Technical Note – RAT: a Robustness Assessment Test for calibrated and uncalibrated hydrological models

Pierre Nicolle[1,3], Vazken Andréassian[1,*], Paul Royer-Gaspard[1], Charles Perrin[1], Guillaume Thirel[1], Laurent Coron[2], Léonard Santos[1]

[1]Université Paris-Saclay, INRAE, UR HYCAR, 92160, Antony, France

[2]EDF, DTG, Toulouse, France

[3]now at Université Gustave Eiffel, Nantes, France

[*]Corresponding author: Vazken Andréassian (vazken.andreassian@inrae.fr)

Key Words
hydrological modelling, split-sample test, differential split-sample test, model evaluation, robustness, climate change

Key Points
- a new method (RAT) is proposed to assess the robustness of hydrological models, as an alternative to the classical split-sample test
- the RAT method does not require multiple calibrations of hydrological models: it is therefore applicable to uncalibrated models
- the RAT method can be used to determine whether a hydrological model cannot be safely used for climate change impact studies
- success at the RAT test is a necessary (but not sufficient) condition of model robustness

Abstract
Prior to their use under future changing climate conditions, all hydrological models should be thoroughly evaluated regarding their temporal transferability (application in different time periods) and extrapolation capacity (application beyond the range of known past conditions). This note presents a straightforward evaluation framework aimed at detecting potential undesirable climate dependencies in hydrological models: the robustness assessment test (RAT). Although it is conceptually inspired by the classic differential split-sample test of Klemeš (1986), the RAT presents the advantage to be applicable to all types of models, be they calibrated or not (i.e. regionalized or physically based). In this note, we present the RAT, illustrate its application on a set of 21 catchments, verify its applicability hypotheses and compare it to previously published tests. Results show that the RAT is an efficient evaluation approach, passing it successfully can be considered a prerequisite for any hydrological model to be used for climate change impact studies.

1 Introduction

1.1 All hydrological models should be evaluated for their robustness

Hydrologists are increasingly requested to provide predictions of the impact of climate change (Wilby, 2019). Given the expected evolution of climate conditions, the actual ability of models to predict the corresponding evolution of hydrological variables should be verified (Beven, 2016). Indeed, when using a hydrological model for climate change impact assessment, we make two implicit hypotheses concerning:

- the **capacity of extrapolation beyond known hydroclimatic conditions**: we assume that the hydrological model used is able to extrapolate catchment behaviour under conditions not or rarely seen in the past. While we do not expect hydrological models to be able to simulate a behaviour which would result from a modification of catchment physical characteristics, we do expect them to be able to represent the catchment response to extreme climatic conditions (and possibly to conditions more extreme than those observed in the past);

- the **independence of the model set-up period**: we assume that the model functioning is independent of the climate it experienced during its set-up/calibration period. For those models which are calibrated, we assume that the parameters are generic and not specific to the calibration period, i.e. they do not suffer from overcalibration on this period (Andréassian et al., 2012).

Hydrologists make the hypothesis that model structure and parameters are well-identified over the calibration period and that parameters remain relevant over the future period, when climate conditions will be different. Unfortunately, the majority of hydrological models are not entirely independent of climate conditions (Refsgaard et al., 2013; Thirel et al., 2015b). When run under changing climate conditions, they sometimes reveal an unwanted sensitivity to the data used to conceive or calibrate them (Coron et al., 2011).

The diagnostic tool most widely used to assess the robustness of hydrological models is the split-sample test (SST) (Klemeš, 1986), which is considered by most hydrologists as a “good modelling practice” (Refsgaard & Henriksen, 2004). The SST stipulates that when a model requires calibration (i.e. when its parameters cannot be deduced directly from physical measurements or catchment descriptors), it should be evaluated twice: once on the data used for calibration and once on an independent dataset. This practice has been promoted in hydrology by Klemeš (1986), who did not invent the concept (Arlot & Celisse, 2010; Larson, 1931; Mosteller & Tukey, 1968), but who formalized it for hydrological modelling. Klemeš proposed initially a four-level testing scheme for evaluating model transposability in time and space: (i) split-sample test on two independent periods, (ii) proxy-basin test on neighbouring catchments, (iii) differential split-sample test on contrasted independent periods (DSST), and (iv) proxy-basin differential split-sample test on neighbouring catchments and contrasted periods.

For model applications in a changing climate context, Klemeš’s DSST procedure is of particular interest. Indeed, when calibration and evaluation are done over climatically-contrasted past periods, the model faces the difficulties it will have to deal with in the future. The power of DSST can be limited by the climatic variability observed in the past, which may be far below the drastic changes expected in the future. However, a satisfactory behaviour during the DSST can be seen as a prerequisite of model robustness.

1.2 Past applications of the DSST method

The DSST received limited attention up to the 2010s, with only a few studies which applied it. The studies by Refsgaard & Knudsen (1996) and Donelly-Makowecki & Moore (1999) investigated to which extent Klemeš's hierarchical testing scheme could be used to improve the conclusions of model intercomparisons. The study by Xu (1999) questioned the applicability of models in nonstationary conditions and was one of the early attempts to apply the Klemeš's testing scheme in this perspective. Similarly, tests carried out by Seibert (2003) explicitly intended to test the ability of a model to extrapolate beyond calibration range and showed limitations of the tested model, stressing the need for improved calibration strategies. Last, Vaze et al. (2010) also investigated the behaviour of four rainfall-runoff models under contrasting conditions, using wet and dry periods on catchments in Australia that experienced a prolonged drought period. They observed different model behaviours when going from wet to dry or dry to wet conditions.

More recently, Coron et al. (2012) proposed a generalized SST (GSST) allowing for an exhaustive DSST to evaluate model transposability over time under various climate conditions. The concept of GSST consists in testing "the model in as many and as varied climatic configurations as possible, including similar and contrasted conditions between calibration and validation.". Seifert et al. (2012) used a differential split-sample approach to test a hydrogeological model (differential being understood with respect to differences in groundwater abstractions). Li et al. (2012) identified two dry and two wet periods in long hydroclimatic series to understand how a model should be parameterised to work under nonstationary climatic conditions. Teutschbein and Seibert (2013) performed differential split-sample tests by dividing the data series into cold and warm as well as dry and wet years, in order to evaluate bias correction methods. Thirel et al. (2015a) put forward an SST-based protocol to investigate how hydrological models deal with changing conditions, which was widely used during an workshop of the International Association of Hydrological Sciences (IAHS), both with physically-oriented models (Gelfan et al., 2015; Magand et al, 2015), conceptual models (Brigode et al., 2015; Efstratiadis et al., 2015; Hughes, 2015; Kling et al., 2015; Li et al., 2015; Yu and Zhu, 2015) or data-based models (Tanaka and Tachikawa, 2015; Taver et al., 2015).

Recently, with the growing concern on model robustness in link with the Panta Rhei decade of the IAHS (Montanari et al., 2013), a slow but steadily increasing interest is noticeable for procedures inspired by Klemeš's DSST (see e.g. the Unsolved Hydrological Problem n° 19 in the paper by Blöschl et al., 2019: *How can hydrological models be adapted to be able to extrapolate to changing conditions?*). A few studies used the original DSST or GSST to implement more demanding model tests (Bisselink et al., 2016; Gelfan and Millionshchikova, 2018; Rau et al., 2019; Vormoor et al., 2018). For example, based on an ensemble approach using six hydrological models, Broderick et al. (2016) investigated under DSST conditions how the robustness can be improved by multi-model combinations.

A few authors also tried to propose improved implementations of these testing schemes. Seiller et al. (2012) used non-continuous periods or years selected on mean temperature and precipitation to enhance the contrast between testing periods. This idea to jointly use these two climate variables to select periods was further investigated by Gaborit et al. (2015), who assessed how the temporal model robustness can be improved by advanced calibration schemes. They showed that the robustness of the tested model was improved when going from humid-cold to dry-warm or from dry-cold to humid-warm conditions when using regional calibration instead of local calibration. Dakhlaoui et al. (2017) investigated the impact of DSST on model robustness by selecting dry/wet and cold/hot hydrological years to increase the contrast in climate conditions between calibration and validation periods. These authors later proposed a bootstrap technique to widen the testing conditions (Dakhlaoui et al. 2019). The investigations of Fowler et al. (2018) identified some limits of the DSST procedure and concluded that "model evaluation based solely on the DSST is hampered due to contingency on the chosen calibration method, and it is difficult to distinguish which cases of DSST failure are truly caused by model structural inadequacy". Last, Motavita et al. (2019) combined DSST with periods of variable length, and conclude that parameters obtained on dry periods may be more robust.

All these past studies show that there is still methodological work needed on the issue of model testing and robustness assessment. This note is a further step in that direction.

