24th International Congress on Modelling and Simulation, Sydney, NSW, Australia, 5 to 10 December 2021
mssanz.org.au/modsim2021

"Naïve" inclusion of diverse climates in calibration is not sufficient to improve model reliability under future climate uncertainty

L. Trotter [a], M. Saft [a], M.C. Peel [a] and K.J.A. Fowler [a]

[a] *Department of Infrastructure Engineering, University of Melbourne, Melbourne, Australia*
Email: l.trotter@unimelb.edu.au

Abstract: Parameter sets of hydrologic models do not transfer well between periods with different climatic conditions. Existing literature shows that model performance is particularly affected when parameters calibrated on wetter conditions are used to project streamflow during drier conditions. In the Australian context, where future projections indicate the climate is likely to become warmer and drier as a result of global climate changes, these limitations of hydrologic models become particularly disquieting, especially with regards to their implications for estimating water availability during dry periods. The Millennium drought, which impacted large parts of south-eastern Australia ca. 1997-2009, exposed these limitations of hydrologic models and their most common calibration methods. During the drought, many catchments in south-eastern Australia underwent changes in their hydrologic behaviour. Extensive research since the end of the drought shows that models calibrated on pre-drought conditions routinely overestimate streamflow when forced with climate data from the years of the drought.

In operational simulation, it is often assumed that once a model is shown a variety of climate conditions in the calibration sequence, it will perform better under future climate variability. In the context of the Millennium drought, it has been theorised that now that we have experienced these conditions, models calibrated on long timeseries that include the Millennium drought will be able to perform well under a future drier climate. In this study, we put this idea to the test. Specifically, we use five commonly used conceptual hydrologic models and evaluate their performance during and after the Millennium drought in 155 Victorian catchments. We test whether their performance (in terms of KGE and bias) improves significantly after inclusion in the calibration period of the drought and the post-drought periods themselves. For calibration we use an objective function specifically designed to optimise models' ability to reproduce both high and low flow conditions while minimising volumetric bias.

Our results show that the "naïve" approach of extending calibration sequences to include as much climate diversity as possible is not sufficient to significantly improve model reliability in the face of future climate uncertainty. We demonstrate that showing models data from the Millennium drought in calibration did not significantly improve their performance across this set of catchments, neither during the drought itself nor, in most cases, in the period after the drought. Further including the post-drought sequence in calibration does significantly improve post-MD KGE in three out of five models, but even in these models, performance is still significantly lower than it is when calibrating on post-drought only and the improvement, albeit statistically significant, is unlikely to make operational difference in most cases. Additionally, bias doesn't significantly change. This is despite drought and post-drought making up a significant proportion of the calibration sequence (at least 30%). Mann-Whitney tests were used to assess whether model performance was significantly different across the set of catchments. Our results also show that, while rarely significant, the extension of the calibration period does provide a marginal improvement in performance for almost all models and both periods tested. This is encouraging and supports the practice to expose models to a variety of climate conditions, however it indicates that additional provisions are needed when training models for use in ungauged climates.

Evidence from literature suggests that more sophisticated calibration methods with explicit and distinct treatment of different climate regimes improve model performance under a transient climate. However, especially in the catchments where drastic shifts were observed, new model structures that are more flexible to such climate-induced changes in hydrologic regime are likely necessary to confidently project streamflow under future climate scenarios. By exposing these limitations, we encourage members of the hydrologic community to exercise caution when applying our existing models and calibration frameworks to project streamflow into unknown and uncertain climate conditions. We also join the numerous community calls for new and more robust approaches for hydrologic modelling and simulation in the face of a changing climate.

