUMMA (Clark et al. 2015a,b, 2021), MARRMoT (Knoben et al. 2019), Raven (Craig et al. 2020) and even airGR with its variants GR5J and GR6J. I think this is a good place to make this point.

A4: We do agree about this consensus and this is what we tried to express here. We will rephrase with the reviewer's proposition 'there is consensus in the community that no universal hydrological model structure exists' and we will add the references the reviewer provides.

5. L24: This is a good place to cite previous studies showing the impact of subjective calibration criteria selection on hydrological modeling applications (e.g., Mendoza et al. 2016; Fowler et al. 2018; Melsen et al. 2019).

A5: Thank you for these suggestions, we will cite these references.

6. L33: I think you should refer to Figure 1a.

A6: We do not understand this comment, as we wrote ‘This is illustrated in Fig. 1, where in panel a, the larger errors’.

7. L52-58: I suggest citing these studies in chronological order.

A7: We will modify the order of sentences to cite these studies in a chronological order.

8. Figure 1: I suggest including the model being used and the simulation year in the figure caption.

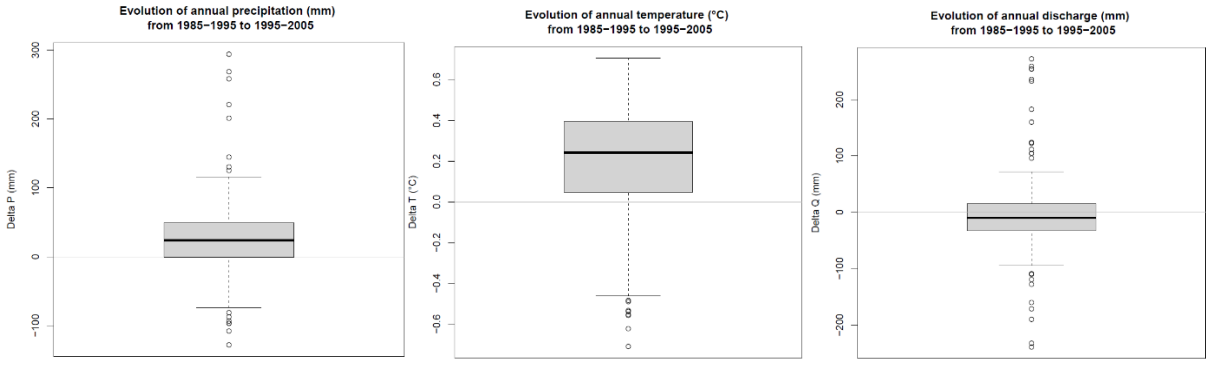
A8: We did not specify the model and simulation year as we believed that this information was not useful here, as what we wanted to show was not actual events, but the behaviour of streamflow transformations. However, we will add it in the revised version of the manuscript.

9. L70-72: This sentence is very confusing. "Alteration" may be interpreted by some readers as human intervention. I suggest re-wording.

A9: Thank you. We propose “with specific streamflow selection procedures such as…”.

10. L82-83: Did the authors examine whether the calibration and evaluation periods are hydroclimatically different? Please clarify.

A10: We wrote in lines 87-88, “This table also shows that the climatic conditions are similar between the two periods, with the evaluation period being only slightly warmer and wetter than the calibration period ». To further investigate potential hydroclimatic differences between the two periods, we propose below boxplots showing the difference of annual precipitation (left), air temperature (middle) and discharge (right) between the two periods. The boxplots are composed of 325 values, i.e. one for each catchment. These boxplots confirm our assertion as the differences for the median are limited for precipitation and air temperature and very low for discharge. Some catchments show larger differences, but those are in a limited number. The boxplots are shown below but will not be included in the manuscript.



11. I think that much of the text in section 2.2 corresponds to methodology, and therefore should be included in the methods section.

A11: We understand the reviewer's concern. We propose some renaming of sections and subsections, as follows:

2. Material and method

2.1. Catchment set and data

2.2. Hydrological model

2.3. Optimization criteria

2.4. Streamflow transformations

2.5. Evaluation methodology (former section 3 Methods)

Then section 4 Results becomes section 3 Results.