1.3 Scope of the technical note

This note presents a new generic diagnostic framework inspired by Klemeš's DSST procedure and by our own previous attempts (Coron et al., 2012; Thirel et al., 2015a) to assess the relative confidence one may have with a hydrological model to be used in a changing climate context. One of the problems of existing methods is the requirement of multiple calibrations of hydrological models: these are relatively easy to implement with parsimonious conceptual models but definitively not with complex models that require long interventions by expert modellers and, obviously, not for those models with a once-for-all parameterisation.

Here, we propose a framework that is applicable with only one long period for which a model simulation is available. Thus, the proposed test is even applicable to those models that do not require calibration (or to those for which only a single calibration exists).

Section 2 presents and discusses the concept of the proposed test, section 3 presents the catchment set and the evaluation method, and section 4 illustrates the application of the test on a set of French catchments, with a comparison to a reference procedure.

2 The robustness assessment test (RAT) concept

The robustness assessment test (RAT) proposed in this note is inspired by the work of Coron et al. (2014). The specificity of the RAT is that it requires only one simulation covering a sufficiently-long period (at least 20 years) with as much climatic variability as possible. Thus, it applies at the same time to simple conceptual models that can be calibrated automatically, to more complex models requiring expert calibration, and to uncalibrated models for which parameters are derived from the measurement of certain physical properties. The RAT consists in computing a relevant numeric bias criterion repeatedly each year and then exploring its correlation with a climatic factor deemed meaningful, in order to identify undesirable dependencies and thus to assess the extrapolation capacity (Roberts et al., 2017) of any hydrological model. Indeed, if the performances of a model are shown to be dependent on a given climate variable, this can be an issue when the model is used on a period with a changing climate. The flowchart in Figure 1 summarizes the concept.

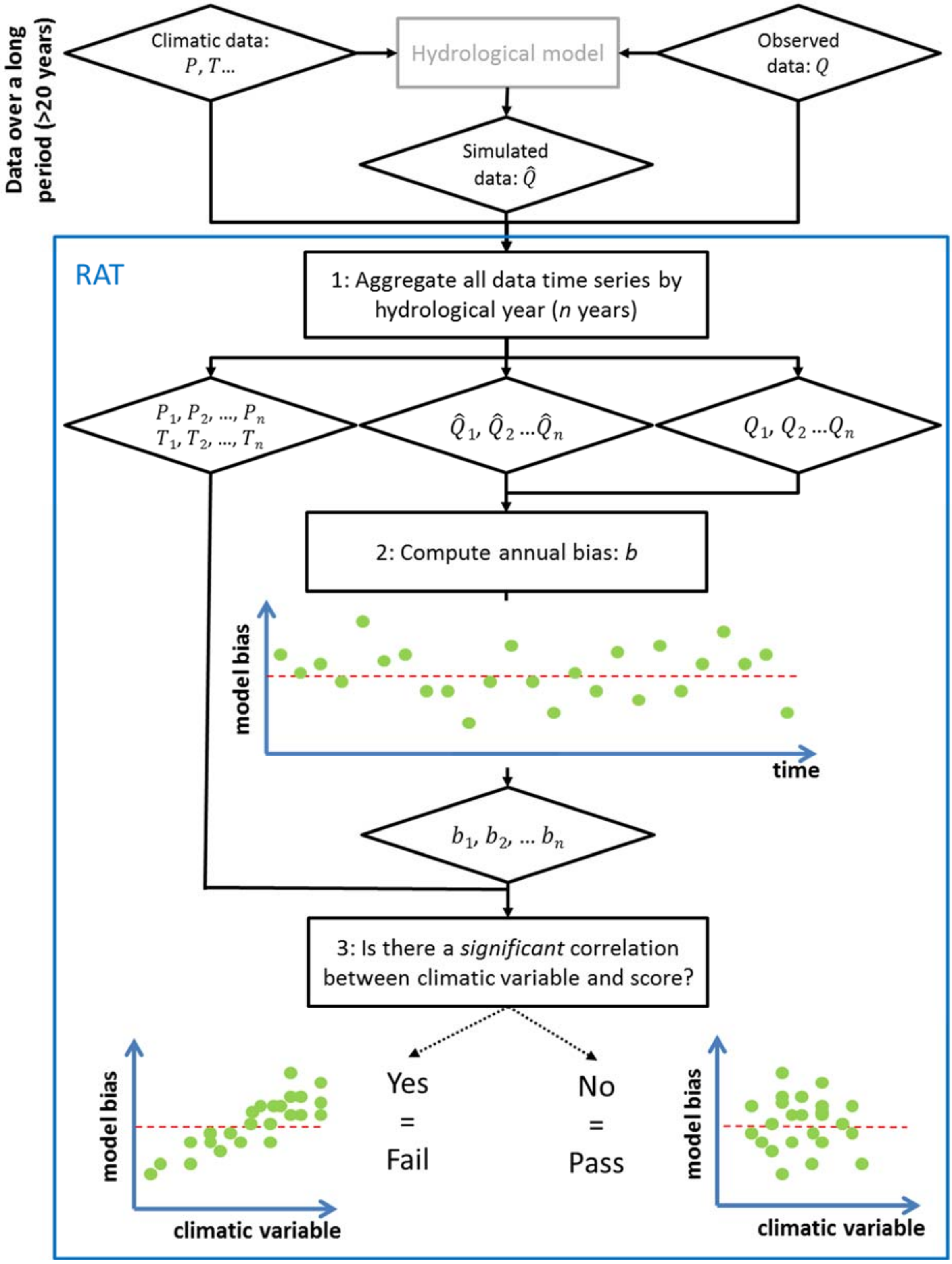

Figure 1. Flow chart of the Robustness Assessment Test

An example is shown in Figure 2, with a daily time step hydrological model calibrated on a 47-year streamflow record. Note that this plot could be obtained from any hydrological model calibrated or not. The relative streamflow bias ($(\overline{Q_{sim}}/\overline{Q_{obs}} - 1)$, with $\overline{Q_{sim}}$ and $\overline{Q_{obs}}$ being the mean simulated and observed streamflows respectively) is calculated on an annual basis (47 values in total). Then, the annual bias values are plotted against climate descriptors, typically the annual temperature absolute anomaly ($T - \bar{T}$, where T is the annual mean and $\bar{T}$ is the long-term mean annual temperature), the annual precipitation relative anomaly $P/\bar{P} - 1$ and the humidity index relative anomaly $HI/\overline{HI} - 1$, where $HI = P/E_0$, E_0 being the potential evaporation). Note that the mean annual values are computed on hydrological years (here from August 1st of year *n-1* to July 31st of year *n*). In this example, there is a slight dependency of model bias on precipitation and humidity index. Clearly, this could be a problem if we were to use this model in an extrapolation mode.

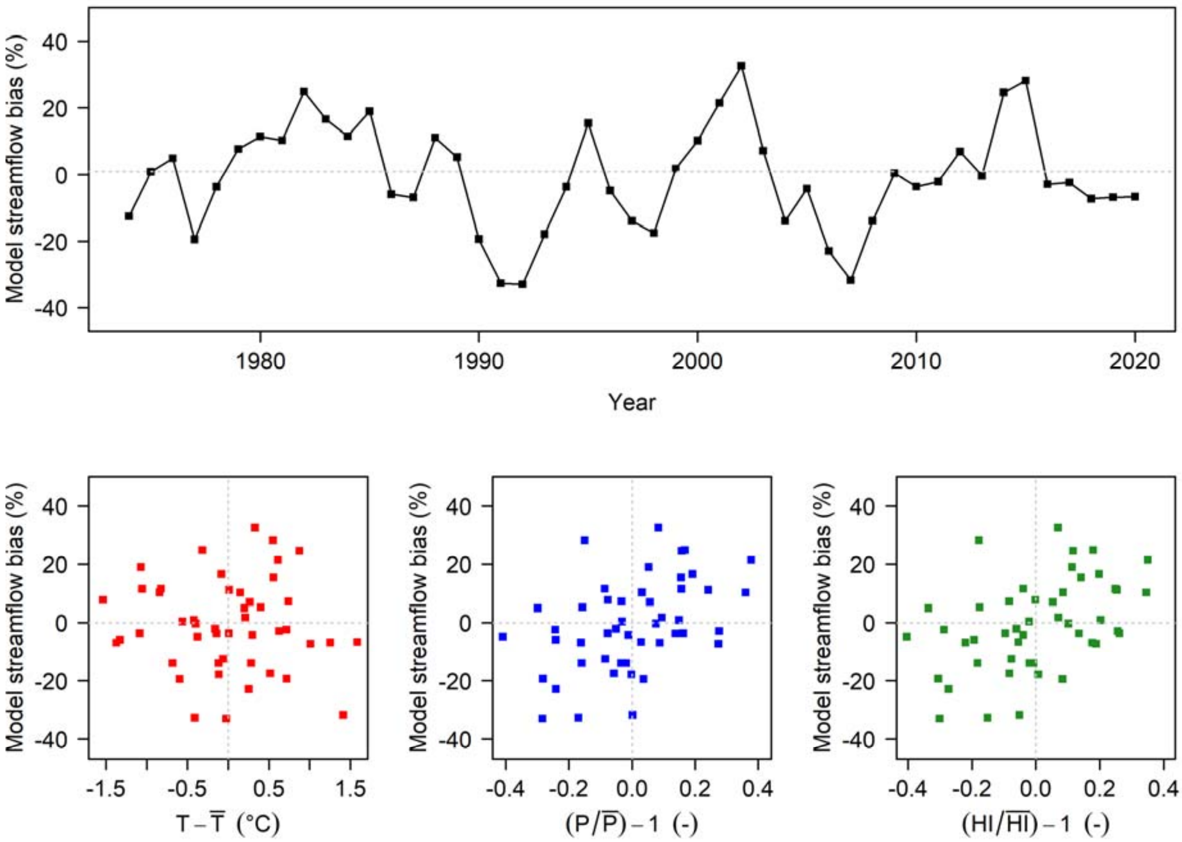

Figure 2. Robustness Assessment Test (RAT) applied to a hydrological model: the upper graph presents the evolution in time (year by year) of model streamflow bias; the lower scatterplots present the relationship between model bias and climatic variables (temperature *T*, precipitation *P* and humidity index *HI*, from left to right)

