Following the suggestion of reviewer 1, we applied RAT with two different precipitation products. We used the CAMELS data set in the USA (Addor et al., 2017):

Table 1 : characteristics of the CAMELS dataset

Number of catchments	**673**
forcing 1	Daymet, daily 1km grid derived solely from temperature and precipitation observations extrapolated through geostatistics dependent of local station) with quality control
forcing 2	NLDAS (National Land Data Assimilation System) 12 km grid product based on North American Regional Reanalysis upscaled using 4 land surface models and adjusted using CPC for precipitations.
period	1980 to 2014 (5 years for warm-up, calibration between 1985 and 2014)

The model used was the same as in our paper (GR4J + Cemaneige snow accounting routine), we used the KGE on the square root of the discharge as objective function and the Oudin formula for PET. RAT was applied on the simulated time series 1985-2014 for both forcing data product. Results are shown in Table 2

Table 2 : number of catchments considered reactive by RAT

Meteorological product	**Average KGE of GR4J**	**Median KGE of GR4J**	**Number of catchments that react to RAT**	
			(predictor: temperature)	**(predictor: precipitation)**
Daymet	0.678	0.775	92	189
NLDAS	0.641	0.739	117	222
Number of catchments which react with both products			25	123

Model performance was better for Daymet (median KGE = 0.775 instead of 0.739; and mean KGE = 0.678 instead of 0.641). Obviously, the type of forcing used does have an impact on RAT results. It is interesting to note that the climatic dataset yielding the best simulation results is also the dataset yielding the less “reactive” catchments (22 % less for temperature and 15% less for precipitation).

It seems unavoidable that forcing data quality will impact the results of RAT since it has a huge impact on model simulation. It would similarly have an impact on the results of a Differential Split Sample Test. We would argue that there is no way to avoid entirely this dependency, evaluating the quality of input data should be done before looking at model robustness. To conclude, this dependency to data quality is not a sufficient reason to say that RAT is useless, even if it is a sufficient reason… to encourage modelers to take the test results with care and complement the RAT analysis with a discussion of data quality.

References

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrol. Earth Syst. Sci., 21, 5293–5313, https://doi.org/10.5194/hess-21-5293-2017, 2017.