12. L98: Can you please clarify how you determined snowfall occurrence in your basins?

A12: We made the approximation that if the daily temperature is below 0 °C, then this is snowfall (as in Knoben et al., 2018). We will clarify this in the manuscript.

13. L101: Why did you choose five elevation bands and not more/less? Did you try other configurations? I think this needs a proper justification, given the large effects that this decision may have on simulated states and fluxes (Murillo et al. 2022).

A13: This choice was based on the previous works led by our colleagues in the past, who assessed the added value of using different numbers of elevation bands and concluded that this number represents a good compromise between time calculation, model efficiency and data quality (see Valéry et al., 2014 and Valéry, 2010). Since the focus of the article is not on

this issue, we think that the cited reference provides sufficient information to the reader to justify this choice.

14. Table 1: I think it would be more informative to show these attributes as maps with a color bar (see, for example, Addor et al. 2017; Alvarez-Garreton et al. 2018).

A14: We made some tests to investigate whether it was possible to replace this table with a figure as requested by the reviewer. We show in the answer A25 these maps. In order to keep the number of figures limited, as requested by the reviewer, and because we believe that maps are less readable than the table, these figures will be inserted as a supplementary material.

15. L111: please specify whether your simulations consider a spin-up period.

A15: We definitely used a spin-up period. A 1-year period, corresponding to the year preceding the calibration or the evaluation period, was used. We will add this information.

16. L160-163: I think this text should be in the methods section.

A16: We saw this text as an introduction about how the results will be presented in the following. As the reviewer thinks that this is clearer to put this information in the Methods section, we will move it there.

17. Figure 3d: the numbers in the y axis are not legible.

A17: Sorry for that, the numbers were somehow cut during the production process, we will correct that.

18. Figure 5 (caption): is CemaNeige implemented in this basin?

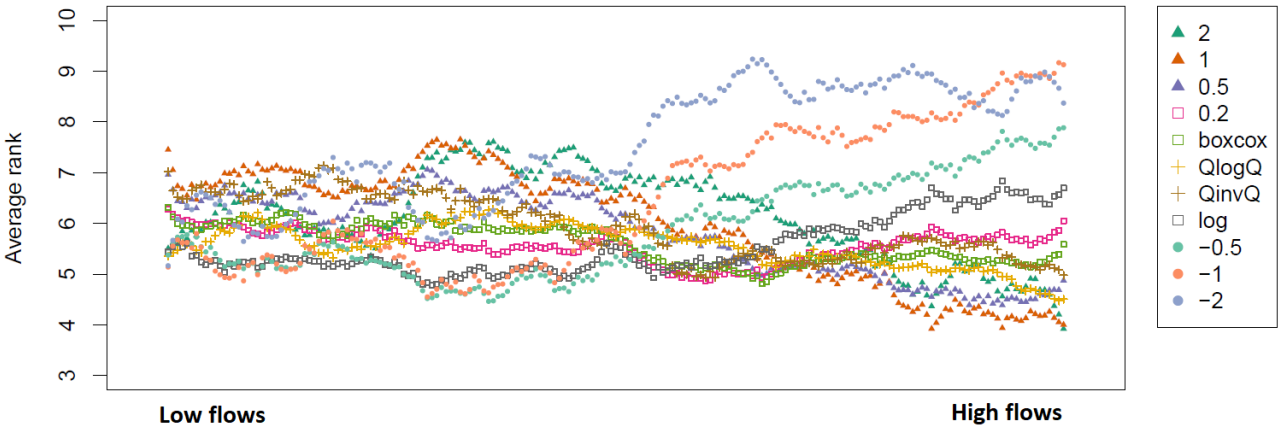
A18: No, we did not use CemaNeige here. We will specify it in the caption.

19. L185: 'average rank of transformations'. How do you compute that average?

A19: Each value of each curve is simply the average of the ranks of transformations over all the time steps of the concerned interval. We will specify it in the text.

20. L185-190: all these comparisons are very hard to visualize. You could use symbols in Figures 6 and 9 to 12 to help readers to see what you want them to see. For example, use X for negative transformations, square for log, circle for Box-Cox, etc.

