Final response on: **"Lack of robustness of hydrological models: A large-sample diagnosis and an attempt to identify the hydrological and climatic drivers"**

Detailed responses

1 Referee 1

General comments

This paper presents an application of the robustness assessment test to a large-sample of catchments across France, Denmark and Sweden. The analysis uses results from three hydrological models and the paper analyses how the robustness varies across the dataset in relation to a selection of hydrological and climatic characteristics. Overall the paper is easy to follow and results are presented clearly, but the manuscript could do with a little more synthesis to bring it all together at the end.

Thank you for your encouraging comments

Specific comments

Section 1.3: There could be a little more detail in this section. For example, instead of mentioning that you use a 'large set of catchments spanning various climate conditions in three European countries' perhaps you could mention how many catchments are simulated, across which conditions and in which countries.

We added more detail in the revised version. In addition to the example given by the reviewer we also listed the three used models.

Modifications: lines 78 to 82 of marked-up article

The RAT (Nicolle et al., 2021) is applied to a ~~large~~ set of 352 catchments spanning ~~various~~ four Köppen climate conditions ~~classes~~ (temperate and continental) in ~~three European countries~~ Denmark, France and Sweden, in order to evaluate the robustness of three rainfall–runoff models with various process representations and parameter estimation approaches (namely GR6J, HYPE and MIKE-SHE). The large test set is used to evaluate how model robustness varies over a wide range of climatic and hydrological conditions and to characterize catchments where models lack robustness.

L91 Did you consider using any metrics other than the model bias to assess the differences between observed and simulated flows? Why did you choose this one and do you think your results are sensitive to this choice?

You are right on this point, bias is only one of the metrics that could be considered (and success at the RAT should only be considered as a « necessary but not sufficient » condition for using a model in a climate evolution context: the same methodology could be applied to bias in different flow ranges (low or high flows) or to statistical indicators describing low-flow characteristics or maximum annual streamflow. And characteristics other than bias could be tested, e.g. ratios pertaining to the variability of flows. We had mentioned this point in the Technical note where we introduced the RAT

methodology (Nicolle et al., 2021). Nonetheless, we believe that bias is the first metric to be considered when looking at robustness in a climate change context. We added one sentence to clarify.

Addition: line 96 and 97 of marked-up article

We chose bias as the score to asses model error every year because we believe that it is the first metric to look at when looking at robustness in a climate change context.

L131 could you quantify how many rivers are affected by hydropower production? It would be interesting to know how many catchments in this sample are affected by this.

During the catchment set selection we avoided as much as possible catchments that are influenced by hydropower production. There is still some diffuse influence of human activities in some catchments since totally natural catchments are very rare in Europe. However, catchments with major dams were rejected during the selection. We added a sentence to be more precise in the manuscript and add some sentences about this selection process.

Addition: lines 138 to 140 of marked-up article

We tried to avoid catchment that were too influenced by hydroelectricity production because it would have distorted the analysis since GR6J does not take any regulation into account.

L129 I understand that perhaps it does not dictate the hydrology as heavily, but since the geology is discussed for France and Denmark is it worth describing the Swedish geology as well?

You are right, we added a description in the revised version.

Addition: lines 135 to 137 of marked-up article

In terms of geology, Sweden is dominated by Precambrian crystalline and metamorphic rocks. Faults are one of the main factors that create topography and so, influence catchment delineation.

Table 1 and 2: Although these tables are useful for listing all the signatures used in this study, I do not think the quantiles are particularly easy to digest, is there a more visual way that this information could be displayed? I like the maps in the supplementary material but understand that there are probably too many to include in the main paper.

Thank you for this comment, we agree that these tables are to visualize but we do not have any better idea to replace it. There are too many maps to plot so we made the decision to keep these maps in Supplementary material. We also tried to replace values of quantiles by boxplots but the result is unreadable.

Her ewe did not change anything in the paper since we did not find any better way to present the information than these tables.

L167: Could the runoff ratios exceeding 100% be related to the hydropower? Often water is imported to support these schemes.

It may be an explanation but, as stated before in our answer, we selected catchments without major hydropower dam. Instead we suspect uncertainties in input precipitation measurement. Especially due to the difficulty to measure snow height in these area.