Whereas the methods based on the split-sample test (i.e. Coron et al, 2012 and Thirel et al., 2015b) evaluate model robustness on periods that are independent of the calibration period, it is not the case for the RAT. Consequently, one could fear that the results of the RAT evaluation may be influenced by the calibration process. However, because the RAT uses a very long period for calibration, we hypothesize that the weight of each individual year in the overall calibration process is small, almost negligible. This assumption can be checked by comparing the RAT with a leave-one-out SST (see Appendix). The analysis showed that this hypothesis is reasonable for long time series, but that the RAT is not applicable when the available time period is too short (less than 20 years).

Last, we would like to mention that the RAT procedure is different from the Proxy metric for Model Robustness (PMR) presented by Royer-Gaspard et al. (2021), even if both methods aim to evaluate hydrological model robustness without employing a multiple calibrations process: the PMR is a simple metric to estimate the robustness of a hydrological model, while the RAT is a method to diagnose the dependencies of model errors to certain types of climatic changes. Thus, the RAT and the PMR may be seen as complementary tools to assess a variety of aspects about model robustness.

3 Material and methods

3.1 Catchment set

We employed the dataset previously used by Nicolle et al. (2014), comprising 21 French catchments (Figure 3), with complementary data until 2020. Catchments were chosen to represent a large range of physical and climatic conditions in France, with sufficiently-long observation time series (daily streamflow from 1974 to 2020) in order to provide a diverse representation of past hydroclimatic conditions. Streamflow data come from the French HYDRO database (Leleu et al., 2014) and with quality control performed by the operational hydrometric services. Catchment size ranges from 380 to 4,300 km² and median elevation from 70 to 1020 m.

The daily precipitation and temperature data originate from the gridded SAFRAN climate reanalysis (Vidal et al., 2010) over the 1959–2020 period. More information about the catchment set can be found in Nicolle et al. (2014). Aggregated catchment files and computation of Oudin potential evaporation (Oudin et al., 2005) was made as described in Delaigue et al. (2018).

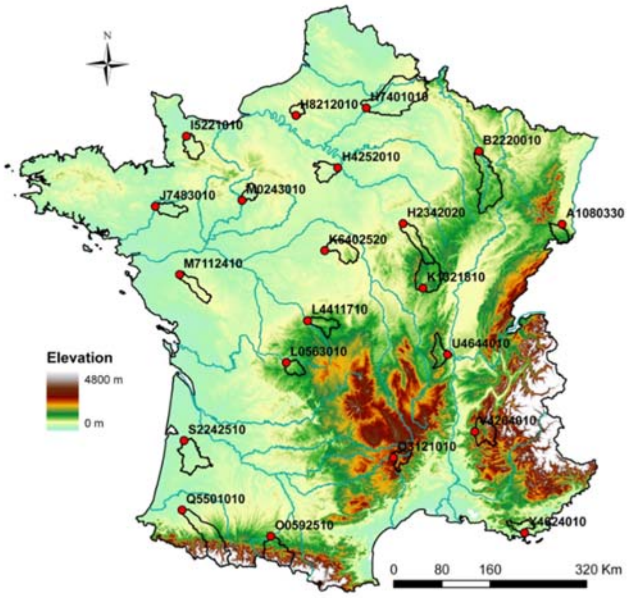

Figure 3. Location of the 21 catchments in France. Red dots represent the catchment outlets

3.2 Hydrological model

The RAT diagnostic framework is generic and can be applied to any type of model. Here daily streamflow was simulated using the daily lumped GR4J rainfall–runoff model (Perrin et al., 2003). The objective function used for calibration is the KGE criterion (Gupta et al., 2009) computed on square-root-transformed flows. Model implementation was done with the airGR R package (Coron et al., 2017, 2018).

3.3 Evaluation of the RAT framework

The RAT was evaluated against the GSST of Coron et al. (2012) used as a benchmark, in order to check whether it yields similar results. The GSST procedure was applied to each catchment using a 10-year period to calibrate the model. For each calibration, each 10-year sliding period over the remaining available period, strictly independent of the calibration one, was used to evaluate the model. The results of the two approaches were compared by plotting on the same graph the annual streamflow bias obtained from the unique simulation period for the RAT, and the average streamflow bias over the sliding calibration-validation time periods for GSST, as a function of temperature, precipitation and humidity anomalies as in Figure 2. The similarity of the trends (between streamflow bias and climatic anomaly) obtained by the two methods was evaluated on the catchment set by comparing the slope and intercept of the linear regressions obtained in each case.

We then identified the catchments where the RAT procedure detected a dependency of streamflow bias to one or several climate variables. The Spearman correlation between model bias and climate variables was computed and a significance threshold of 5% was used (p-value 0.05).

4 Results

4.1 Comparison between the RAT and the GSST procedure

Figure 4 presents an example for the Orge River at Morsang-sur-Orge: GSST points are represented by black dots and RAT points by red squares. Let us first note that since red points represent only each of the N years of the period for the RAT and black points represent all GSST possible independent calibration-validation pairs (a number close to *N(N-1)*), black points are much more numerous. We can observe that the amplitude of both streamflow bias and climatic variable change is larger for the GSST than for the RAT as there are more calibration periods, whatever the climatic variable (P, T or HI). However, the trends in the scatterplot are quite similar. Graphs for all catchments are provided as supplementary material.

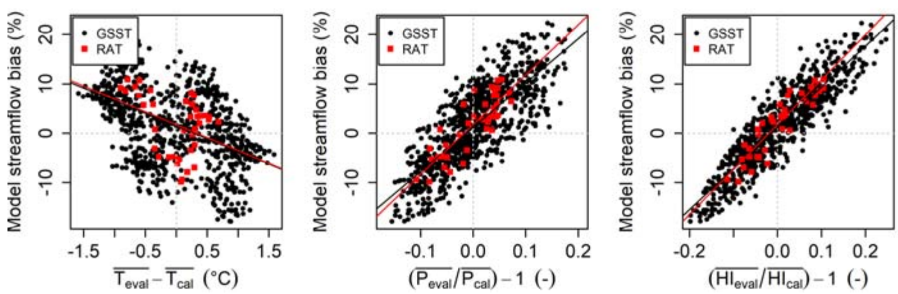

Figure 4. Streamflow bias obtained with the RAT (red squares) and the GSST (black dots), as a function of temperature, precipitation and humidity index anomalies, for the Orge River at Morsang-sur-Orge (H4252010) (934 km²).

To summarize the results on the 21 catchments, we present on Figure 5 the slope and intercept of a linear regression computed between model streamflow bias and climatic variable anomaly, for the GSST and the RAT over the 21 catchments: the slope of the regressions obtained for both methods are very similar and the intercept also exhibits a good match (although somewhat larger differences).