A20: Thank you for this suggestion. We propose the following visualization, here replacing Fig. 6 of the manuscript. All similar figures will be replaced with this visualization.



21. L261: this sentence is unclear. What do you mean?

A21: We mean that over the calibration periods, the transformations can lead to simulations that are relatively good for their supposed target (e.g. transformation -2 has a low average rank over low flows), but since this average rank is higher for the evaluation period, we can consider that the transformations are less specific, i.e. worse for their supposed target, and closer to each other. We will rephrase and we propose the following:

'To phrase it differently, over the evaluation period, the transformations lead to simulations that are less specific, i.e. closer to each other.'

22. L279: 'to behave much worse'. Note that you are judging based on the ranks, and not on the actual sum of absolute errors. I think it would be much more honest if you showed the latter.

A22: We understand the concern of the reviewer. We will rephrase this sentence to be fairer with what is actually shown. We propose the following: "which appears to show much worse ranks than…".

23. L296: You have ranks for 9/11 transformations. Did you obtain the same number of correlations?

A23: In this section, we discuss results for the GR4J model calibrated with NSE, which means that we have 11 ranks. Correlations were also calculated for GR4J calibrated with KGE, resulting indeed into 9 correlation values, but we did not discuss these results as no further informative result arose.

24. L301: Are these correlations statistically significant?

A24: We thank the reviewer for this suggestion.

First, we must mention that we calculated the correlations for the new characteristics suggested by the reviewers (see the Table provided in A25). Unfortunately, no outstanding correlations were found. We observed that the central slope of the flow duration curve shows correlations of the same order of magnitude, but opposite to the one shown by the BFI, with exactly the same transformations. This indicates that these indicators are somehow well related.

Second, the p-values were calculated for all the calculated correlations using the stats::cor.test() function in R. We found that the correlations lower than -0.11 and higher than 0.11 were all significant (i.e. p-values < 0.05), whereas none of the other ones were significant. Consequently, this does not change the related analysis.
We will add this information in the revised manuscript.

25. Section 4.3: I suggest the authors adding to their analysis the aridity index, the seasonality of aridity (Knoben et al. 2018) and maybe the center of time of runoff (Stewart et al. 2005).

A25: We calculated these indicators as suggested (in addition to flow signatures already mentioned) and added them to our analysis and to Table 1. We assume that the “center of time of runoff” mentioned by the reviewer corresponds to the “timing of the center of mass of the annual runoff” as introduced by Stewart et al. (2005).

Table 1 will therefore be modified as shown below. We can see that for the four new indicators the catchments show some variability between each other, but also that the two periods seem to face rather similar conditions, as was already the case for other indicators. This will be mentioned in the paper.

Characteristic	Period	Minimum	Median	Maximum
Surface area [km^2]	-	5.3	225.5	13 483.5
Min. altitude [m a.s.l.]	-	6.0	209.0	2 154.0
Median altitude [m a.s.l.]	-	53.0	368.0	2 741.0
Max. altitude [m a.s.l.]	-	93.0	784.0	3 997.0
Median slope [deg]	-	1.1	7.4	35.8
Median hydraulic length [km]	-	2.1	19.0	200.7
Artificial land cover [%]	-	0.0	2.1	18.2
Agricultural land cover [%]	-	0.0	54.2	97.7
Forest land cover [%]	-	0.0	43.5	100.0
Mean annual precipitation [$mm\ y^{-1}$]	Calibration	651	1 009	2 204
	Evaluation	691	1 025	2 077
Fraction of solid precipitation [%]	Calibration	0.3	2.5	59.1
	Evaluation	0.0	2.2	50.3
Mean air temperature [°C]	Calibration	-1.1	10.0	13.9
	Evaluation	-0.9	10.3	14.2
Mean annual potential evapotranspiration [$mm\ y^{-1}$]	Calibration	252	661	858
	Evaluation	267	678	871
Mean annual streamflow [$mm\ y^{-1}$]	Calibration	101	405	2485
	Evaluation	123	410	2250
Baseflow index [−]	Calibration	0.01	0.22	0.68
	Evaluation	0.01	0.23	0.76
Aridity index [−]	Calibration	0.03	0.33	0.74
	Evaluation	0.01	0.33	0.77
Aridity seasonality [−]	Calibration	0.69	1.33	1.64
	Evaluation	0.62	1.36	1.72
Center mass of annual runoff [doy]	Calibration	117	152	248
	Evaluation	113	145	244
Central slope of the flow duration curve [−]	Calibration	0.39	1.05	5.05
	Evaluation	0.40	1.01	5.26

