Thank you very much for your review. In addition to our general answer, we provide here specific answers to the points you raised (in blue while your comments remain in black).

1. General comments

This technical note by Nicolle et al. presents an evaluation method for hydrological model robustness and is called the Robustness Assessment Test (RAT). The main assumption of the RAT is that the bias in the model output (i.e. streamflow) should not be correlated with the climatic input data (e.g. precipitation, temperature, air humidity), in which case, the model has a dependency on climactic variables, thereby not suitable for use in climate change impact studies. The manuscript is relatively well structured and written. The development of a framework for hydrological model evaluation is relevant, and can potentially be of interest for the readers of HESS if the RAT is thoroughly evaluated and some of its limitations are addressed. The RAT seems to be developed to detect deficiencies in model structure but it is not clearly demonstrated that model rejection by the RAT is not also due to input data. This is a key point that the authors have to address. Please refer to my comments and suggestions below.

2. Specific comments

2.1 Major Comments

2.1.1 Contradictions

The main assumption is that there should not be a dependency between model output bias and the input climatic variable for a hydrological model to be considered robust (Lines 152-157). However, in the conclusion, the authors clearly recognize that "Detecting a relationship between model bias and a climate variable using the RAT does not allow to directly conclude on a lack of model robustness". They further mention in the key points that "success at the RAT test is a necessary (but not sufficient) condition of model robustness". Therefore, I wonder why the RAT should be used for model evaluation.

We consider that given that the input data are reliable (which is to be checked before any study), a fail in passing the RAT is very likely indicating a model robustness issue. We will rephrase the mentioned sentences to avoid any misunderstanding.

2.1.2 Input data

How do you know if the model failure at the RAT is due to the dependency on input data or to the dependency on model structure or parametrization?

We assume that the reviewer means here "dependency on input data bias", not solely "dependency on input data". As mentioned below, we consider that checking the quality of input data has to be done prior to assessing the robustness of models. The RAT does not allow to decipher the sources of these dependencies.

L170-171: Is the model rejected because the parametrization is wrong or because the input data is wrong? Would you get the same results with a different precipitation data? A clear answer to this question is essential for this work, as we do not want to reject a model that is not wrong.

We agree with you that "we do not want to reject a model that is not wrong", but like with any test, this risk (type I error) will exist. This is the risk of judging guilty an innocent person, which, in the case of a model is not so worrying: it is in the very nature of hydrological models to be guilty or imperfect. A test such as

RAT may yield a false positive for a few catchments, but it will nonetheless offer a possibility for comparing modelling alternatives.

In their work, the authors assume that in case there is a correlation between the model output bias and the input data, that correlation is due to the model structure/parametrization, as the authors do not investigate the potential contribution of the input data to the model output bias. This is a clear limitation of the evaluation of the RAT method. Without testing different input data, the authors implicitly assume that the only input data that they are using is right or at least is not the source of errors in the model outputs. But we know that input data remain a main source of uncertainty in hydrological modelling (Gupta, 1998 https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/97WR03495). The authors recognize in the conclusion that the lack of model robustness can also be due to meteorological causes (L324-325). I urge the authors to test different input datasets, which are largely available nowadays because of the availability of satellite and reanalysis products. Different precipitation and temperature datasets must be tested to demonstrate that model rejection based on the RAT is not false negative, i.e. Type 2 error that we want to avoid in model evaluation (Beven, 2010 https://onlinelibrary.wiley.com/doi/epdf/10.1002/hyp.7718).

We do agree on the rationale of these comments. However, we believe that addressing data quality in this paper would deviate the aim of this work, which is presenting a method to assess model robustness. Also, we think that global datasets will be unavoidably much more imprecise than Meteo-France's ground-based interpolated data, which will increase considerably the amplitude of bias (we have a long experience of using Meteo France's SAFRAN product as catchment forcing, and we consider it to be the best option in France).

Another option to test if the model performance depends on precipitation, for instance, is to take a precipitation product and gradually apply some perturbation of [-30, -20, -10, 0, 10, 20, 30] percent bias and check if the bias in streamflow is correlated to precipitation for the different scenarios.