Keywords: *Hydrologic models, calibration, climate variability, Millennium drought*

1. INTRODUCTION

Parameter sets of hydrologic models do not transfer well between periods with different climatic conditions. This is a well-known fact, supported by an extensive corpus of literature (e.g. Coron et al., 2012; Li et al., 2012; Seibert, 2003; Vaze et al., 2010). Transferability is particularly poor when the evaluation period is drier than the calibration period; in these conditions, models tend to overestimate streamflow during evaluation. These limitations of hydrologic models are especially disquieting in the context of climate change (Peel and Blöschl, 2011). Under a changing climate, hydrologic models play a central role in assessing risks derived from water availability and extreme hydroclimatic events (Xu, 1999). Nevertheless, future climate conditions (i.e. model forcing data) may be too uncertain or too different from any period in the observed series to be able to define and identify an adequate calibration period. In practice, to project streamflow under uncertain climates, it is standard procedure to subject models to differential split-sample testing – DSST (Klemeš, 1986) – to ensure their adequacy to perform under varying conditions. Models are then recalibrated on the entirety of the available data prior to their operational use to ensure their parameters have been tuned by exposing them to the maximum available climate variability.

In the Australian context, the Millennium drought (MD) exposed serious limitations of hydrologic models and their most common calibration methods (Fowler et al., 2016; Saft et al., 2016a). The MD impacted large parts of south-eastern Australia ca. 1997-2009 and contributed to a shift in the hydrologic behaviour of many catchments in the area, often persisting years after the end of the drought itself (Peterson et al., 2021; Potter and Chiew, 2011; Saft et al., 2016b). This, in turn, affected the reliability and adequacy of common frameworks for modelling and calibration (e.g. Fowler et al., 2020, 2016; Saft et al., 2016a). In particular, exposing hydrologic models to the many short dry spells which are present in the instrumental record prior to the MD is not sufficient for them to perform well during the MD itself (Saft et al., 2016a). Given the extreme conditions posed by this event and its unprecedented nature, however, it is often assumed for operational simulation that even if models fail this DSST, inclusion of the MD in the calibration sequence may be sufficient to train models to perform under future climate, as long as it isn't more extreme than what was observed during the MD (Chiew et al., 2014, 2009).

In this study, we investigate whether inclusion diverse climatic conditions such as the Millennium drought in the calibration period, without their explicit and distinct treatment, is sufficient to produce a set of parameters that will performs satisfactorily under future climate variability. We refer to this approach as "naïve" to distinguish it from more sophisticated calibration methods aiming at maximising the amount of information extracted from the calibration sequence through differential treatment of different periods or aspects of the flow regime – e.g. through meta-objective functions or multi-objective optimisation (Fowler et al., 2018, 2016). We use five commonly used conceptual hydrologic models and evaluate their performance during and after the Millennium drought in 155 Victorian catchments. We test whether their performance improves significantly after inclusion in the calibration period of the drought and the post-drought periods themselves.

2. METHODS

2.1. Catchments and data

Models are calibrated and forced with data from 155 drought-affected catchments in the southern Australian state of Victoria. Located on both sides of the Great Dividing Range, these catchments experienced a range of hydrologic responses to the Millennium drought: two-thirds of them were shown to shift their rainfall-runoff relationship during the drought and only half of the shifted catchments recovered from these shifts in the decade after the end of the dry spell (Peterson et al., 2021). Meteorological data were extracted from daily gridded interpolated records: rainfall data comes from the Australian Gridded Climate Data collection (Jones et al., 2009), while for temperatures (HBV only) and potential evapotranspiration (Morton's wet environment) the SILO database was used (Jeffrey et al., 2001). Catchment-level daily average values were calculated from the gridded records. All meteorological data is complete at the daily timestep from before 1950 to 2019 inclusive.

Streamflow data were obtained from the WMIS portal of the Victorian Department of Environment, Land, Water and Planning. All available daily data between 1950 and 2019 were extracted and used for this study. Within the 155 catchments, 29 had streamflow data starting on or prior to 1950; for the majority of gauges, monitoring began in the 1950's or 1960's (110 gauges in total). The shortest record starts in 1981. All except 15 of the catchments have streamflow records running up to the end of the 2019 water year (i.e. 29 February 2020). This set of catchments was selected to exclude catchments with incomplete or unreliable records and significant external impacts on their flow regime; additionally, quality codes were used to additionally filter out problematic data points.