We did not modify anything here

L231: on L432 you mention that there is a regulation module in HYPE, could this be briefly described here?

We will add a short description of this module but, since we tried to avoid catchments with large reservoir this module is barely used in the catchments analyzed in this paper.

Addition: lines 244 to 246 of marked-up article
Regulation of dams is taken into account using simple regulation rules. However, this module has low impact on the results since the catchments used for this study are not affected by major dams.

Figure 3: # of stations isn't a particularly intuitive label, perhaps you could instead write 'Reactive catchments: '

Thank you for this remark, we made the change for Fig.3, Fig.8 and Fig.11.

L282: instead of saying 'especially numerous' could you instead quantify how many catchments are reactive in France?

Thank you, we gave the number

Modifications: lines 291 and 292 of marked-up paper
(i) when temperature is used as a predictor, 70 reactive catchments over a total of 99 are located ~~especially numerous~~ in France;

L385 which is section 0?

Thank you for pointing out this error

Modification: line 394 of marked-up paper
Here, we compare the robustness of the three models presented in Sect. 2.~~40~~.

L410: Have you thought about using any signatures which describe the degree of flow regulation by reservoirs/ hydropower? This might help to identify whether the flaws in the GR4J model are linked to this and could be included as a signature in Figure 10.

It would have been interesting, but we have excluded catchment with major dams to the analysis. It would also have been interesting to use signature that characterize diffuse impact of human activities on each catchment. However, we do not have enough information to characterize it for all of our catchments.

Here we did not modify anything since we said that regulated catchments are excluded (see specific comment number 3 above)

L453: could you elaborate on what you mean when you say the calibration of S-HYPE could be responsible for the seemingly random reactivity? Perhaps this could be done on L533.

We agree that this sentence is unclear. To clarify, we added two sentences.

Addition: lines 465 to 467 of marked-up file
Since S-HYPE is calibrated primarily for flood forecasting, the long-term bias is taken into account in a second time which may influence RAT results for some catchment. Manual tuning specific to some catchments may introduce differences that make difficult to identify a type of catchment that has robustness issues.

L472: what do you mean by differs from the rest of the dataset? If the same can be said for the Swedish catchments then do you just mean that the results differ from those associated with the French data? This seems to be contradicted by L504.

Here again, we agree that this sentence is unclear. We will modify it to be better understandable.

Modification: lines to 485 to 487 of marked-up paper

Here, as for Sweden, we can identify differences between the part of the dataset on which GR6J reacts to RAT and the part of the dataset on which it does not. ~~Here, as for Sweden, the profile of catchments on which GR6J reacts to RAT differs from the rest of the dataset.~~

Table 4: Could you perhaps shade the last three columns so that we can see the patterns visually?

It is a good idea ~~and we will do it~~but the journal template does not allow to add shade in table. Thus, we decide not to do anything here.

Table 4 and its caption were not modified.

L528: Again, perhaps you could consider using a signature to quantify the degree of dam regulation in each catchment to confirm or reject this hypothesis.

See our answer before. We will remove this sentence to avoid misunderstanding.

Removed: lines 543 and 544 of marked-up paper
~~The last observation from this catchment characterization is that GR6J seems to have robustness issues on catchments in which streamflow is regulated by dams.~~

L564 what did you do with the catchments where the KGE was less than 0.7?

We did not remove any basin based for reasons of low model performance, as we consider that it would have biased our analysis. We will try to be clearer in the text.

Modification: lines 582 to 584 of marked-up file
In the case of our study, performances are good overall. All the three models have a KGE value higher than 0.7 on 329 catchments over the 352. ~~good performance in the majority of the catchments (the KGE value is greater than 0.7 on almost all the catchments).~~

Section 5.1: This section feels like a lot of repetition of results/ discussion and doesn't really feel like it achieves much synthesis. It would be good to make the implications of your work clearer here. The start of the paper makes it clear that this work is useful for understanding the implications of using models such as HYPE, GR6J and MIKE SHE for climate change applications, but I don't feel like you ever quite bring together your findings here and discuss what your results mean for using these models for climate change applications. 'Our analysis pointed out flaws in the models in terms of robustness to changing climate.'. Although I can see that the idea is that you use the results from catchments with different climatic conditions as proxies for how the models will perform under climate change, it would be good to make this link clearer.