Table 1: Same as in the article, with the addition of the suggested characteristics

In addition, these characteristics will be represented as maps, and provided in the supplementary material.

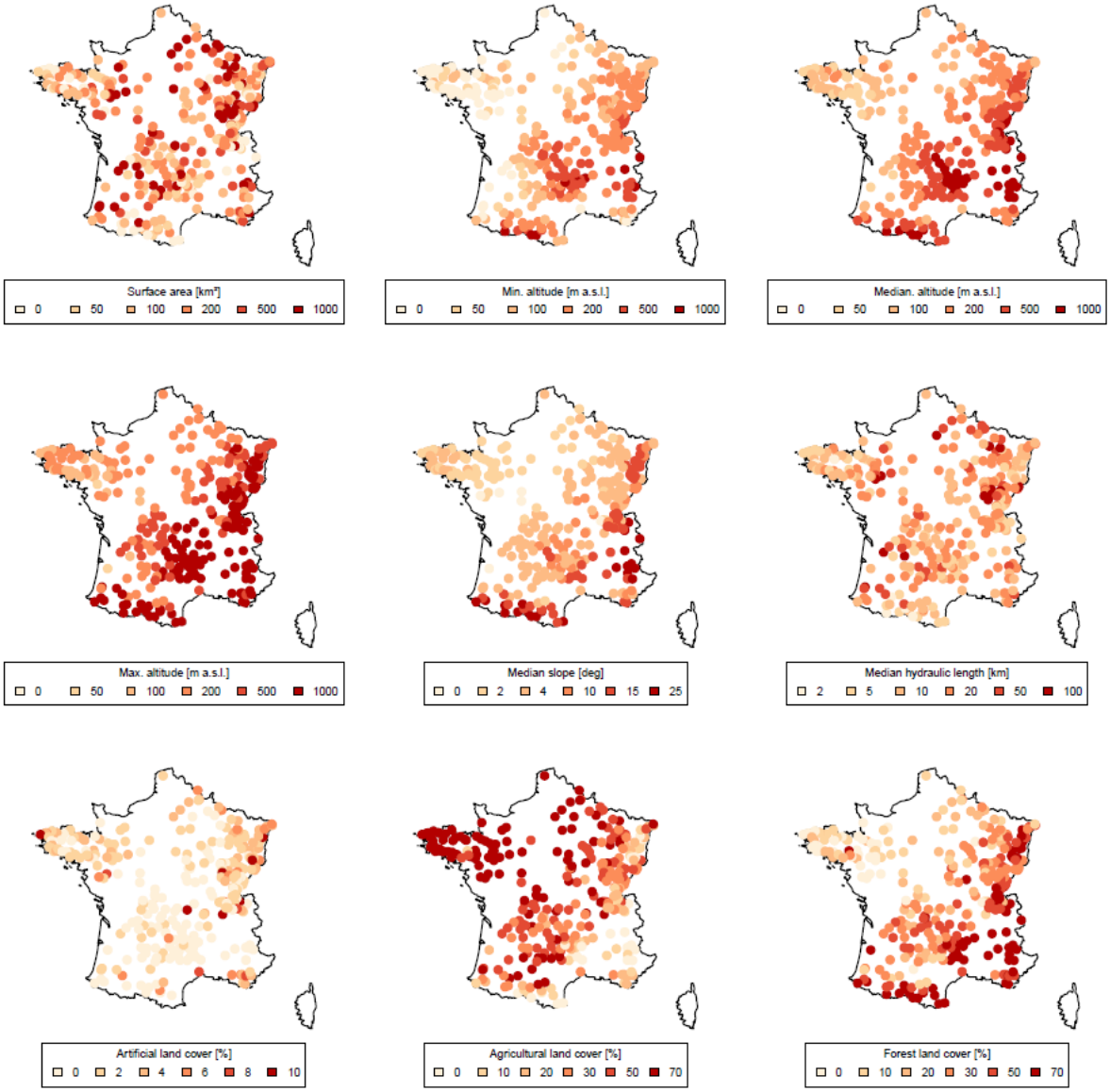


Fig: Maps of physical characteristics

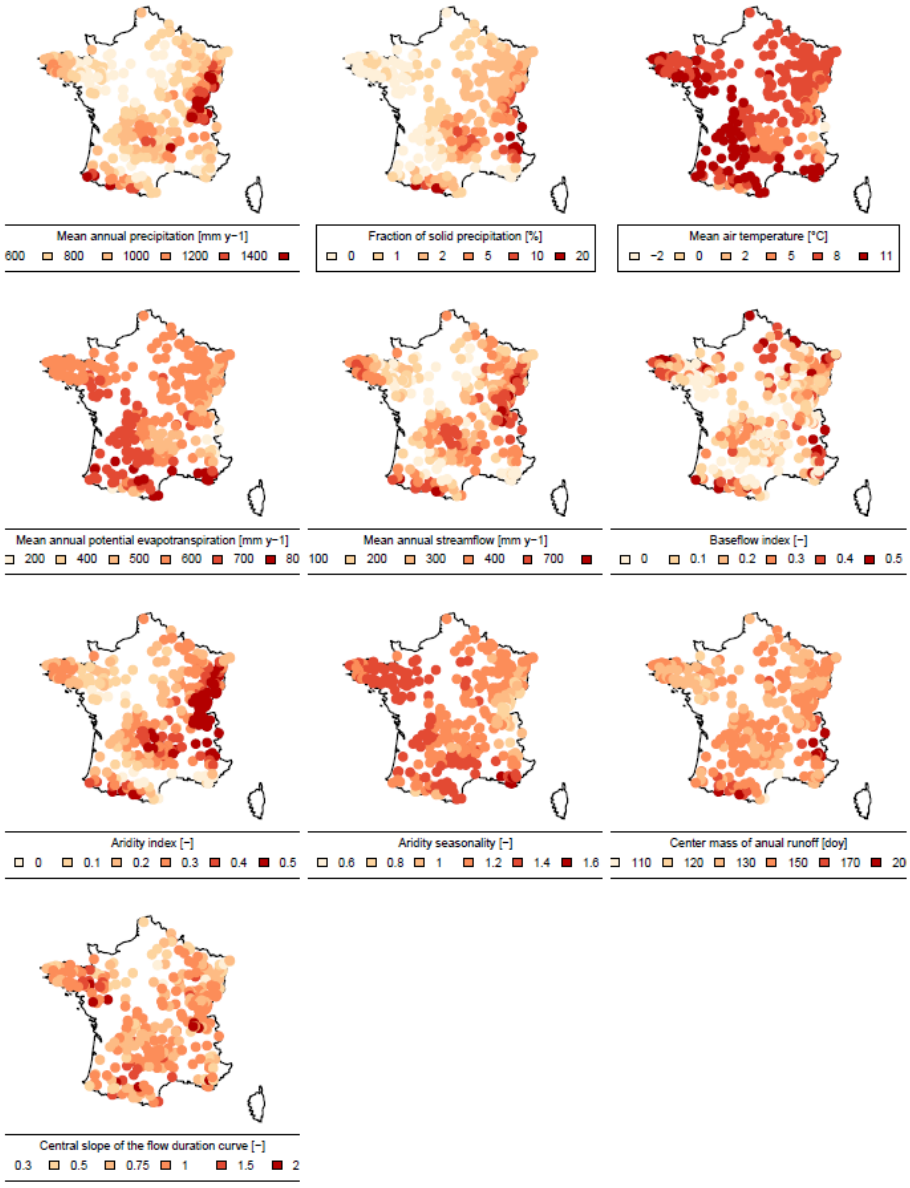


Fig: Maps of indicators on the calibration period