We distinguish three periods of interest: (1) PreMD is the period before the onset of the drought, i.e. up to and including 1996; (2) MD is the period during which the drought occurred, here fixed between 1997 and 2009 water years inclusive for all catchments; and (3) PostMD refers to the years after the drought, from 2010 to the end of the record. Based on available streamflow, the average (median) length of the PreMD period in these catchments is 33.5 (32) years; whereas the MD is of a fixed 13 years for all catchments. The minimum length of the PostMD period is 5 years, but as already stated, all except for 15 catchments have 10 years of record in this period. When models are calibrated on the period including PreMD and MD (see section 2.2 below, calibration 2), the drought represents on average 29.2% of the calibration length. When the PostMD period is added to that (hence calibrating on the entire record, calibration 3 in section 2.2) the period after the drought makes up on average 17.9% of the calibration sequence and MD and PostMD together sum up to on average 41.7% and at least 29.9% of it.

2.2. Experimental set-up

We use five commonly used, conceptual, spatially-lumped hydrologic models, namely IHACRES (Croke and Jakeman, 2004; Jakeman et al., 1990), GR4J (Perrin et al., 2003), SimHyd (Chiew et al., 2002), Sacramento (Burnash, 1995) and HBV (Lindström et al., 1997). This set of models encompasses a range of complexities (1–5 stores and 4–15 parameters) and are all widespreadly used in hydrologic modelling studies both in Australia and overseas, including in the same area and period used for this study (e.g. Fowler et al., 2020, 2016; Saft et al., 2016a). All models are used in their implementation within the MARRMoT framework (Knoben et al., 2019; Trotter et al., in preparation).

Models are calibrated using the Covariance Matrix Adaptation Evolution Strategy – CMA-ES (Hansen et al., 2003). The objective function optimized in the calibration process is designed to capture the ability of models to reproduce high and low flow conditions while maintaining minimal volumetric bias. The objective function is comprised of the average of two KGEs (Gupta et al., 2009), one calculated on observed and simulated streamflow and one on their transformation using the fifth-root to enhance the weight given to smaller flows, and of a logarithmic bias penalisation factor (Viney et al., 2009), which reduces the value of the objective function as the volumetric bias (B) deviates from zero. The formula to calculate model efficiency (E) under this objective function is given in equation 1 below.

$$E = \frac{1}{2}\left(KGE_Q + KGE_{Q^{0.2}}\right) - 5 \cdot |\ln(B+1)|^{2.5} \quad (1)$$

Meteorological data of the five years prior to the beginning of the calibration sequence are used to warm-up the models and stabilise the stores.

Each of the five models was calibrated in each of the 155 catchments using data from five different periods:

1. Pre-drought period (PreMD), from the beginning of the streamflow record up to 1996;
2. Pre-drought and drought (PreMD+MD), from the beginning of the streamflow record up to 2009;
3. Entire record (AllTime), using all of the available streamflow record for each catchment;
4. Millennium drought only (MD), from 1997 to 2009;
5. Post-drought only (PostMD), from 2010 and up to the end of the streamflow record.

Once a calibration run returned a set of parameters, this was used to simulate streamflow for the entire available record and the streamflow output by this simulation is used to evaluate model performance during and after the drought. KGE and volumetric bias are used as metrics for model performance, representing respectively how well a model reproduces the hydrograph overall and the water balance specifically.

