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).

This manuscript introduces an alternative to the generalised split sample test, which is less demanding computationally yet still provides similar insights into model robustness, as illustrated by Figures 4 and 5. This is an important outcome, as this new approach, named the RAT, has the potential to make the crash-testing of hydrological models more widely used before they are employed to assess future climate change impacts. Of course, more detailed tests of model realism exist, but there is, I believe, a need for tests that can be readily applied using typically available simulation data and provide a first-order assessment of the robustness of a model in a changing climate.

We thank the reviewer for their positive assessment of the interest of the method.

There is some ambiguity in the paper about what can and cannot be inferred from the RAT. This is for instance clear from the key point 3 ("the RAT method can be used to determine whether a hydrological model can be safely used for climate change impact studies") and 4 ("success at the RAT test is a necessary (but not sufficient) condition of model robustness"), which in my view, are in contradiction. While I agree with key point 4, I would argue that the RAT does not enable us to declare that it is "safe" to use the model for climate studies (at most, one could argue that the RAT is useful to identify models that are "unsafe" to use). Like the vast majority of model evaluation techniques, RAT can only falsify a model but not guarantee its validity. This is recognised by the authors on L322 L335 and in several other places, but it needs to be clearer throughout the key points, abstract and paper. Also, as the authors use the word ""climateâproof"" L136, they should clarify that a methodology like the RAT is not enough to declare that a model is "climate-proof".

You are right, there is a logical contradiction with key point 4. Indeed, the RAT is useful to identify models that are "unsafe" to use. Or even (because we believe that no model can be considered as perfectly safe), RAT can be used to compare models and identify "the less unsafe" one. We will modify this sentence as follows "the RAT method can be used to determine whether a hydrological model cannot be safely used for climate change impact studies".

While I find section 4.1 quite convincing, I feel that section 4.2 needs more work. At the moment, it essentially illustrates that biases in streamflow simulations depend on different climatic variables for different catchments, which could be expected. There is scope for a more substantial discussion on what could be done when these different types of model failure happen. Should models be excluded as soon as streamflow errors are significantly correlated with one climate variable? Or two? Are correlations with some climatic variables more detrimental than others? Could the model be re-calibrated to improve robustness (if so, how?) or should other model structure(s) be used? Of course, one could simply say that "it depends on the study". But I think that answering these questions that users of the RAT are likely to face, or at least, proposing an approach to answer them, is essential for the RAT to be used widely, effectively, and in a consistent way across studies. I agree with the authors that "the RAT should not be seen as the only basis for model choice" L345

We would argue that we present here RAT as a tool. Once a tool is proved useful, different users may still be willing to use it in different ways.

As the RAT is mostly data driven (in contrast to other tests focussed more on process representation), clearer recommendations on the input data should be provided. It is mentioned, almost in passing, that "some of the lacks of robustness detected among the 21 catchments presented here could be in fact due to metrological causes" (second bullet point of the conclusions). Hence, robust models can (wrongly) be rejected because of artefacts in the input data. Could the authors illustrate this, ideally using data from one of their catchments? Sadly, in large-sample datasets (often in contrast to studies relying on a few research catchments), there is rarely detailed enough metadata/knowledge on individual catchments to catch these artefacts in the input data. Furthermore, many large-sample datasets rely on meteorological gridded data products to produce a catchment average, and these products typically favour accuracy (using data from all available stations at each time step) over temporal homogeneity (sticking to the same set of stations for the entire period), so the risk of inhomogeneities is real. There may also be inhomogeneities in the streamflow time series, for instance caused by changes in rating. This is of course not something I am expecting the authors to solve, but I encourage them to discuss these data-related challenges earlier than in the conclusions (maybe in section 2 or 3). Ideally, we would like to differentiate between failures of RAT caused by the lack of robustness of the hydrological model (model inadequacy) and failures caused by trends/inhomogeneities in the data. Can the authors elaborate on this?

Honestly, we see more perspectives for learning from RAT on large sample experiments than on a catchment-by-catchment application. RAT may not offer a definitive answer but it can be used for comparing modelling alternatives.

Point 5 of the conclusions: "Although it would be tempting to transform the RAT into a post-processing method, we do not recommend it", what do the authors mean by a post-processing method? "What we do recommend is to work as much as possible on the model structure, to turn it less climate dependent", yes, but where to start?

A last precision on our use of the term "post-processing": if we identify a linear dependency between model bias and temperature for example, one could be tempted to fit a linear correction model in order to unbias the results (this is what we call "post-processing"). And when writing that we do not recommend it, we wanted to stress that we should not focus on the symptoms but rather try to identify the causes of the bias.

Thank you for developing this method.