Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-192-RC3, 2020

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

Interactive comment on "Flood hazard and change impact assessments may profit from rethinking model calibration strategies" *by* Manuela I. Brunner et al.

Anonymous Referee #3

Received and published: 2 July 2020

Though I agree with the title of the manuscript and with the main conclusions on l. 254-259 (see below), there are several serious deficiencies in the approach and interpretation of results. The study looks like an initial stage only: let us calibrate four models for many relatively small catchments in USA using one metric, EKG, to see, how well the models will reproduce local flood characteristics and spatial aspects of flooding, and how well would they be prepared for climate impact assessment. The conclusion is that the models calibrated on the Kling–Gupta efficiency alone have limited reliability in flood hazard assessments.

Such "negative" result could be expected, as there are several recent publications

Printer-friendly version

Discussion paper

pointing on a necessity of comprehensive approaches for hydrological model calibration and evaluation (for mean flow and for extremes) and especially if they are intended for climate impact assessment (see e.g. Choi and Beven, 2007, Coron et al., 2012, Refsgaard et al., 2013, Thirel et al., 2015, Krysanova et al., 2018). Therefore, such an "initial stage“ of the study should be supplemented by application of an extended approach: for example, including at least some of the further steps suggested in the papers listed above, like multi-site and multi-variable calibration (mentioned in the manuscript), DSS test checking for contrasting climate sub-periods, testing specifically for indicators of interest, i.e. for high flows and floods. Then the study would be much more valuable.

There are also other deficiencies in the applied approach and in the interpretation of the obtained results. Therefore, the manuscript should be rejected in its present form.

Other major concerns:

APPROACH

l. 81-82: were driven with Daymet meteorological forcing (Thornton et al., 2012) and mHM with the forcing by Maurer et al. (2002): → how are they comparable with the observed climate? Was the comparison done or not? If not, it would be reasonable to do.

l. 82: SAC, HBV, and VIC were evaluated on the period 1985–2008: → and calibrated for which period?

l. 110-112: → would be good to express the relative error in %, and define thresholds for acceptable performance (e.g. based on literature) for all 3 indicators. For example, is a relative error of 25% acceptable or not? The thresholds could be shown in Fig. 3 by horizontal lines to enable distinguishing the good/acceptable and poor performances.

Sec. 3.1: → to discuss performance based on the pre-defined thresholds

l. 116-118: a catchment is connected to another catchment if they share a certain

Printer-friendly version

Discussion paper

number of events, i.e. at least 1% of the total or seasonal number of events: → is 1% of shared events really sufficient to define their connectivity??? Due to that, the whole section 3.2 is questionable.

l. 127: we generate surrogate time series of temperature, precipitation, and streamflow for each catchment by resampling the available hydrological years with replacement: → the procedure is not quite clear, and should be better explained!

l. 202-203: "to assess each model' suitability for climate impact assessments on floods": → how the resampling could help to assess suitability? It would be better to test for contrasting climate subperiods, or to compare trends in discharge, high flows and POT series.

Sec. 3.3 and Fig.5: → Maybe to add correlation coefficients to better characterize the relationships?

INTERPRETATION

l. 145-146: "For most catchments, the number of flood events is relatively well simulated by most models": → this is not evident, if a threshold is not defined. It is only visible that medians are close to zero for three models, and there is no under- or overestimation for the whole set of 40 – 176 catchments, but nothing more! After defining the threshold, the interpretation could be different! Besides, it would make sense to normalize over the number of catchments in every regime? And it would be reasonable to cut Y scale for (i) at -50 and +50, even if one box for HBV will not be fully visible.

l. 158: Over all seasons, most models show an acceptable performance (i.e. median error close to zero): → if the median error is close to zero, it does not mean that most models show an acceptable performance!!! It only means that there is no tendency to over- or underestimation for catchments in five regimes, nothing more!

l. 156: Over all, there is no clear tendency of one model to perform better than the

other ones. → Based on thresholds, this could be better visible.

l. 224-225: reliance on an individual calibration metric (EKG) rather than a broader suite of performance metrics can lead to simulation performance deficits for phenomena of interest, including an underestimation of streamflow variability: → Not only the metric, but the calibration approach is general!!!

l. 238: the number of flood events in a simulation time series, which tend to verify well: → disagree, see above!

l. 245-246: Such focus could be improved by giving more weight to the variability component of EKG → or including indicators of extremes in the calibration/validation!!!

Minor corrections needed:

Fig. 1: catchments are indicated by the gauge location?

Fig. 2: for which period(s) is this statistics?

l. 112: circular statistics???

Fig. 3: to explain what is represented by each box with whiskers: comparison for all catchments in a regime over which period: 1981-2008? To add this to the caption.

l. 159-160: particularly in the Western part of the US: → not, in the middle part (intermittent regime)

———

I agree with the authors on the following:

l. 222-223: The results of this study indicate that the limited capability of hydrological models used in this study to reproduce observed hydrologic sensitivities during flooding may be related to insufficient model calibration: FULLY AGREE!

l. 247: The spatial concern could be addressed by applying spatial calibration procedures: → Agree!

l. 254-256: We conclude that calibration using only an individual model performance metric or variable can result in model implementations that have limited value for specific model applications, such as local and in particular spatial flood hazard analyses and change impact assessments: AGREE!

l. 258: more comprehensive multi-objective and multi-variable calibration strategies are needed: AGREE!

References Choi and Beven, 2007, doi:10.1016/j.jhydrol.2006.07.012 Coron et al., 2012, doi:10.1029/2011WR011721 Refsgaard et al., 2013, doi:10.1007/s10584-013-0990-2 Thirel et al., 2015, doi:10.1080/02626667.2015.1050027 Krysanova et al., 2018, DOI: 10.1080/02626667.2018.1446214

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-192, 2020.

Printer-friendly version

Discussion paper