For each calibration period, results from individual catchments are combined into a distribution reflecting performance across all catchments. Mann-Whitney tests (Mann and Whitney, 1947) are computed to compare the distributions from the different calibrations. We use $\alpha = 0.05$ as threshold for significance and evaluate global significance using the false discovery rate (FDR) method (Benjamini and Hochberg, 1995). We compare the performance during the drought of models calibrated on periods 1, 2 and 4 above; where comparison between the performance using parameter sets from calibrations 1 and 2 indicate whether the performance improves significantly by adding the MD in the calibration sequence, while calibration 4 is used as a benchmark indicating the maximum expected performance of a model during the drought. Similarly, for the period after the end of the drought, we compare performance with parameter sets from calibrations 1, 2, 3 and 5: here calibration 5 is the benchmark and the comparison of calibrations 1, 2 and 3 are useful to identify whether performance improves significantly by adding the drought or the post-drought to the calibrations sequence.

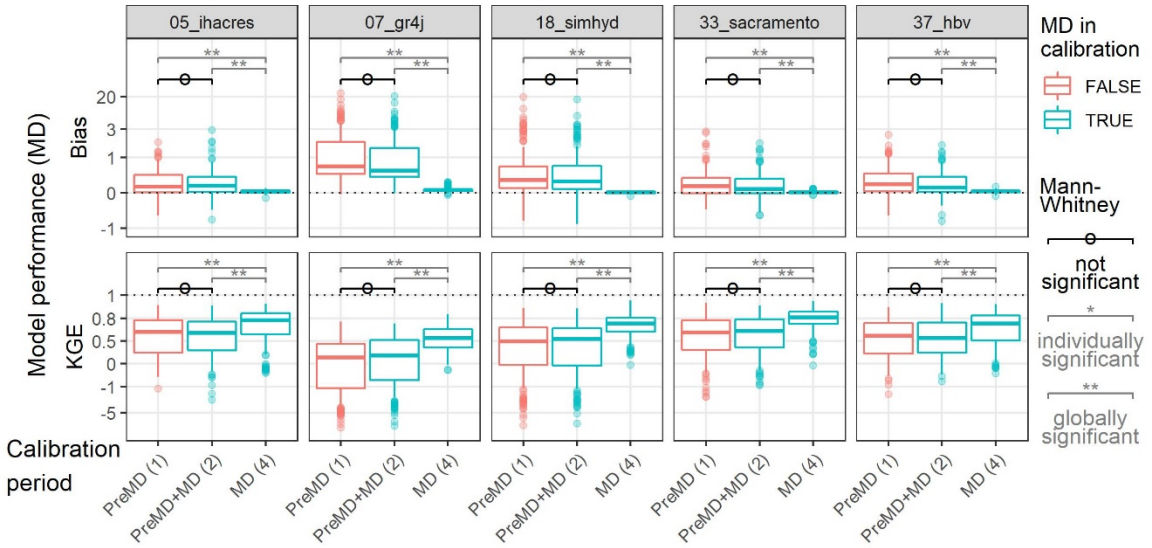

Figure 1. Model performance during the Millennium drought (MD), depending on the period used for calibration. Significance of Mann-Whitney test indicates significant difference in performance.

3. RESULTS

Including the Millennium drought in the calibration sequence does not significantly improve the ability of models to project streamflow during the Millennium drought itself. Model performance during the drought for calibration periods 1 (PreMD), 2 (PreMD and MD) and 4 (MD) is shown in Figure 1. When models are calibrated on the PreMD period, their median KGE in evaluation during the MD across this set of catchments ranges from 0.17 (GR4J) to 0.63 (IHACRES). Model performance increases an average of 0.012 points on the KGE scale when the Millennium drought itself is shown to the models in calibration; the model with the biggest improvement in performance is GR4J, where the median KGE goes up by 0.048 points. Based on the results of the Mann-Whitney tests, these improvements in performance are never significant (Figure 1). The addition of the Millennium drought to the calibration sequence contributes on average to closing 1.9% of the gap in performance from the benchmark (i.e. calibration 4), with SimHyd closing the biggest gap (14.5% of the gap filled). In terms of bias, the same considerations apply: models overestimate streamflow volumes during the drought, regardless of whether the parameter sets used were optimised on the pre-drought period only or also including the drought itself. Again, we see a marginal improvement in the estimation of flow volumes– on average, median bias decreases by 5.3%, with GR4J again leading the other models with a 14.0% median bias reduction – but this again does not add up to a significant improvement according to the Mann-Whitney tests.

