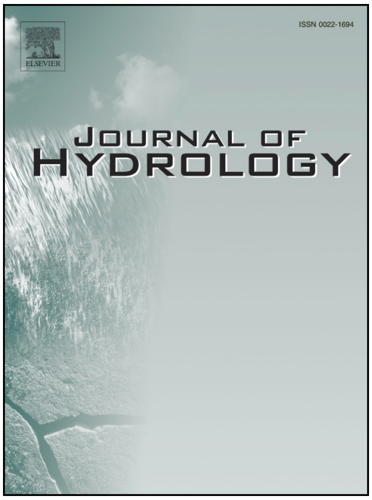
Journal Pre-proofs



Research papers

Investigating strategies to improve hydrologic model performance in a changing climate

C.M. Stephens, L.A. Marshall, F.M. Johnson

PII: S0022-1694(19)30954-0
DOI: https://doi.org/10.1016/j.jhydrol.2019.124219
Reference: HYDROL 124219

To appear in: *Journal of Hydrology*

Received Date: 29 July 2019
Revised Date: 27 September 2019
Accepted Date: 7 October 2019

Please cite this article as: Stephens, C.M., Marshall, L.A., Johnson, F.M., Investigating strategies to improve hydrologic model performance in a changing climate, *Journal of Hydrology* (2019), doi: https://doi.org/10.1016/j.jhydrol.2019.124219

Investigating strategies to improve hydrologic model performance in a changing climate

C.M. Stephens*, L.A. Marshall, F.M. Johnson

Water Research Centre, School of Civil and Environmental Engineering, UNSW Sydney, NSW 2052, Australia

*Corresponding author.

Highlights

- Catchment model experiments assessed performance shifts under climatic change.
- We tested parameter transferability, model weighting and dynamic parameters.
- Transferability depended more on validation than calibration conditions.
- Approaches tested improved average validation performance, but results varied.

Abstract

It has been repeatedly shown that conceptual hydrologic model performance degrades under conditions that deviate from those of the calibration period. In this study, we describe three experiments that aim to understand and address this problem using the conceptual model GR4J over 164 Australian catchments. The first is an investigation of model transferability, where parameters calibrated under certain conditions are applied to simulate both similar and contrasting conditions. We find that model performance degradation is more dependent on the conditions under which the model is tested than the conditions under which it is calibrated. Because dry periods are typically more difficult to simulate than wet periods, this means that transferring dry-calibrated parameters to wet periods is more successful than transferring wet-calibrated parameters to dry periods. For both wet and dry periods the best results were obtained when climatically similar calibration periods were used, suggesting that targeted use of climatically similar calibration data could improve predictive capacity. To this end, a second experiment was designed that preferentially weights modeled series calibrated under different conditions, with series associated with more climatically similar calibration periods weighted more heavily. While this improved model performance in most cases, the success was variable across the different catchments. Given that the model weighting scheme could not easily be generalized to all catchments in the sample, a third experiment was conducted where each model parameter was defined dynamically as a function of climate. The dynamic parameters were calibrated separately for each model, so the individual sensitivities of each catchment to climate conditions could be captured. While this also gave performance improvements, especially under drier testing conditions, the results continued to vary between catchments and there was no clear pattern in the parameter variation. This suggests that nonstationarity can be captured in different parameters for models of different catchments. While both model weighting and dynamic parameters can benefit overall conceptual model performance, it seems that reliable improvements across large samples of catchments may be difficult to achieve without more physically realistic model structures.