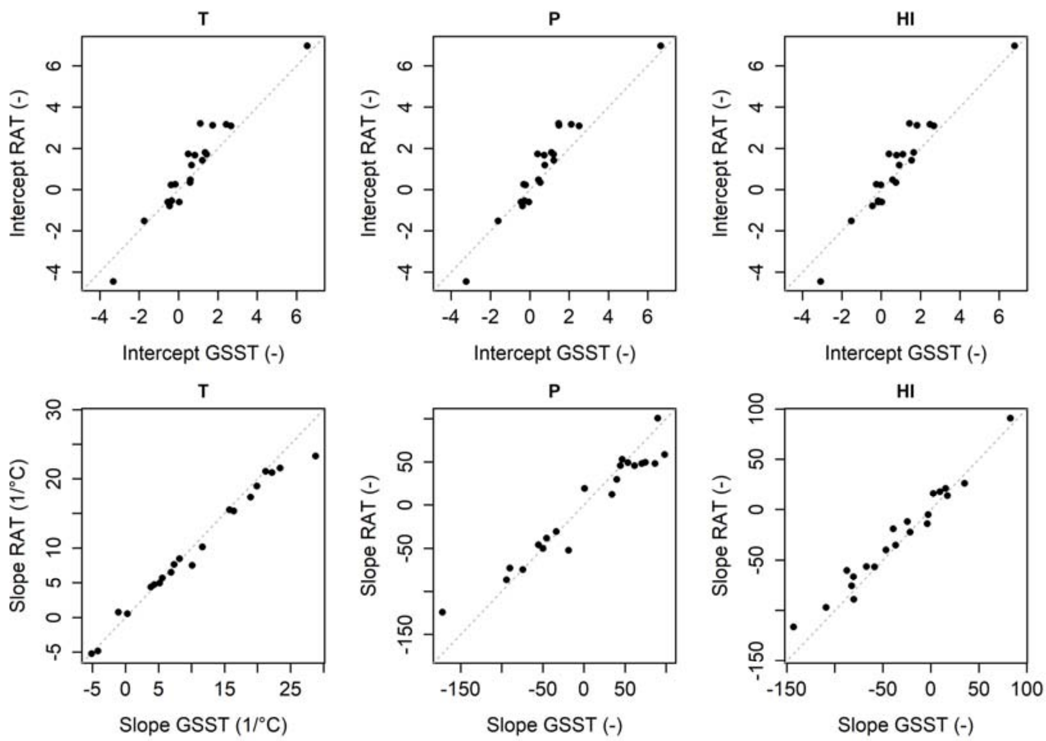

Figure 5. Comparison of slopes and intercept of linear regressions between streamflow bias and temperature (T), precipitation (P) and humidity index (HI) anomalies (from left to right) obtained by the GSST and the RAT procedures (each point represents one of the 21 test catchments)

We can thus conclude that the RAT reproduces the results of GSST, but at a much lower computational cost, and this is what we were aiming at. One should however acknowledge that switching from the GSST to the RAT unavoidably reduces the severity of the climate anomalies we can expose the hydrological models to: indeed, the climate anomalies with the RAT are computed with respect to the mean over the whole period, whilst with the GSST they are computed between two shorter (and hence potentially more different) periods.

4.2 Application of the RAT procedure to the detection of climate dependencies

We now illustrate the different behaviours found among the 21 catchments when applying the RAT procedure. The significance of the link between model bias and climate anomalies was based on the Spearman correlation and a 5 % threshold. Five cases were identified:

1. **No climate dependency** (Figure 6): This is the case for 6 catchments out of 21 and the expected situation of a "robust" model. The different plots show a lack of dependence, for temperature, precipitation and humidity index alike. For the catchment of Figure 6, the p-value of the Spearman correlation is high (between 0.23 and 0.98) and thus not significant.

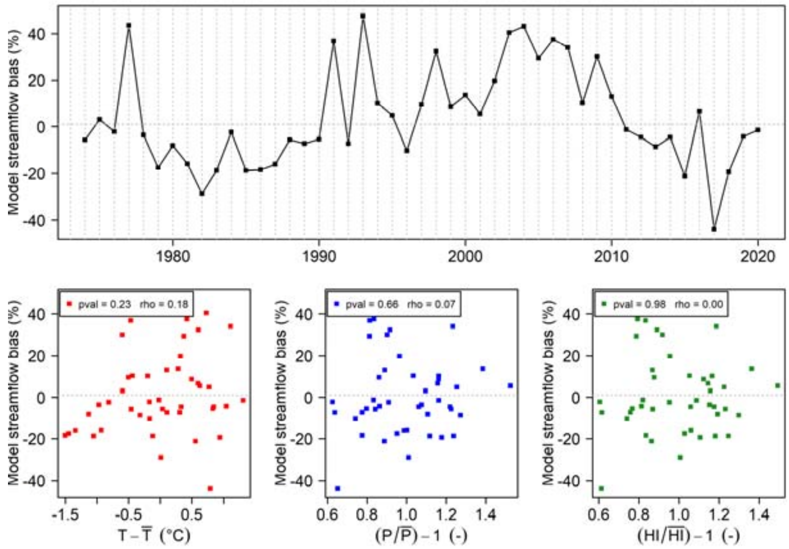

Figure 6. Streamflow annual bias obtained with the RAT function of time (top), temperature absolute anomalies (bottom left), and precipitation P (bottom centre) and humidity index P/E_0 (bottom right) anomalies, for the Orne Saosnoise River at Montbizot (M0243010) (510 km²).

2. **Significant dependency on annual temperature, precipitation and humidity index** (Figure 7): This is a clearly undesirable situation illustrating a lack of robustness of the hydrological model. It happens on only two catchments out of 21. The Spearman correlation between model bias and temperature, precipitation and humidity index anomalies (respectively 0.49, -0.36 and -0.46) is significant (i.e. below the classic significance threshold of 5%). In Figure 7, the annual bias shows an increasing trend with annual temperature and a decreasing trend with annual precipitation and humidity index.

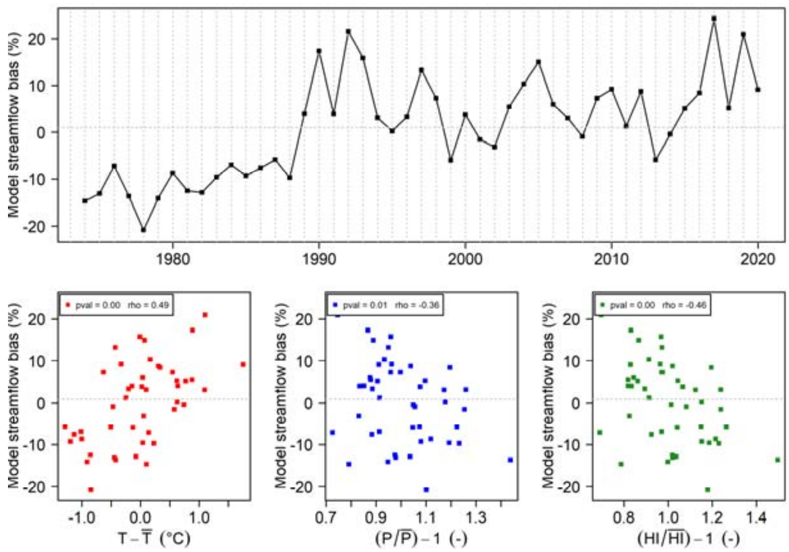

Figure 7. Streamflow annual bias obtained with the RAT function of time (top), temperature absolute anomalies (bottom left), and precipitation P (bottom center) and humidity index P/E_0 (bottom right) anomalies, for the Arroux River at Etang-sur-Arroux (K1321810) (1790 km²)

3. **Significant climate dependency on precipitation and humidity index but not on temperature** (Figure 8). This case happens on 5 of the 21 catchments.

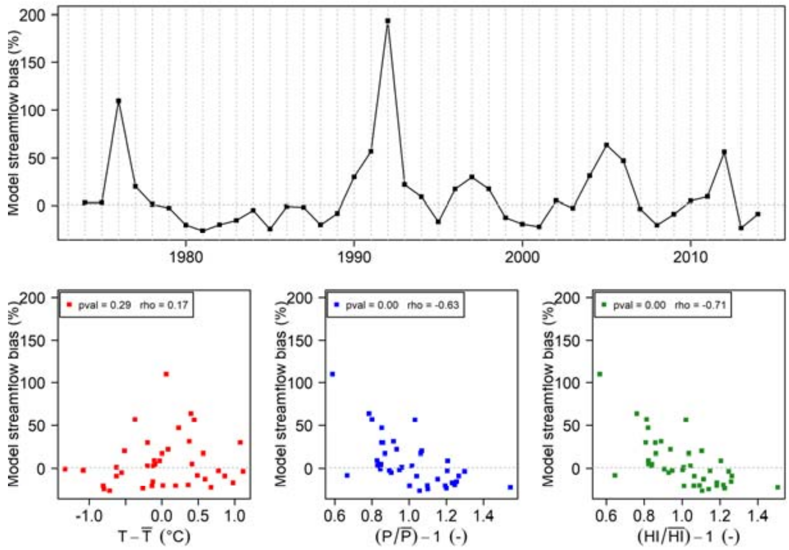

Figure 8. Streamflow annual bias obtained with the RAT function of time (top), temperature absolute anomalies (bottom left), and precipitation P (bottom center) and humidity index P/E_0 (bottom right) anomalies, for the Seiche River at Bruz (J7483010) (810 km²)

4. **Significant climate dependency on temperature but not on precipitation and humidity index** (Figure 9). This case happens on 3 of the 21 catchments.