This suggestions is interesting, but the data quality issues that modellers are the most concerned with are not random errors, they are the trends or the sudden changes linked with modifications of the observation network (we could still add noises with a trend but it would be a rather obvious result that these are detectable).

2.1.3 Seasonality

Is the RAT valid for catchments with a strong seasonality in climatic variables? In case of bias in the input data, wouldn't that easily be reflected in the model output for catchments with strong seasonality? Thereby, misleading the conclusion of the RAT as the RAT would reject the model assuming that it is its parametrization that is problematic?

We believe that the RAT is still valid for catchments with a strong seasonality. However, we recognize seasonal biases if they do compensate each other would not be detected by the RAT in its present form. A seasonal bias could indeed help.

2.1.4 Annual aggregation - Hydrological year

L159: Fig.1 - What is the reasoning behind the annual aggregation of the data by hydrological year to compute the bias? What are the implications of the choice of the hydrological year on the calculation of the bias?

You are right, a perfect model does not need a specific season to compute bias. But in the case of snow-affected catchments, it seems always more careful to avoid separating the snow accumulation and the snow melt seasons. We applied the RAT both with calendar and hydrological years, and obtained a difference on only one catchment (see table below).

Table 1. Comparison of RAT results depending of the time step chosen for bias calculation (calendar year vs. hydrological year)

Catchment	Significant dependency with calendar years	Significant dependency with hydrological years
A1080330	FALSE	TRUE
B2220010	TRUE	TRUE
H2342020	TRUE	TRUE
H4252010	TRUE	TRUE
H7401010	FALSE	FALSE
H8212010	TRUE	TRUE
I5221010	TRUE	TRUE
J7483010	TRUE	TRUE
K1321810	TRUE	TRUE
K6402520	TRUE	TRUE
L0563010	TRUE	TRUE
L4411710	TRUE	TRUE
M0243010	FALSE	FALSE
M7112410	FALSE	FALSE
O0592510	FALSE	FALSE
O7101510	TRUE	TRUE
Q5501010	TRUE	TRUE
S2242510	FALSE	FALSE
U4644010	TRUE	TRUE
V4264010	TRUE	TRUE
Y4624010	FALSE	FALSE
TOTAL	**14**	**15**

The model output bias and climatic variables may not be dependent at daily scale but show dependency at coarser temporal scale, or vice versa. How would that be captured by the RAT? The RAT might reject the model at annual scale while there is no dependency at daily scale (temporal resolution of the hydrological simulations), or inversely.

We consider that because hydrological models are full of imperfections, it is more reasonable to work on bias at coarse time steps first, and then only to move to fine time steps. But we agree that the absence of bias for any given year, may well hide a positive bias during half of the year and a negative year during the other half. This is why we underline in conclusion that "the RAT should not be seen as the only basis for model choice."

2.2 Minor comments

L15: I see the RAT as a “complement” rather than an alternative to the “split-sample test”.

We disagree on this comment. Indeed, the split-sample test cannot be applied to all models, e.g. physically-based models that do not use calibration. For these models, the RAT is therefore an alternative, not a complement.

A model might pass an SST but fail the RAT, or vice versa. What would happen in those cases? Should the model be rejected or accepted? Which of the SST and RAT outcomes would you give a preference?

This is a tough question, and it also depends on the threshold chosen (the p-value in the case of the RAT, the drop in criterion value between the calibration and the validation period in the case of the SST). Usually, people trust more a method that they have been using for a long time (experience with a method is an important factor of adoption). We have a little more than one year of experience with RAT, much more with SST... However we like the visualization opportunity that the RAT offers, this is why we are tempted to answer that we would give our preference to the RAT results.

L16: “the RAT method does not require multiple calibrations”. This sentence can be misinterpreted because it seems like the RAT method could be calibrated while you are referring to the hydrological model. Should be “…multiple calibrations of hydrological models”. Also correct this at line 137.

We agree and will make the modifications.