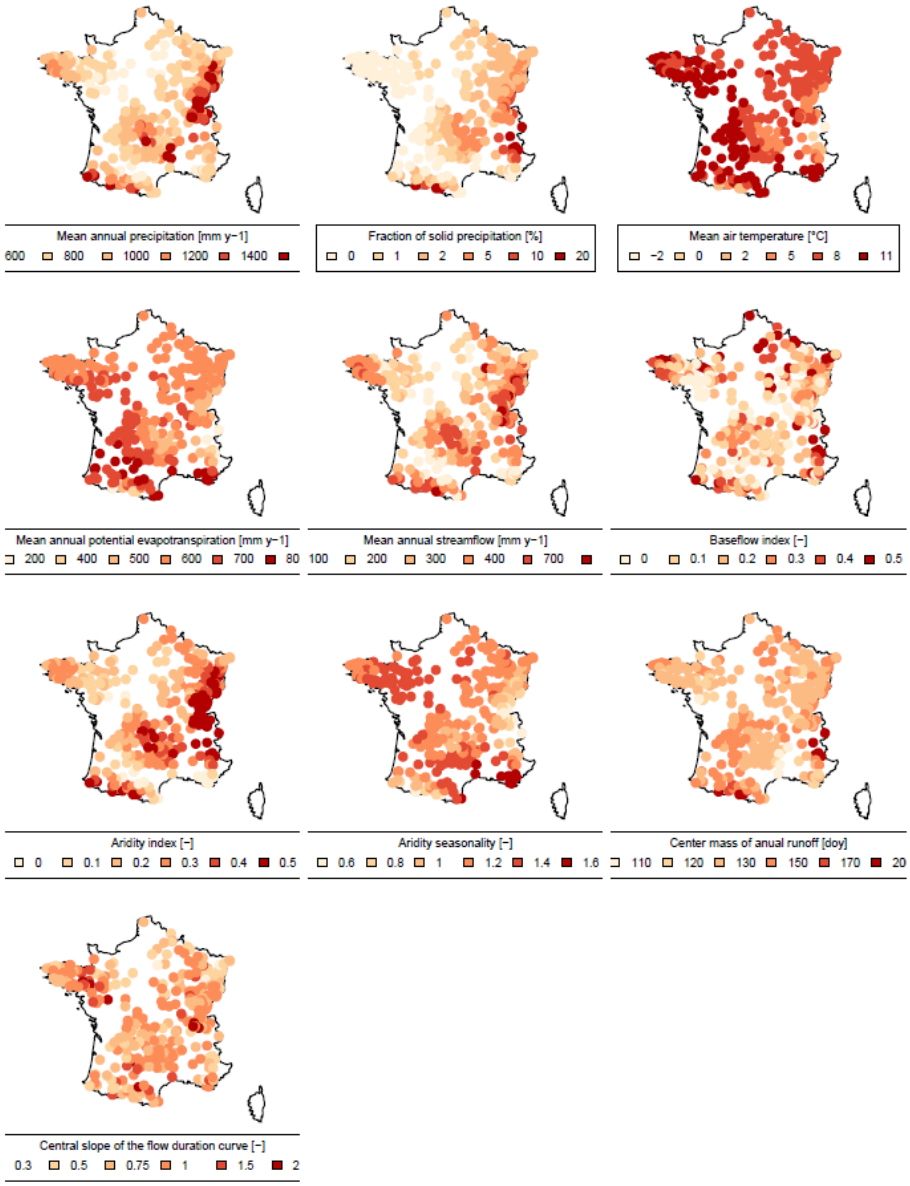


Fig: Maps of indicators on the evaluation period

Some suggested edits

26. L30: ‘have been’ -> ‘has been’ (‘a wide panel’ is singular).

27. L36: I suggest deleting ‘more specifically’.

28. L44: ‘some other works’ -> ‘other studies’.

29. L48-49: delete ‘Nevertheless, some authors tried to investigate this issue. For instance,’.

30. L59: ‘Still, most of the time’ -> ‘To the best of our knowledge’.

31. L59: ‘are not’ -> ‘have not been’.

32. L61-62: I strongly encourage the authors to write that finding with their own words instead of quoting.

33. L63 and anywhere else: I recommend the authors using past tense (i.e., ‘used’ and ‘justified’) when referring to previous studies.