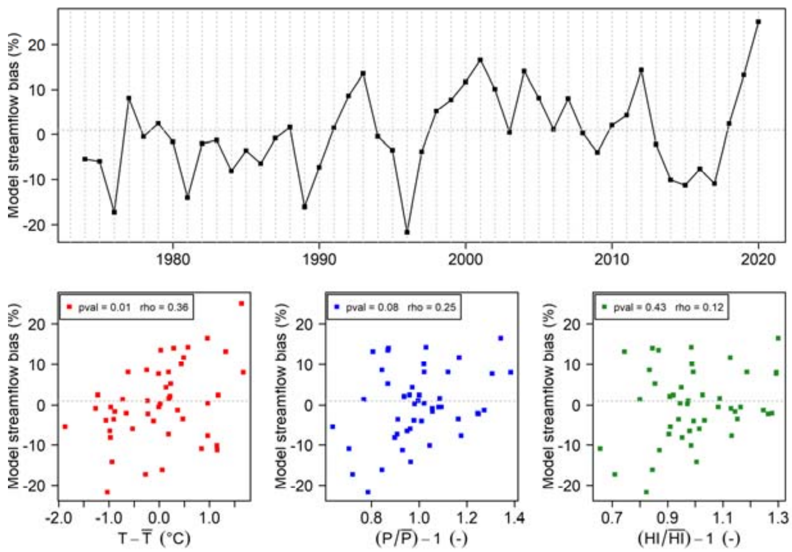

Figure 9. Streamflow annual bias obtained with the RAT function of time (top), temperature absolute changes (bottom left), and precipitation P (bottom center) and humidity index P/E_0 (bottom right) anomalies, for the Ill at Didenheim (A1080330) (670 km²)

5. **Significant climate dependency on temperature and humidity index but not on precipitation** (Figure 10). This case happens on 5 of the 21 catchments.

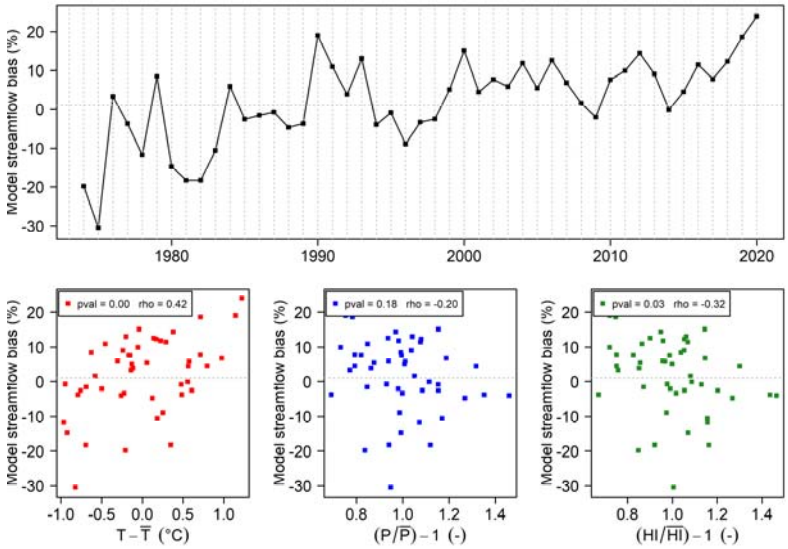

Figure 10. Streamflow annual bias obtained with the RAT function of time (top), temperature absolute changes (bottom left), and precipitation P (bottom center) and humidity index P/E_0 (bottom right) anomalies, for the Briance River at Condat-sur-Vienne (L0563010) (597 km²)

4.3 How to use RAT results?

A question that many modelers may ask us is *what can be done when different types of model failure are identified?* Some of the authors of this paper have long be fond of the concept of Crash test (Andréassian et al., 2009), and we would like to argue here that the RAT too can be seen as a kind of crash-test. As all crash tests, it will end up identifying failures. But the fact that a car may be destroyed when projected against a wall does not mean that it is entirely unsafe, it rather means that it is not entirely safe. Although we are conscious of this, we keep driving cars… but, we are also willing to pay (invest) more for a safer car (even if this safer-and-more-expensive toy did also ultimately fail the crash test). We believe that the same will occur with hydrological models: The RAT may help identify safer models, or safer ways to parameterize models. If applied on large datasets, it may help identify model flaws, and thus help us work to eliminate them. It will not however help identify perfect models: these do not exist.

5 Conclusion

The proposed robustness assessment test (RAT) is an easy-to-implement evaluation framework that allows robustness evaluation from all types of hydrological models to be compared, by using only one long period for which model simulations are available. The RAT consists in identifying undesired dependencies of model errors to the variations of some climate variables over time. Such dependencies can indeed be detrimental for model performance in a changing climate context. This test can be particularly useful for climate change impact studies where the robustness of hydrological models is often not evaluated at all: as such, our test can help users to discriminate alternative models and select the most reliable models for climate change studies, which ultimately should reduce uncertainties on climate change impact predictions (Krysanova et al., 2018).

The proposed test has obviously its limits, and a first difficulty that we see in using the RAT is that it is only applicable in cases where the hypothesis of independence between the 1-year subperiods and the whole period is sufficient. This is the case when long series are available (at least 20 years, see last graph in appendix). If it is not the case, the RAT procedure should not be used. Therefore, we would indeed recommend its use in cases where modellers cannot "afford" multiple calibrations, or where the parameterisation strategy is considered (by the modeller) as 'calibration free' (i.e. physically-based models). A few other limitations should be mentioned:

1. In this note, the RAT concept was illustrated with a rank-based test (Spearman correlation) and a significance threshold of 0.05. Like all thresholds, this one is arbitrary. Moreover, other non-parametric tests could be used and would probably yield slightly different results (we also tested the Kendall tau test, with very similar results, but do not show the results here);
2. 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, because even a robust model will be affected by a trend in input data, yielding the impression that the hydrological model lacks robustness. Such an erroneous conclusion could also be due to widespread changes in land use, construction of an unaccounted storage reservoir or the evolution of water uses. Some of the lacks of robustness detected among the 21 catchments presented here could be in fact due to metrological causes;
3. Also, because of the ongoing rise of temperatures (over the last 40 years at least), we have a correlation between temperature and time since the beginning of streamgaging. If for any reason, time is having an impact on model bias, this may cause an artefact in the RAT in the form of a dependency between model bias and temperature;
4. Similarly to the Differential Split Sample Test, the diagnostic of model climatic robustness is limited to the climatic variable against which the bias is compared. As such, the RAT should not be seen as an *absolute* test, but rather as a *necessary but not sufficient* condition to use a model for climate change studies: because the climatic variability present in the past observations is limited to the historic range, so is the extrapolation test. With Popper's words (Popper, 1959), the RAT can only allow falsifying a hydrological model… but not proving it right;
5. Although it would be tempting to transform the RAT into a post-processing method, we do not recommend it. Indeed, detecting a relationship between model bias and a climate variable using the RAT does not necessarily mean that a simple (linear) debiasing solution can be proposed to solve the issue (see e.g. the paper by Bellprat et al. (2013) on this topic). What we do recommend is to work as much as possible on the model structure, to turn it less climate dependent;
6. Some of the modalities of the RAT, that we initially thought of importance, are not really important: this is for example the case with the use of hydrological years. We tested the twelve possible annual aggregations schemes (see https://doi.org/10.5194/hess-2021-147-AC6) and found no significant impact;
7. Upon recommendation by one of the reviewers, we tried to assess the possible impact of the quality of the precipitation forcing on RAT results (see https://doi.org/10.5194/hess-2021-147-AC5) and found that the type of forcing used does have an impact on RAT results (interestingly, the climatic dataset yielding the best simulation results was also the dataset yielding the less catchments failing the robustness test). It seems unavoidable that forcing data quality will impact the results of RAT, but we would argue that it would similarly have an impact on the results of a Differential Split Sample Test. We believe that there is no way to avoid entirely this

dependency, and that evaluating the quality of input data should be done before looking at model robustness;

8. Last, we could mention that a model showing a small overall annual bias (but linked to a climate variable) could still be preferred to one showing a large overall annual bias (but independent of the tested climate variables): the RAT should not be seen as the only basis for model choice.

Beyond the limitations, we also see the perspective for further development of the method: although this note only considered overall model bias (as the most basic requirement for a model to be used to predict the impact of a future climate), we think that this 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. Further, while we only tested the dependency to mean annual temperature, precipitation and humidity index, other characteristics, such as precipitation intensity or fraction of snowfall, could be considered in this framework.

6 Acknowledgments

This work was funded by the project AQUACLEW, which is part of ERA4CS, an ERA-NET initiated by JPI Climate, and funded by FORMAS (SE), DLR (DE), BMWFW (AT), IFD (DK), MINECO (ES), ANR (FR) with co-funding by the European Commission [Grant 690462].