Keywords

Hydrologic modeling

model robustness

model weighting

dynamic parameters

climate variability

1 Introduction

Hydrologic models are used for a wide range of purposes including flood risk management, water resource planning and environmental flow assessments. They typically require calibration to observed streamflow to determine parameter values that reflect the runoff behavior of the catchment. This means the model effectively projects past behavior forward, assuming stationarity over the long term. However, there is now overwhelming evidence that the Earth's climate is undergoing substantial change due to anthropogenic activities [IPCC, 2013]. It follows that changes in catchment characteristics and hence response may also occur, threatening the assumption of stationarity and the validity of models calibrated to past streamflow observations [Milly et al., 2008]. Uncertainty in future climate drivers and associated catchment property shifts presents a key challenge for developing improved modelling methods.Fig 1.fig1Fig 2.fig2Fig 3.fig3Fig 4.fig4Fig 5.fig5Fig 6.fig6Table 1.tbl1Table 2.tbl2Table 3.tbl3Table 4.tbl4Table 5.tbl5Table 6.tbl6Table 7.tbl7Table 8.tbl8Table 9.tbl9

A number of studies have shown that the performance of conceptual hydrologic models tends to degrade when climatic conditions become increasingly different to those of the calibration period [Brigode et al., 2013; Coron et al., 2014; Coron et al., 2012; Merz et al., 2011; Osuch et al., 2015; Vaze et al., 2010]. *Vaze et al.* [2010] defined eight calibration periods based on average rainfall (wet versus dry), then evaluated how calibrated parameters transferred to time periods with different climatic conditions. They found that model skill decreased as the contrast between the calibration and testing periods increased, with the effect being more pronounced for drier testing periods. This result was supported by several subsequent studies [Coron et al., 2014; Coron et al., 2012; Dakhlaoui et al., 2017]. *Merz et al.* [2011] calibrated models of 273 Austrian catchments over six consecutive five-year periods and found significant trends in parameters related to snow and soil moisture. These trends were associated with changes in climate observed in the study area, indicating that the parameters (being conceptual without a precise physical interpretation) do relate to environmental conditions in the

Commented [A1]: AUTHOR: Please note that Fig 1. was not cited in the text. Please check that the citation suggested by the copyeditor is in the appropriate place, and correct if necessary.

Commented [A2]: AUTHOR: Please note that Fig 2. was not cited in the text. Please check that the citation suggested by the copyeditor is in the appropriate place, and correct if necessary.

Commented [A3]: AUTHOR: Please note that Fig 3. was not cited in the text. Please check that the citation suggested by the copyeditor is in the appropriate place, and correct if necessary.

Commented [A4]: AUTHOR: Please note that Fig 4. was not cited in the text. Please check that the citation suggested by the copyeditor is in the appropriate place, and correct if necessary.

Commented [A5]: AUTHOR: Please note that Fig 5. was not cited in the text. Please check that the citation suggested by the copyeditor is in the appropriate place, and correct if necessary.

Commented [A6]: AUTHOR: Please note that Fig 6. was not cited in the text. Please check that the citation suggested by the copyeditor is in the appropriate place, and correct if necessary.

Commented [A7]: AUTHOR: Please note that Table 1. was not cited in the text. Please check that the citation suggested by the copyeditor is in the appropriate place, and correct if necessary.

Commented [A8]: AUTHOR: Please note that Table 2. was not cited in the text. Please check that the citation suggested by the copyeditor is in the appropriate place, and correct if necessary.

Commented [A9]: AUTHOR: Please note that Table 3. was not cited in the text. Please check that the citation suggested by the copyeditor is in the appropriate place, and correct if necessary.

Commented [A10]: AUTHOR: Please note that Table 4. was not cited in the text. Please check that the citation suggested by the copyeditor is in the appropriate place, and correct if necessary.

Commented [A11]: AUTHOR: Please note that Table 5. was not cited in the text. Please check that the citation suggested by the copyeditor is in the appropriate place, and correct if necessary.

Commented [A12]: AUTHOR: Please note that Table 6. was not cited in the text. Please check that the citation suggested by the copyeditor is in the appropriate place, and correct if necessary.

Commented [A13]: AUTHOR: Please note that Table 7. was not cited in the text. Please check that the citation suggested by the copyeditor is in the appropriate place, and correct if necessary.

Commented [A14]: AUTHOR: Please note that Table 8. was not cited in the text. Please check that the citation suggested by the copyeditor is in the appropriate place, and correct if necessary.