L20: As you said that success at the RAT is a necessary BUT not a sufficient condition of model robustness, can you call your approach a “robustness” assessment test? if a model is robust it would be successful at the RAT, but success at the RAT does not necessarily mean the model is robust. This key point highlights a strong limitation of RAT because you cannot confirm the robustness of the model but just have a hint that it might be robust.

We don’t know of perfect tests (there is always a type I and type II error possibility, as you mentioned it). RAT is definitely not a “Robustness Assessment Certificate”, but we do think that it is fair to call it a test.

L319-320: “Detecting a relationship between model bias and a climate variable using the RAT does not allow to directly conclude on a lack of model robustness.” Isn’t this statement in direct contradiction with your research hypothesis? Thereby, highlighting that the proposed RAT is not mature enough as a method for model robustness evaluation. From these contradictions, is it still relevant to use the RAT?

We believe that it is relevant, because it can help us compare models, it can help us identify unwanted/unexpected behaviors. RAT cannot reveal all the flaws of models but some: this is already a valid information useful for model selection.

L57: “…considered by all hydrologists as a good modelling practice”. This statement is speculative. Something like “…most hydrologists…” would be acceptable.

Ok.

L74: I found the section 1.2 a bit too long.

We will consider reducing this section when revising the manuscript.

L101: IAHS should be defined here at its first occurrence, instead of at line 106.

We agree, this will be done.

L127-128: “it is difficult to distinguish which cases of DSST failure are truly caused by model structural inadequacy”. Do you think that the RAT can address that limitation of the DSST? This is not demonstrated in your manuscript.

We think that because of the exhaustive nature of the RAT (examining the results of a long simulation period), it is less sensitive to the “contingency on the chosen calibration method” mentioned by Fowler et al. (2018).

L148-149: “The specificity of the RAT is that it requires only one calibration (or one parameterization)”. The use of the term “calibration” is confusing. Shouldn’t you use the term “simulation” here as you did at lines 140-141?

This will be changed when revising the manuscript: “*The specificity of the RAT is that it requires only one simulation covering…”*

L149: “at least 30 years”. In Fig.1 it’s “> 20 years”. Is it 20 or 30 years? Be coherent through the manuscript (check lines 184, 310).

The reviewer is right, we will modify the text.

L270: “It happens on only two catchments out of 21”. What are those two catchments? Do they have any similarities (climate, elevation, etc.) that might explain this result? Same questions for catchments identified under the other dependency tests (see L281, L287, L293).

We will consider developing these analyses in the revised version of the manuscript.

L300: “robustness evaluation from all types of hydrological models”. This is not explicitly demonstrated in the manuscript. Only the GR4J is used in your methodology.

Since the RAT only needs one simulation time series, we believe that it is applicable to any kind of model. We remind that this article should more be seen as the description of a method than as the real assessment of a model. For demonstrating a drawback of hydrological models, we found it more elegant to demonstrate it on our own model rather than to include other models, which explains why only GR4J is applied here.

L320-322: “Indeed…robustness”. This statement needs clarifications.

We will clarify the statement.

2.3 Conclusion

In conclusion, I think the RAT is subject to many limitations that challenge its own robustness and validity, which might hinder its large adoption by the scientific community. Therefore, I recommend that the authors develop strategies to address most of the limitations and thoroughly test the robustness of the RAT before it can potentially be released in the public.

We thank the reviewer for their comments and useful remarks and we hope that our answers will help to convince them about the benefit of the RAT method and will help to address potential misunderstandings

3. Technical corrections

L53: Thirel et al., 2015. Please specify if its “a” or “b”. Also check line 100.

OK

L153: “numeric criterion” ---> “numeric bias criterion”

OK

L163: one missing closing parenthesis “)”.

The missing parenthesis is actually located at l. 164

L250: “computation price” ---> “computation cost”

OK

L325: “metrological” ---> “meteorological”

We do not agree: metrology is used here in the sense of linked to measurement.

L335: “it true” ---> “it is true”

We rather changed to “prove it right”