Using parameter sets calibrated including the Millennium drought also doesn't improve model performance after the drought. Figure 2 shows model performance in the PostMD period, here we compare performance using parameter sets from calibrations 1, 2, 3 (entire period) and 5 (PostMD only). Changes in post-drought median KGE between calibrations 1 and 2 (i.e. by adding the MD in the calibration sequence) are between 0.09 points (GR4J) and −0.07 points (HBV), averaging 0.01 across all models. None of these changes are significant according to the Mann-Whitney tests. This very marginal improvement results in 3.2% of the performance gap during the post-drought period being filled by adding the drought to the calibration period – this value jumps to 18.2% if only the models where the PostMD median performance improved in calibration 2 compared to calibration 1 are considered, i.e. excluding HBV and Sacramento. Again, we see similar results for the bias. The only model where adding the MD to the calibration sequence significantly improved bias estimation after the drought is GR4J, whose PostMD median bias reduced by 9.4% by using parameters from calibration 2 instead of calibration 1, closing 28.8% of the bias gap and bringing its post-MD bias in line to that of the other models. For all other models, however, the reduction in bias is not significant and is on average 3.2%, ranging between 0.8% (Sacramento) to 5.8% (SimHyd).

Showing models the data after the drought during calibration resulted in some marginal PostMD performance improvement for the KGE, but never for the bias. Comparing the second and third boxes of the plots in Figure 2 reveals that, adding the PostMD period itself to calibration did significantly improve PostMD KGE scores for SimHyd, Sacramento and HBV. Median KGE for these three models improved on average by 0.09 points from calibration 2 to calibration 3: albeit statistically significant, this is unlikely to make operational difference in most cases. Adding the PostMD to the calibration sequence of SimHyd and Sacramento closed 32.4% and

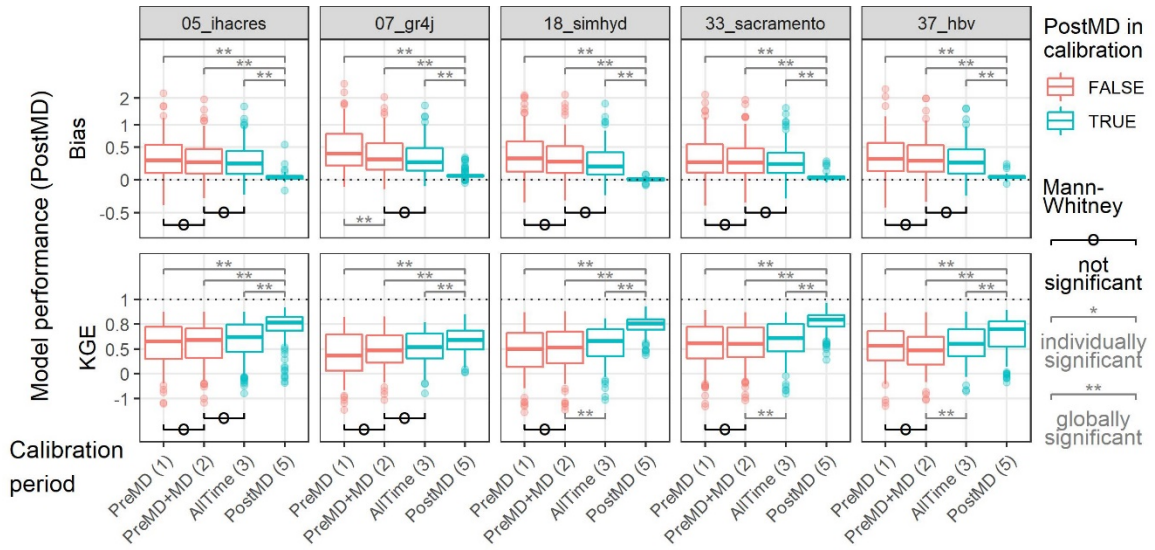

Figure 2. Model performance after the Millennium drought (PostMD), depending on the period used for calibration. Significance of Mann-Whitney test indicates significant difference in performance.

