Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC1
https://doi.org/10.5194/hess-2021-147-RC1, 2021

Comment on hess-2021-147

Anonymous Referee #1

Referee comment on "Technical note: RAT – a robustness assessment test for calibrated and uncalibrated hydrological models" by Pierre Nicolle et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-147-RC1, 2021

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.

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?

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.

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).

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.

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?

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?

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.

2.2 Minor comments

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

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?

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.

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.

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?

L57: "…considered by all hydrologists as a good modelling practice". This statement is

speculative. Something like "…most hydrologists…" would be acceptable.

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

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

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.

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?

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).

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).

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.

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

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.

3. Technical corrections

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

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

L163: one missing closing parenthesis “)”.

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

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

L335: “it true” ---> “it is true”