We modified this section to try to achieve a real synthesis, as this is what will be useful to the reader. The section is now much shorter.

Modifications: lines 588 to 620 of marked-up file. The section now reads:
This paper presented a large-sample analysis of the robustness of three models to a changing climate. The RAT allowed us to evaluate the robustness of the three different models without controlling their calibration process, and the analysis of the hydrological signatures of the catchments that react to the RAT suggested some potential issues specific to each model. Our objective was not to compare models, as we have shown that they all suffered of a lack of robustness to be safely applied in a changing climate context, but to identify the hydrological features that could be the cause of this lack of robustness. Overall, the models reacted to the RAT on a significant number of catchments (between 33% and 42% depending on the model and the datasets), and this indicates that much work is needed to make models more robust in the context of climate change.

It would be good to also have some discussion surrounding how transferable your results are to other hydrological models. Are your findings only relevant for the models used in this study? Or is it likely that your findings will be relevant for other models in used in other countries too?

Thank you for suggesting this. We plan added short section to the conclusion and the reference that correspond.

Addition: lines 621 to 636 of marked-up file

5.2 How generic are our results?

The issue of genericity is central in science. With an application of the RAT over 3 models, in 3 countries and on a total of 352 catchments, the work presented in this paper presents a significant improvement over what had initially been done in the note describing the method (Nicolle et al., 2021). Because models are more than ever used to predict the impact of a changing climate, we believe more than ever in the need to test them more thoroughly, in the need to challenge their extrapolation capacity. Because the RAT is so simple to apply, because it can be applied to models requiring calibration that run in seconds and to models which do not and need hours to produce a single run, we consider that it is a useful investment for a modeler as well as for a model user, one that is likely to « increase their confidence » in their results as de Marsily et al. (1992) were recommending.

Of course, we keep in mind the advice that the late Vit Klemeš (personal communication) had sent to one of us. Asked how he was looking back at the impact of his famous paper discussing the different options of split-sample test (Klemeš, 1986), he answered that he had in fact always been skeptical about the capacity of hydrologists to validate rigorously their models : he said he knew in advance that the tests he had suggested would be « avoided under whatever excuses available because modelers, especially those who want to 'market' their products, know only too well that they would not pass it », adding that he had « no illusions in this regard » when he wrote his paper. We do not have any illusions either, and we do not wish to fight against windmills. We modestly think it is part of our scientific duty to keep expressing our concerns on this topic.

Addition to the reference list

de Marsily, G., Combes, P., and Goblet, P. 1992. Comment on 'Ground-water models cannot be validated', by L.F. Konikow and J.D. Bredehoeft. Adv. Water Resour., 15, 367-369.

2 Reviewer 2

The study by Santos et al. explores the application of the robustness assessment test (RAT) for three hydrological models of varying complexity. They tested the RAT in 352 catchments across Denmark, France, and Sweden. The topic is very interesting and indeed worth studying. The methodology is well-explained, and the writing is clear. However, my main concern is that since these three models were calibrated separately, each at different times and by different research institutes, I am worried about the comparability of the results. Additionally, some of the explanations for the results appear somewhat strained and lack adequate data support; for example, linking robustness issues to dams regulation. If the authors adequately address these issues, I believe this paper is suitable for publication in the HESS journal. My detailed comments can be found below.

Thank you for your encouraging comments

Detailed comments:

Line88: In this line 88, it says at least 30 years of data is needed, yet in the Figure 1, it labelled with '> 20 years'. So what is the minimum requirement for data?

Thank you for pointing this. It is a mistake, we will replace "30 years" by "20 years"

Modification: lines 89 to 91 of marked-up file

The RAT only requires observed climatic variables (to be used as a potential predictor for the model bias), as well as simulated and observed flows covering a sufficiently long time period (at least 23~~0~~ years, as shown in the study by Nicolle et al., 2021).