28.4% of the KGE gap to calibration 5, respectively – this is less than the gap filled in GR4J's KGE (31.7%) where the improvement, however, is not significant. For HBV, the gap filled is larger (34.4%). Significance of the improvement in KGE value of HBV, however, is largely due to the fact that the model with parameters from calibration 2 saw a decrease in model performance after the drought compared to calibration 1 parameters. In fact, the difference in PostMD median KGE between calibrations 1 and 3 is again not significant (not shown). With regards to bias, again the inclusion the PostMD period in calibration provides only a marginal improvement to PostMD median bias for all models – on average bias is reduced by 3.7% and up to 6.7% (SimHyd) – however, this never results in a significant change from the bias resulting from simulations with parameter sets from calibration 2.

4. DISCUSSION AND CONCLUSIONS

The results presented clearly indicate that the "naïve" approach of extending calibration sequences to include as much climate diversity as possible is not sufficient to significantly improve model reliability in the face of future climate uncertainty. This is particularly striking given that, in the methodology used for this study, models were given the advantage of being evaluated on data they had already seen during the calibration process. Nevertheless, our results also show that, albeit rarely significant, the extension of the calibration period does provide a marginal improvement in performance for almost all models and both periods tested. This indicates that while it is good practice to expose models to a variety of climate conditions, further provisions are necessary when training models for use in ungauged climates. This may be caused by a faulty design of the calibration framework, which fails to give enough emphasis on the dry periods in the calibration sequence, or may be the results of inadequate model structures, which, no matter how they are trained, result in models unable to perform satisfactorily in all three periods: before, during and after the drought.

Evidence suggests that different calibration methods which treat dry and non-dry periods distinctly can be helpful to improve model performance under changing conditions. Fowler et al. (2016) demonstrated that in some cases models can perform well during both dry and non-dry climate when these two conditions are given equal weight in the calibration process – in their case, using a multi-criteria optimiser. Another possible approach that could be implemented without the use of multi-objective optimisers is to average values of an objective function calculated on specific subperiods, similarly to the split-KGE approach which proved superior to others according to Fowler et al. (2018), this explicitly prevents the generation of a parameter set biased towards the wettest years or periods of the calibration sequence.

Despite our best efforts, however, it is possible and likely that new model structures that are more flexible to changes in hydrologic conditions caused by climate variability are necessary for models to be able to satisfactorily extrapolate into future climate scenarios. Many of the catchments in this study underwent changes in their hydrologic behaviour induced or exacerbated by the prolonged dry period during the MD, which persisted after the end of the dry spell (Peterson et al., 2021). In the catchments where drastic behavioural changes occurred, improvements in calibration may not be sufficient to train models to simulate both regimes

and new model structures which include time variant parameters or store sizes, or otherwise more flexible components are likely to be needed to appropriately reproduce such changes in behaviour. This is likely to become more common as climatic changes exacerbate in the future, however, the lack of a comprehensive understanding of the causes of the observed shifts entails this approach is challenging both in terms of model design and eventual calibration of such modified models. In this context, fully distributed, process-based models might be better at modelling future conditions thanks to more physically based components that better reproduce catchment changes. Nevertheless, uncertainty due to overparametrisation and equifinality are grave limits of these more complex models and their calibration methods (Franks et al., 1998).