The authors gratefully acknowledge the comments of Prof. Jens-Christian Refsgaard and Dr Nans Addor on a preliminary version of the note, and the reviews by Dr Bettina Schäfli and by two anonymous reviewers.

The gridded SAFRAN climate reanalysis data can be ordered from Météo-France.

Observed streamflow data are available on the French HYDRO database (http://www.hydro.eaufrance.fr/).

The GR models, including GR4J, are available from the airGR *R* package.

7 References

Andréassian, V., Le Moine, N., Perrin, C., Ramos, M.-H., Oudin, L., Mathevet, T., Lerat, J., Berthet, L (2012). All that glitters is not gold: the case of calibrating hydrological models. Hydrological Processes, 26(14), 2206–2210. https://doi.org/10.1002/hyp.9264

Andréassian, V., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., Loumagne, C., Oudin, L., Mathevet, T., Ramos, M.H., Valéry, A. (2009). Crash tests for a standardized evaluation of hydrological models. Hydrology and Earth System Sciences, 13, 1757-1764. https://doi.org/10.5194/hess-13-1757-2009

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics Surveys, 4(0), 40–79. https://doi.org/10.1214/09-SS054

Bellprat, O., Kotlarski, S., Lüthi, D., & Schär, C. (2013). Physical constraints for temperature biases in climate models: limits of temperature biases. Geophysical Research Letters, 40(15), 4042–4047. https://doi.org/10.1002/grl.50737

Beven, K. (2016). Facets of uncertainty: epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication. Hydrological Sciences Journal, 61(9), 1652–1665. https://doi.org/10.1080/02626667.2015.1031761

Bisselink, B., Zambrano-Bigiarini, M., Burek, P., & de Roo, A. (2016). Assessing the role of uncertain precipitation estimates on the robustness of hydrological model parameters under highly variable climate conditions. Journal of Hydrology: Regional Studies, 8, 112–129. https://doi.org/10.1016/j.ejrh.2016.09.003

Blöschl, G., et al. 2019. Twenty-three Unsolved Problems in Hydrology – a community perspective. Hydrological Sciences Journal, https://doi.org/10.1080/02626667.2019.1620507

Brigode, P., et al. (2015). Dependence of model-based extreme flood estimation on the calibration period: case study of the Kamp River (Austria). Hydrological Sciences Journal, 60 (7–8). https://doi.org/10.1080/02626667.2015.1006632

Broderick, C., Matthews, T., Wilby, R. L., Bastola, S., & Murphy, C. (2016). Transferability of hydrological models and ensemble averaging methods between contrasting climatic periods. Water Resources Research, 52(10), 8343–8373. https://doi.org/10.1002/2016WR018850

Coron, L., Andréassian, V., Bourqui, M., Perrin, C., & Hendrickx, F. (2011). Pathologies of hydrological models used in changing climatic conditions: a review. In S. Franks, E. Boegh, E. Blyth, D. Hannah, & K. K. Yilmaz (Eds.), Hydro-climatology: Variability and Change. IAHS Red Books Series 344 (pp. 39–44). Wallingford: IAHS.

Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., & Hendrickx, F. (2012). Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments. Water Resources Research, 48, W05552. https://doi.org/10.1029/2011WR011721

Coron, L., Andréassian, V., Perrin, C., Bourqui, M., & Hendrickx, F. (2014). On the lack of robustness of hydrologic models regarding water balance simulation: a diagnostic approach applied to three models of increasing complexity on 20 mountainous catchments. Hydrol. Earth Syst. Sci., 18(2), 727–746.

Coron, L., Thirel, G., Delaigue, O., Perrin, C., & Andréassian, V. (2017). The Suite of Lumped GR Hydrological Models in an R package. Environmental Modelling and Software, 94, 337. https://doi.org/10.1016/j.envsoft.2017.05.002

Coron, L., Delaigue, O., Thirel, G., Perrin, C. and Michel, C. (2020). airGR: Suite of GR Hydrological Models for Precipitation-Runoff Modelling. R package version 1.4.3.65. DOI: 10.15454/EX11NA. URL: https://CRAN.R-project.org/package=airGR.

Dakhlaoui, H., Ruelland, D., Tramblay, Y., & Bargaoui, Z. (2017). Evaluating the robustness of conceptual rainfall-runoff models under climate variability in northern Tunisia. Journal of Hydrology, 550, 201–217. https://doi.org/10.1016/j.jhydrol.2017.04.032

Dakhlaoui, H, Ruelland, D., & Tramblay, Y. (2019). A bootstrap-based differential split-sample test to assess the transferability of conceptual rainfall-runoff models under past and future climate variability. Journal of Hydrology, 575, 470–486. https://doi.org/10.1016/j.jhydrol.2019.05.056

Delaigue, O., Génot, Lebecherel, L., Brigode, P., & Bourgin, P. Y. (2018). Base de données hydroclimatiques observées à l'échelle de la France. IRSTEA. IRSTEA, UR HYCAR, Équipe Hydrologie des bassins versants, Antony. Retrieved from https://webgr.inrae.fr/en/activities/database-1-2/

Donelly-Makowecki, L. M., & Moore, R. D. (1999). Hierarchical testing of three rainfall-runoff models in small forested catchments. Journal of Hydrology, 219(3–4), 136–152.

Efstratiadis, A., Nalbantis, I., and Koutsoyiannis, D. (2015). Hydrological modelling of temporally-varying catchments: facets of change and the value of information. Hydrological Sciences Journal, 60 (7–8). https://doi.org/10.1080/02626667.2014.982123

Fowler, K., Coxon, G., Freer, J., Peel, M., Wagener, T., Western, A., et al. (2018). Simulating Runoff Under Changing Climatic Conditions: A Framework for Model Improvement. Water Resources Research, 54(12), 9812–9832. https://doi.org/10.1029/2018WR023989

Gaborit, É., Ricard, S., Lachance-Cloutier, S., Anctil, F., & Turcotte, R. (2015). Comparing global and local calibration schemes from a differential split-sample test perspective. Canadian Journal of Earth Sciences, 52(11), 990–999. https://doi.org/10.1139/cjes-2015-0015

Gelfan, A.N., & Millionshchikova, T.D. (2018). Validation of a Hydrological Model Intended for Impact Study: Problem Statement and Solution Example for Selenga River Basin. Water Resour 45, 90–101. https://doi.org/10.1134/S0097807818050354

Gelfan, A., et al. (2015). Testing robustness of the physically-based ECOMAG model with respect to changing conditions. Hydrological Sciences Journal, 60 (7–8). https://doi.org/10.1080/02626667.2014.935780

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of Hydrology, 377(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Klemeš, V. (1986). Operational testing of hydrologic simulation models. Hydrological Sciences Journal, 31(1), 13–24. https://doi.org/10.1080/02626668609491024

Krysanova, V., Donnelly, C., Gelfan, A., Gerten, D., Arheimer, B., Hattermann, F., & Kundzewicz, Z. W. (2018). How the performance of hydrological models relates to credibility of projections under climate change. Hydrological Sciences Journal, 63(5), 696–720. https://doi.org/10.1080/02626667.2018.1446214

Hughes, D.A. (2015). Simulating temporal variability in catchment response using a monthly rainfall–runoff model. Hydrological Sciences Journal, 60 (7–8). https://doi.org/10.1080/02626667.2014.909598

Kling, H., et al. (2015). Performance of the COSERO precipitation–runoff model under non-stationary conditions in basins with different climates. Hydrological Sciences Journal, 60 (7–8). https://doi.org/10.1080/02626667.2014.959956

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. Journal of Educational Psychology, 22(1), 45–55. https://doi.org/10.1037/h0072400

Leleu, I., Tonnelier, I., Puechberty, R., Gouin, P., Viquendi, I., Cobos, L., et al. (2014). La refonte du système d'information national pour la gestion et la mise à disposition des données hydrométriques. La Houille Blanche, (1), 25–32. https://doi.org/10.1051/lhb/2014004

Li, C. Z., Zhang, L., Wang, H., Zhang, Y. Q., Yu, F. L., and Yan, D. H. (2012). The transferability of hydrological models under nonstationary climatic conditions, Hydrol. Earth Syst. Sci., 16, 1239–1254, https://doi.org/10.5194/hess-16-1239-2012

Li, H., Beldring, S., and Xu, C.-Y. (2015). Stability of model performance and parameter values on two catchments facing changes in climatic conditions. Hydrological Sciences Journal, 60 (7–8). https://doi.org/10.1080/02626667.2014.978333

Magand, C., et al. (2015). Parameter transferability under changing climate: case study with a land surface model in the Durance watershed, France. Hydrological Science Journal, 60 (7–8). https://doi.org/10.1080/02626667.2014.993643

Montanari, A., Young, G., Savenije, H. H. G., Hughes, D., Wagener, T., Ren, L. L., et al. (2013). “Panta Rhei—Everything Flows”: Change in hydrology and society—The IAHS Scientific Decade 2013–2022. Hydrological Sciences Journal, 58(6), 1256–1275. https://doi.org/10.1080/02626667.2013.809088