34. L68: ‘tends to illustrate’ -> ‘illustrates these assertions to some degree’. Delete ‘we feel that’.

35. L69: delete ‘in this article’.

36. L75: ‘Data’ -> ‘We used data from…’. I strongly motivate the authors to use active voice.

37. L95: ‘Maximal’ -> ‘Maximum’.

38. L101: ‘take into account the catchment heterogeneity’ -> ‘consider intra-catchment variability’.

39. L103: delete ‘while GR4J is the main model used’ and write ‘In this work, we also use the GR6J model to assess the transferability...’.

40. L124: ‘with N the total number’ -> ‘being N the total number’.

41. L131: ‘as this focuses’ -> ‘as it focuses’.

42. L300: Delete ‘Unfortunately, only a few correlations could be identified’.

43. L301: ‘Anti-correlations’ reads really awkward. I suggest writing ‘negative correlations’ instead.

A26-43: we will consider all these edits during the revision phase, thank you for suggesting them.

References

Addor, N., A. J. Newman, N. Mizukami, and M. P. Clark, 2017: The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrol. Earth Syst. Sci.*, doi:10.5194/hess-21-5293-2017.

Addor, N., G. Nearing, C. Prieto, A. J. Newman, N. Le Vine, and M. P. Clark, 2018: A Ranking of Hydrological Signatures Based on Their Predictability in Space. *Water Resour. Res.*, **54**, 8792–8812, doi:10.1029/2018WR022606.

Alvarez-Garreton, C., and Coauthors, 2018: The CAMELS-CL dataset: Catchment attributes and meteorology for large sample studies-Chile dataset. *Hydrol. Earth Syst. Sci.*, **22**, 5817–5846, doi:10.5194/hess-22-5817-2018.

Clark, M. P., A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener, and L. E. Hay, 2008: Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. *Water Resour. Res.*, **44**, W00B02, doi:10.1029/2007WR006735.

——, D. Kavetski, and F. Fenicia, 2011: Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resour. Res.*, **47**, W09301, doi:10.1029/2010WR009827.

Clark, M. P., and Coauthors, 2015a: A unified approach for process-based hydrologic modeling: 1. Modeling concept. *Water Resour. Res.*, doi:10.1002/2015WR017198.

——, and Coauthors, 2015b: A unified approach for process-based hydrologic modeling: 2. Model implementation and case studies. *Water Resour. Res.*, doi:10.1002/2015WR017200.

Clark, M. P., and Coauthors, 2021: The numerical implementation of land models: Problem formulation and laugh tests. *J. Hydrometeorol.*, **22**, 1627–1648, doi:10.1175/JHM-D-20-0175.1.

Craig, J. R., and Coauthors, 2020: Flexible watershed simulation with the Raven hydrological modelling framework. *Environ. Model. Softw.*, **129**, 104728, doi:10.1016/j.envsoft.2020.104728. https://doi.org/10.1016/j.envsoft.2020.104728.

Fenicia, F., D. Kavetski, and H. H. G. Savenije, 2011: Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. *Water Resour. Res.*, **47**, W11510, doi:10.1029/2010WR010174.

Fowler, K., M. Peel, A. Western, and L. Zhang, 2018: Improved Rainfall-Runoff Calibration for Drying Climate: Choice of Objective Function. *Water Resour. Res.*, **54**, 3392–3408, doi:10.1029/2017WR022466.

Knoben, W. J. M., R. A. Woods, and J. E. Freer, 2018: A Quantitative Hydrological Climate Classification Evaluated With Independent Streamflow Data. *Water Resour. Res.*, **54**, 5088–5109, doi:10.1029/2018WR022913. https://onlinelibrary.wiley.com/doi/abs/10.1029/2018WR022913.