With this short study, we demonstrated that long time series which include extensive climate variability are not sufficient to calibrate models able to perform under future climate conditions. We used climate and streamflow data from catchments affected by the Millennium drought in south-eastern Australia and compared the performance of models in these catchments during and after the drought. Model performance during these periods of interest was almost never significantly improved even when the drought or the post-drought period themselves made up a significant proportion (at least 30%) of the calibration sequence. We argue that distinct treatment of different climatic conditions observed in the calibration period is more appropriate to properly inform models and train them to perform in a variety of future climates. Nevertheless, we recognise that it is also probable that, given the observed hydrologic non-stationarity that developed during the drought in these catchments, existing models may not altogether have the appropriate flexibility and resilience to perform in such drastically changing conditions. We hope with this piece of research to increase awareness of the limits our existing frameworks for modelling and calibration have when projected into unknown, uncertain and previously unseen conditions. While we encourage modellers, researchers and hydrologists to exercise humility and constraint in this regard, we also hope to inspire them by joining the numerous calls to spur action towards the development of novel and innovative approaches to hydrologic modelling and simulation under uncertainty.

ACKNOWLEDGEMENTS

This study received support from the Australian Research Council via Linkage Project LP180100796 *Observed streamflow generation changes: better understanding and modelling*, which was co-funded by the Victorian Department of Environment, Land, Water and Planning and Melbourne Water. KF acknowledges support from LP170100598.

REFERENCES

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B 57, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Burnash, R.J.C., 1995. The NWS River Forecast System-catchment modeling. Comput. Model. watershed Hydrol. 311–366.

Chiew, F.H.S., Peel, M.C., Western, A.W., 2002. Application and testing of the simple rainfall runoff model Simhyd. Math. Model. Small Watershed Hydrol. Appl.

Chiew, F.H.S., Potter, N.J., Vaze, J., Petheram, C., Zhang, L., Teng, J., Post, D.A., 2014. Observed hydrologic non-stationarity in far south-eastern Australia: Implications for modelling and prediction. Stoch. Environ. Res. Risk Assess. 28, 3–15. https://doi.org/10.1007/s00477-013-0755-5

Chiew, F.H.S., Teng, J., Vaze, J., Post, D.A., Perraud, J.M., Kirono, D.G.C., Viney, N.R., 2009. Estimating climate change impact on runoff across southeast Australia: Method, results, and implications of the modeling method. Water Resour. Res. 45, 1–17. https://doi.org/10.1029/2008WR007338

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 Resour. Res. 48. https://doi.org/10.1029/2011WR011721

Croke, B.F.W., Jakeman, A.J., 2004. A catchment moisture deficit module for the IHACRES rainfall-runoff model. Environ. Model. Softw. 19, 1–5. https://doi.org/10.1016/j.envsoft.2003.09.001

Fowler, K.J.A., Knoben, W.J.M., Peel, M.C., Peterson, T., Ryu, D., Saft, M., Seo, K., Western, A.W., 2020. Many commonly used rainfall-runoff models lack long, slow dynamics: implications for runoff projections. Water Resour. Res. 56. https://doi.org/10.1029/2019wr025286

Fowler, K.J.A., Peel, M.C., Western, A., Zhang, L., 2018. Improved Rainfall-Runoff Calibration for Drying Climate: Choice of Objective Function. Water Resour. Res. 54, 3392–3408. https://doi.org/10.1029/2017WR022466

Fowler, K.J.A., Peel, M.C., Western, A.W., Zhang, L., Peterson, T.J., 2016. Simulating runoff under changing climatic conditions: Revisiting an apparent deficiency of conceptual rainfall-runoff models. Water Resour. Res. 52, 1820–1846. https://doi.org/10.1002/2015WR018068

Franks, S.W., Gineste, P., Beven, K.J., Merot, P., 1998. On constraining the predictions of a distributed model: The incorporation of fuzzy estimates of saturated areas into the calibration process. Water Resour. Res. 34, 787–797. https://doi.org/10.1029/97WR03041

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. J. Hydrol. 377, 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Hansen, N., Müller, S.D., Koumoutsakos, P., 2003. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). Evol. Comput. https://doi.org/10.1162/106365603321828970

Jakeman, A.J., Littlewood, I.G., Whitehead, P.G., 1990. Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments. J. Hydrol. https://doi.org/10.1016/0022-1694(90)90097-H