Mosteller, F., & Tukey, J. W. (1968). Data Analysis, Including Statistics. The Collected Works of John W. Tukey (1988) Graphics pp. 1965-1985, vol. 5 (123)

Motavita, D.F., R. Chow, A. Guthke & W. Nowak. (2019). The comprehensive differential split-sample test: A stress-test for hydrological model robustness under climate variability, Journal of Hydrology, 573: 501-515, https://doi.org/10.1016/j.jhydrol.2019.03.054

Nicolle, P., Pushpalatha, R., Perrin, C., François, D., Thiéry, D., Mathevet, T., et al. (2014). Benchmarking hydrological models for low-flow simulation and forecasting on French catchments. Hydrol. Earth Syst. Sci., 18(8), 2829–2857. https://doi.org/10.5194/hess-18-2829-2014

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., & Loumagne, C. (2005). Which potential evapotranspiration input for a lumped rainfall–runoff model?: Part 2-Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling. Journal of Hydrology, 303(1–4), 290–306. https://doi.org/10.1016/j.jhydrol.2004.08.026

Perrin, C., Michel, C., & Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. Journal of Hydrology, 279(1–4), 275–289. https://doi.org/10.1016/S0022-1694(03)00225-7

Popper, K. (1959). The logic of scientific discovery. London: Routledge.

Rau, P., Bourrel, L., Labat, D., Ruelland, D., Frappart, F., Lavado, W., et al. (2019). Assessing multidecadal runoff (1970–2010) using regional hydrological modelling under data and water scarcity conditions in Peruvian Pacific catchments. Hydrological Processes, 33(1), 20–35. https://doi.org/10.1002/hyp.13318

Refsgaard, J. C., & Henriksen, H. J. (2004). Modelling guidelines––terminology and guiding principles. Advances in Water Resources, 27, 71–82. https://doi.org/10.1016/j.advwatres.2003.08.006

Refsgaard, J. C., & Knudsen, J. (1996). Operational validation and intercomparison of different types of hydrological models. Water Resources Research, 32(7), 2189–2202. https://doi.org/10.1029/96WR00896

Refsgaard, J. C., Madsen, H., Andréassian, V., Arnbjerg-Nielsen, K., Davidson, T. A., Drews, M., et al. (2013). A framework for testing the ability of models to project climate change and its impacts. Climatic Change, 122(1–2), 271–282. https://doi.org/10.1007/s10584-013-0990-2

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913–929. https://doi.org/10.1111/ecog.02881

Royer-Gaspard, P., Andréassian, V., and Thirel, G. (2021) Technical note: PMR – a proxy metric to assess hydrological model robustness in a changing climate. In review for Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-58

Seibert, J. (2003). Reliability of model predictions outside calibration conditions. Nordic Hydrology, 34(5), 477–492. https://doi.org/10.2166/nh.2003.0019

Seifert, D., Sonnenborg, T. O., Refsgaard, J. C., Højberg, A. L., and Troldborg, L. (2012). Assessment of hydrological model predictive ability given multiple conceptual geological models, Water Resour. Res., 48, W06503, doi:10.1029/2011WR011149.

Seiller, G., Anctil, F., & Perrin, C. (2012). Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions. Hydrology and Earth System Sciences, 16(4), 1171–1189. https://doi.org/10.5194/hess-16-1171-2012

Tanaka, T. and Tachikawa, Y. (2015). Testing the applicability of a kinematic wave-based distributed hydrological model in two climatically contrasting catchments. Hydrological Sciences Journal, 60 (7–8). https://doi.org/10.1080/02626667.2014.967693

Taver, V., et al. (2015). Feed-forward vs recurrent neural network models for non-stationarity modelling using data assimilation and adaptivity. Hydrological Sciences Journal, 60 (7–8). https://doi.org/10.1080/02626667.2014.967696

Teutschbein, C. & Seibert, J. 2013. Is bias correction of regional climate model (RCM) simulations possible for non-stationary conditions? Hydrology and Earth System Sciences, 17, 5061–5077., https://doi.org/10.5194/hess-17-5061-2013

Thirel, G., Andréassian, V., Perrin, C., Audouy, J.-N., Berthet, L., Edwards, P., et al. (2015a). Hydrology under change: an evaluation protocol to investigate how hydrological models deal with changing catchments. Hydrological Sciences Journal, 60(7–8), 1184–1199. https://doi.org/10.1080/02626667.2014.967248

Thirel, G., Andréassian, V., & Perrin, C. (2015b). On the need to test hydrological models under changing conditions. Hydrological Sciences Journal, 60(7–8), 1165–1173. https://doi.org/10.1080/02626667.2015.1050027

Vaze, J., Post, D. A., Chiew, F. H. S., Perraud, J. M., Viney, N. R., & Teng, J. (2010). Climate non-stationarity - Validity of calibrated rainfall-runoff models for use in climate change studies. Journal of Hydrology, 394(3–4), 447–457. https://doi.org/10.1016/j.jhydrol.2010.09.018

Vidal, J.-P., Martin, E., Franchistéguy, L., Baillon, M., & Soubeyroux, J.-M. (2010). A 50-year high-resolution atmospheric reanalysis over France with the Safran system. International Journal of Climatology, 30(11), P. 1627–1644. DOI: 10.1002/joc.2003. https://doi.org/10.1002/joc.2003

Vormoor, K., Heistermann, M., Bronstert, A., & Lawrence, D. (2018). Hydrological model parameter (in)stability – "crash testing" the HBV model under contrasting flood seasonality conditions. Hydrological Sciences Journal, 63(7), 991–1007. https://doi.org/10.1080/02626667.2018.1466056

Wilby, R. L. (2019). A global hydrology research agenda fit for the 2030s. Hydrology Research, 50(6): 1464-1480. https://doi.org/10.2166/nh.2019.100

Xu, C. (1999). Climate Change and Hydrologic Models: A Review of Existing Gaps and Recent Research Developments. Water Resources Management, 13(5), 369–382. https://doi.org/10.1023/A:1008190900459

Yu, B. and Zhu, Z. (2015). A comparative assessment of AWBM and SimHyd for forested watersheds. Hydrological Sciences Journal, 60 (7–8). https://doi.org/10.1080/02626667.2014.961924

8 Appendix – Checking the impact of the partial overlap between calibration and validation periods in the RAT

In this appendix, we deal with calibrated models, for which we verify that the main hypothesis underlying the RAT is reasonable, i.e. that when considering a long calibration period, the weight of each individual year in the overall calibration process is almost negligible. We then explore the limits of this hypothesis when reducing the length of the overall calibration period.

- **Evaluation method**

In order to check the impact of the partial overlap between calibration and validation periods in the RAT, it is possible (provided one works with a calibrated model) to compare the RAT with a "leave-one out" version of it, which is a classical variant of the Split Sample Test (SST): instead of computing the annual bias after a single calibration encompassing the whole period (RAT), we compute the annual bias with a different calibration each time, encompassing the whole period minus the year in question ("leave-one-out SST").

The comparison between the RAT and the SST can be quantified using the root mean square difference (RMSD) of annual biases:

$$RMSD_{Bias} = \sqrt{\overline{\left(Bias_{RAT} - Bias_{SST}\right)^2}} \qquad \text{Eq.1}$$

where $Bias_{RAT}$ is the bias of validation year n when calibrating the model over the entire period (RAT procedure), and $Bias_{SST}$ the bias of validation year n when calibrating the model over the entire period minus year n (leave-one-out SST procedure).

The difference between the two approaches is schematized in Figure 11: the leave-one-out procedure consists in performing N calibrations over (N-1)-year-long periods followed by an independent evaluation on the remaining 1-year-long period. As shown in Figure 11, the two procedures result in the same number of validation points (N). Eq. 1 provides a way to quantify whether both methods differ, i.e. whether the partial overlap between calibration and validation periods in the RAT makes a difference.