——, J. E. Freer, K. J. A. Fowler, M. C. Peel, and R. A. Woods, 2019: Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) v1.2: an open-source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous state-space formulations. *Geosci. Model Dev.*, **12**, 2463–2480, doi:10.5194/gmd-12-2463-2019.

McMillan, H., 2020: Linking hydrologic signatures to hydrologic processes: A review. *Hydrol. Process.*, **34**, 1393–1409, doi:10.1002/hyp.13632.

Melsen, L., A. J. Teuling, P. J. J. F. Torfs, M. Zappa, N. Mizukami, P. A. Mendoza, M. P. Clark, and R. Uijlenhoet, 2019: Subjective modeling decisions can significantly impact the simulation of flood and drought events. *J. Hydrol.*, **568**, 1093–1104, doi:10.1016/j.jhydrol.2018.11.046.

Mendoza, P. A., M. P. Clark, N. Mizukami, E. D. Gutmann, J. R. Arnold, L. D. Brekke, and B. Rajagopalan, 2016: How do hydrologic modeling decisions affect the portrayal of climate change impacts? *Hydrol. Process.*, **30**, 1071–1095, doi:10.1002/hyp.10684.

Murillo, O., P. A. Mendoza, N. Vásquez, N. Mizukami, and Á. Ayala, 2022: Impacts of Subgrid Temperature Distribution Along Elevation Bands in Snowpack Modeling: Insights From a Suite of Andean Catchments. *Water Resour. Res.*, **58**, under review, doi:10.1029/2022WR032113.

Newman, A. J., and Coauthors, 2015: Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrol. Earth Syst. Sci.*, **19**, 209–223, doi:10.5194/hess-19-209-2015. http://www.hydrol-earth-syst-sci.net/19/209/2015/.

Niu, G.-Y., and Coauthors, 2011: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *J. Geophys. Res.*, **116**, D12109, doi:10.1029/2010JD015139.

Perrin, C., C. Michel, and V. Andréassian, 2003: Improvement of a parsimonious model for streamflow simulation. *J. Hydrol.*, **279**, 275–289, doi:10.1016/S0022-1694(03)00225-7.

Pool, S., M. J. P. Vis, R. R. Knight, and J. Seibert, 2017: Streamflow characteristics from modeled runoff time series - Importance of calibration criteria selection. *Hydrol. Earth Syst. Sci.*, **21**, 5443–5457, doi:10.5194/hess-21-5443-2017.

Pushpalatha, R., C. Perrin, N. Le Moine, T. Mathevet, and V. Andréassian, 2011: A downward structural sensitivity analysis of hydrological models to improve low-flow simulation. *J. Hydrol.*, **411**, 66–76, doi:10.1016/j.jhydrol.2011.09.034. http://dx.doi.org/10.1016/j.jhydrol.2011.09.034.

Stewart, I. T., D. R. Cayan, and M. D. Dettinger, 2005: Changes toward earlier streamflow timing across western North America. *J. Clim.*, **18**, 1136–1155, doi:10.1175/JCLI3321.1.

Valéry, A., V. Andréassian, and C. Perrin, 2014: 'As simple as possible but not simpler': What is useful in a temperature-based snow-accounting routine? Part 2 – Sensitivity analysis of the Cemaneige snow accounting routine on 380 catchments. *J. Hydrol.*, **517**, 1176–1187, doi:https://doi.org/10.1016/j.jhydrol.2014.04.058.

References:

Valéry, A., 2010. Modélisation précipitations – débit sous influence nivale. Élaboration d'un module neige et évaluation sur 380 bassins versants. Thèse de Doctorat, Cemagref (Antony), AgroParisTech (Paris), 405 pp.

Valéry, A., V. Andréassian, and C. Perrin, 2014: 'As simple as possible but not simpler': What is useful in a temperature-based snow-accounting routine? Part 2 – Sensitivity analysis of the Cemaneige snow accounting routine on 380 catchments. *J. Hydrol.*, **517**, 1176–1187, https://doi.org/10.1016/j.jhydrol.2014.04.058.