Jeffrey, S.J., Carter, J.O., Moodie, K.B., Beswick, A.R., 2001. Using spatial interpolation to construct a comprehensive archive of Australian climate data. Environ. Model. Softw. 16, 309–330. https://doi.org/10.1016/S1364-8152(01)00008-1

Jones, D.A., Wang, W., Fawcett, R., 2009. High-quality spatial climate data-sets for Australia. Aust. Meteorol. Oceanogr. J. 58, 233–248. https://doi.org/10.22499/2.5804.003

Klemeš, V., 1986. Operational testing of hydrological simulation models. Hydrol. Sci. J. 31, 13–24. https://doi.org/10.1080/02626668609491024

Knoben, W.J.M., Freer, J.E., Fowler, K.J.A., Peel, M.C., Woods, R.A., 2019. Modular Assessment of Rainfall-Runoff Models Toolbox (MARRMoT) v1.2: An open-source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous state-space formulations. Geosci. Model Dev. 12, 2463–2480. https://doi.org/10.5194/gmd-12-2463-2019

Li, C.Z., Zhang, L., Wang, H., Zhang, Y.Q., Yu, F.L., 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

Lindström, G., Johansson, B., Persson, M., Gardelin, M., Bergström, S., 1997. Development and test of the distributed HBV-96 hydrological model. J. Hydrol. https://doi.org/10.1016/S0022-1694(97)00041-3

Mann, H.B., Whitney, D.R., 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Ann. Math. Stat. 18, 50–60. https://doi.org/10.1214/aoms/1177730491

Peel, M.C., Blöschl, G., 2011. Hydrological modelling in a changing world. Prog. Phys. Geogr. 35, 249–261. https://doi.org/10.1177/0309133311402550

Perrin, C., Michel, C., Andréassian, V., 2003. Improvement of a parsimonious model for streamflow simulation. J. Hydrol. https://doi.org/10.1016/S0022-1694(03)00225-7

Peterson, T.J., Saft, M., Peel, M.C., John, A., 2021. Watersheds may not recover from drought. Science 372, 745–749. https://doi.org/10.1126/science.abd5085

Potter, N.J., Chiew, F.H.S., 2011. An investigation into changes in climate characteristics causing the recent very low runoff in the southern Murray-Darling Basin using rainfall-runoff models. Water Resour. Res. 47. https://doi.org/10.1029/2010WR010333

Saft, M., Peel, M.C., Western, A.W., Perraud, J.M., Zhang, L., 2016a. Bias in streamflow projections due to climate-induced shifts in catchment response. Geophys. Res. Lett. 43, 1574–1581. https://doi.org/10.1002/2015GL067326

Saft, M., Peel, M.C., Western, A.W., Zhang, L., 2016b. Predicting shifts in rainfall-runoff partitioning during multiyear drought: Roles of dry period and catchment characteristics. Water Resour. Res. 52, 9290–9305. https://doi.org/10.1002/2016WR019525

Seibert, J., 2003. Reliability of Model Predictions Outside Calibration Conditions. Nord. Hydrol. 34, 1–13. https://doi.org/10.2166/nh.2003.0019

Trotter, L., Knoben, W.J.M., Fowler, K.J.A., Saft, M., Peel, M.C., n.d. Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) v2.0: an object-oriented implementation of all your favourite hydrologic models for improved speed and readability., In preparation.

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. J. Hydrol. 394, 447–457. https://doi.org/10.1016/j.jhydrol.2010.09.018

Viney, N.R., Perraud, J., Vaze, J., Chiew, F.H.S., Post, D.A., Yang, A., 2009. The usefulness of bias constraints in model calibration for regionalisation to ungauged catchments, in: Proceedings of the 18 Th World IMACS / MODSIM Congress. Cairns, Australia.

Xu, C.Y., 1999. Climate change and hydrologic models: A review of existing gaps and recent research developments. Water Resour. Manag. 13, 369–382. https://doi.org/10.1023/A:1008190900459