Line 116: Could you could use Köppen-Geiger classes as a background map in Figure 2? This provides readers with a more intuitive understanding of the climate zones to which each watershed belongs.

This will be a good improvement to the map, we added it with data from Beck et al. (2018). We also cited the paper.

Additions and modifications:
Figure 2 has been modified, a reference to the figure has been added at line 119 of the marked-up file, and the following text has been added to the figure caption:
Background colours represent the Koppen-Geiger climate classes (data and legend are described in Beck et al., 2018).
Beck et al. (2018) has been added to reference list (line 670).

Line 222-223: What do you mean by 'free parameters'? Please clarify.

The free parameters are those which are allowed to be adjusted during the model calibration process (by opposition to the many fixed parameters which have been chosen to describe the hydrological processes and remain the same for all the catchments). We modified the sentence to turn it clearer.

Addition: from line 228 to 230 of marked-up file
GR6J (Pushpalatha et al., 2011) is a lumped bucket-type model that simulates catchment runoff response to rainfall using six free parameters which are adjusted during calibration.

Line 234: Does this 'ca.' represent the catchment area?

In reality, ca. stood for "circa". We removed it since it is not absolutely necessary in this sentence.

Removed: line 241 of marked-up file
In the version used here (S-HYPE-2016b) the whole country is divided into sub-catchments of an average size of ~~ca.~~ 13 km^2.

Table 3: Please add the explanation of what do you mean by 'OF'?

We modified the caption accordingly by adding (OF: Objective Function).

Table 3: You mentioned in the discussion that these 3 models were calibrated on different temporal period. Could you add in this table about the specific time periods during which each of these models was calibrated?

Thank you for this suggestion, we added the information in the table 3.

Line 385: Can you clarify what do you mean by 'Sect. 0'?

Thanks for pointing out this mistake. We corrected it.

Modification: line 394 of marked-up paper
Here, we compare the robustness of the three models presented in Sect. 2.4~~0~~.

Line 430: I don't think the large river catchments will necessarily be higher than the average level. This sentence is not rigorous. Please correct.

We modified the sentence.

Modification: lines 440 and 441 of marked-up file

GR6J seems to react to RAT on a specific type of catchments (which are at the same time large and which have a higher than average specific flow)~~It seems that GR6J reacts to RAT on large river catchments where the flow is higher than the average~~.

Line 523-525: Can you provide the details of these tests on the choice of the evaporation formula in the supplement file?

Yes, we added a figure that shows our results in supplementary material.

Addition: we added a supplementary material number 5 and some words to include it on lines 650 and 651 of marked-up file.

Line 527: Can you add the location of these dams in one of your figures? It would nice and more convinced to see the spatial distribution of both dams locations and the GR6J robustness issues to draw this conclusion. Moreover, the presence of dams in a catchment does not necessarily mean they are impacted by the dams. So how do you know the streamflow of these catchments are actually affected by the dams?

See discussion with reveiwer number 1. We avoided major dams in our catchment set selection. We will try to be clearer here.

Removed: lines 542 and 543 of marked-up paper
~~The last observation from this catchment characterization is that GR6J seems to have robustness issues on catchments in which streamflow is regulated by dams.~~

Line 534-539: I'm a bit concerned about the different calibration strategies were used and also model calibrated over different time periods. Adopting different calibration methods may introduce uncertainty. I'm not sure whether the calibration results of models using different calibration methods are comparable or not? More justifications are needed here.

We understand your point: in order to keep « all other things equal », and to ease the interpretation, your argue that it would have been better to have the same calibration strategy for all models. The problem is that the recommended calibration strategies vary widely among models, and very often, we do face another critic: "you did not respect the calibration that the modelers recommend, therefore we do not trust your results" or even "you don't have the necessary expertise to perform the calibration of this model, therefore we do not trust your results". We believe that the only way to avoid this critic is to ask each "specialist" to perform his calibration/parameterization, to do his best, and to judge of the robustness of the model when placed in the "best" conditions.

Addition: lines 558 to 560 of marked-up file
These differences in terms of calibration processes are difficult to overcome since the models have different structures that requires different calibration processes. It is, then, difficult to avoid here since one of the scopes of the paper is to compare models with different modelling philosophies.