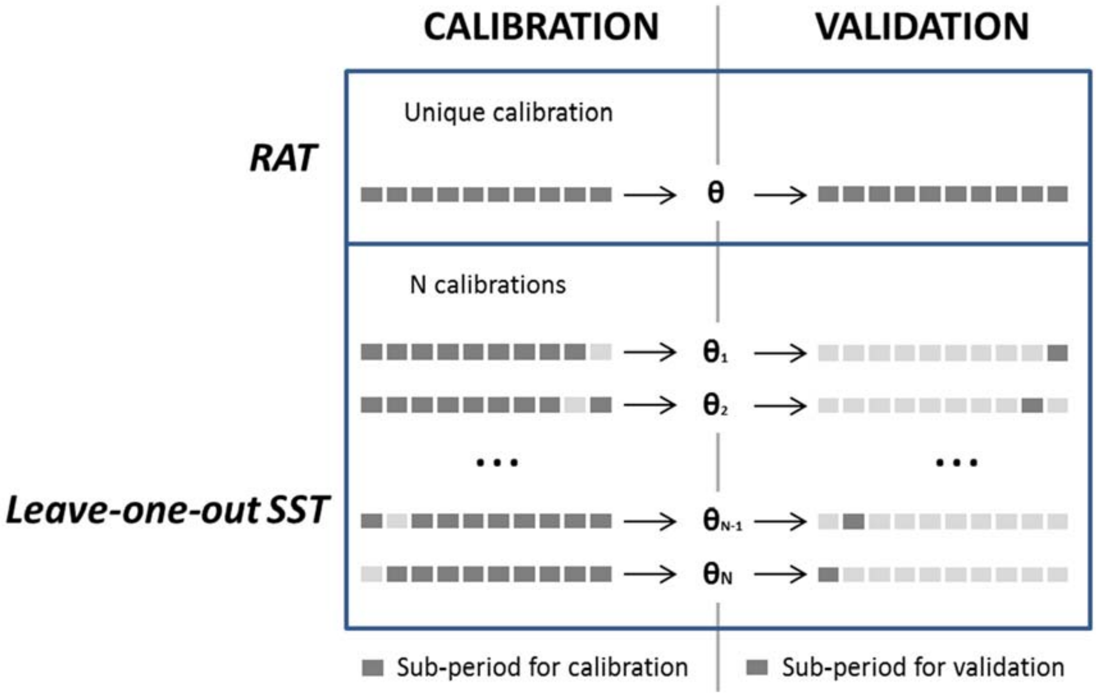

Figure 11. Comparison of the RAT procedure with a leave-one-out split-sample test (SST). Both methods have N validation periods (one per year). The RAT needs only one calibration, whereas the SST requires N calibrations. Dark grey squares represent the years used for calibration or validation

- **Comparison between the RAT and the leave-one-out SST**

Figure 12 plots the annual bias values obtained with the RAT versus the annual bias obtained with the leave-one-out SST for the 21 test catchments, showing a total of 21x47 points. The almost perfect alignment confirms that our underlying "negligibility" hypothesis is reasonable (at least on our catchment set).

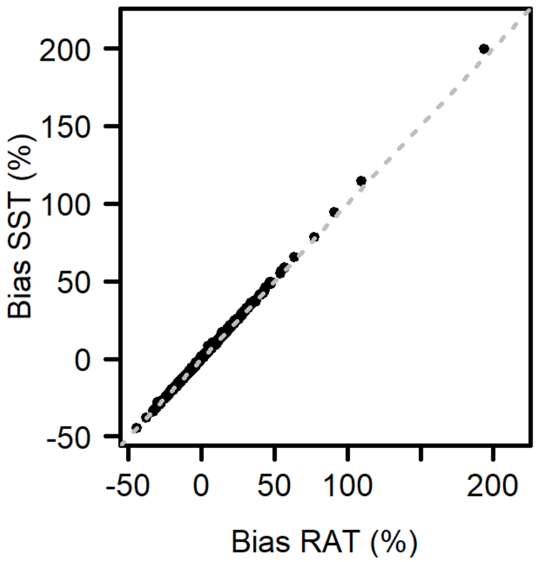

Figure 12. Comparison of the annual bias obtained with the RAT with the annual bias obtained with the leave-one-out SST. Each of the 21 catchments is represented with annual bias values (47 points by catchment, 21x47 points in total)

Figure 13 presents the Spearman correlation p-values for the correlation between annual bias and changes in annual temperature, precipitation, and humidity index (*P/E_0*), for the RAT and the leave-one-out SST. The results from the RAT and the SST show the same dependencies on climate variables (similar *p*-values).

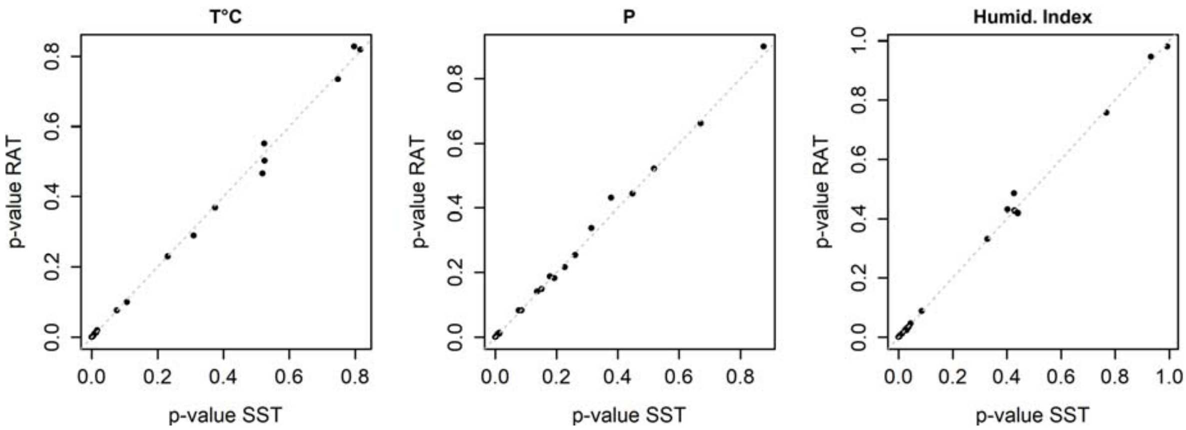

Figure 13. Spearman correlation *p*-value from the correlation for annual bias and annual temperature, precipitation, and humidity index (P/E_0). Comparison between RAT and SST (one point per catchment)

- **Sensitivity of the RAT procedure to the period length**

It is also interesting to investigate the limit of our hypothesis (i.e. that the relative weight of one year within a long time series is very small) by progressively reducing the period length: indeed, the shorter the data series available to calibrate the model, the more important the relative weight of each individual year. Figure 14 compares the annual bias obtained with the RAT procedure with the annual bias obtained with the leave-one-out SST, for 10-, 20-, 30-, and 40-year period lengths (selection of the shorter periods was realized by sampling 10, 20, 30, and 40 years regularly among the complete time series). The shorter the calibration period, the larger the differences between both approaches (wider points scatter): there, we reach the limit of the single calibration procedure. We would not advise to use RAT with time series of less than 20 years.

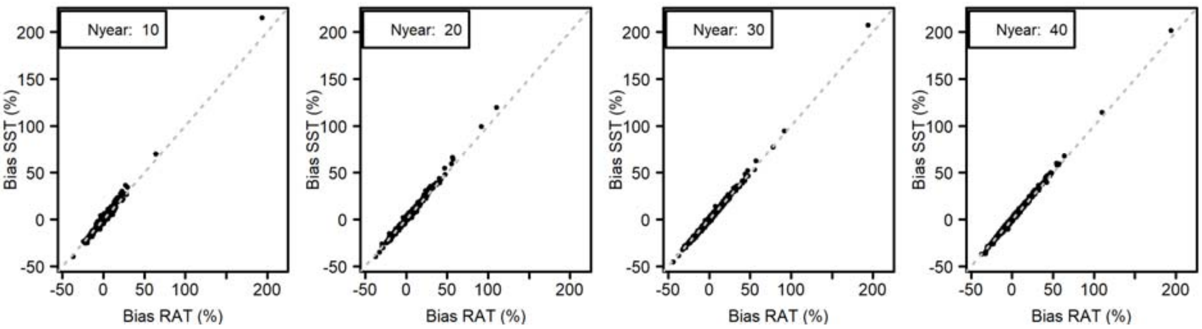

Figure 14. Annual bias obtained with the RAT procedure vs. annual bias obtained with leave-one-out SST. Shorter time periods are obtained by sampling 10, 20, 30, and 40 years regularly among the complete time series. Each of the 21 catchments is represented with annual bias values

These differences can be quantitatively measured by computing the RMSD (see Eq.1) between the annual bias obtained with the RAT procedure and with the SST for different calibration period lengths (see Figure 15). The RMSD tends to increase when the number of years available to calibrate the model decreases, but it seems to be stable for periods longer than 20 years.

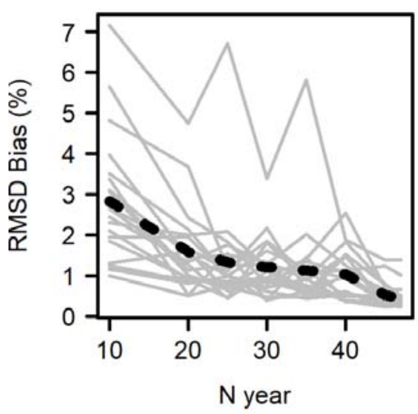

Figure 15. RMSD between annual bias obtained with the RAT procedure and with the leave-one-out SST for different calibration period lengths for each catchment. The dotted line represents the mean RMSD for all catchments. Each grey line represents one of the 21 catchments.