Commented [A15]: AUTHOR: Please note that Table 9. was not cited in the text. Please check that the citation suggested by the copyeditor is in the appropriate place, and correct if necessary.

catchment. These findings were supported by further work in the same region by Sleziak et al. [2018].

Other studies have addressed non-stationarity in climatic conditions by investigating the utility of non-stationary parameterizations. Westra et al. [2014] evaluated time-varying parameters in the common conceptual model GR4J [Perrin et al., 2003] that were nominally representative of antecedent conditions, climatic cycles and long-term trends. The model's predictive capacity improved when the parameter representing storage capacity was allowed to vary over time, with particular improvement under dry conditions. This suggests that there may be potential to enhance future predictions by including time-varying parameters in hydrologic models. Later studies have used data assimilation to automatically update parameters based on climate variability [Xiong et al., 2019] or land cover change [Pathiraja et al., 2016]. *Grigg and Hughes* [2018] tested two model structural changes in GR4J aiming to reproduce the effects of catchment memory and vegetation cover shifts. The updated model was better able to capture hydrologic dynamics associated with groundwater declines and clearing followed by revegetation [Grigg and Hughes, 2018].

Another recommendation that has been put forward to improve model performance under nonstationary conditions is improving calibration methods to identify more widely-applicable parameters [Fowler et al., 2016]. We put forward an additional idea: using the available observations such that climatically similar calibration data is preferenced to simulate a projected future climate state with particular hydrologic statistics. Given that optimal model calibration parameters have been shown to correlate with climatic conditions [Merz et al., 2011], we hypothesize that targeted use of the available data for model calibration could improve performance under climatic conditions outside the existing record. To this end, we use data from 164 Australian catchments to conduct an exploratory analysis of a model weighting strategy that preferences data from climatically similar calibration subsets. We also test an updated model structure where parameters are defined dynamically based on climate conditions at a given time. The ultimate goal of this study was to test whether model robustness in a future climate could be targeted through improved calibration strategies with available current climate data.

2 Data and Methods

The initial steps for this study involved identifying catchments and obtaining data; selecting a hydrologic model; calculating a statistic to describe climate conditions, and calibrating the model over climatically distinct subsets of the observations. The Australian Bureau of Meteorology (BoM) provides high quality streamflow data at 222 Hydrologic Reference Stations (HRS) around Australia. We identified catchments with data available over 40 water years commencing in 1974 and whose catchment areas were between 25 km^2 and 10,000 km^2, totaling 164 catchments (Figure 1). Note that the definition of the water year (by month) is provided by the BoM for each HRS, so the data period begins in different months for catchments in different regions. Rainfall and temperature data were obtained from the Australian Water Availability Project daily grids at 0.05 degree resolution [Raupach et al., 2009; Raupach et al., 2012]. Potential evaporation (PET) was estimated using the McGuinness Bordne method [McGuinness and Bordne, 1972] in the R package *Evapotranspiration* [Guo et al., 2016], which requires only temperature data as an input and has been identified as a suitable PET formulation for conceptual hydrologic modeling [Oudin et al., 2005].

The streamflow, rainfall and PET data were used to calibrate the conceptual hydrologic model GR4J [Perrin et al., 2003] for each catchment. GR4J is parsimonious, appropriate for a range of climate conditions [Anshuman et al., 2018; Perrin et al., 2003] and commonly used in Australia [Grigg and Hughes, 2018; Guo et al., 2017; Humphrey et al., 2016; Stephens et al., 2018; Zhou et al., 2015]. It contains a production store (or soil moisture accounting store, controlled by parameter X_1), a non-linear routing store (controlled by parameter X_3) and an inter-catchment water exchange allowance (controlled by parameter X_2). The fourth parameter (X_4) controls the time bases of two unit hydrographs. Of the water that either bypasses or percolates through the production store, 90% is routed through the first unit hydrograph (with time base X_4) and into the routing store, where it contributes to the routed flow component. The remaining 10% is routed through the second unit hydrograph (with time base $2X_4$) and contributes to the direct flow component. Further detail on the model components and equations is provided by *Perrin et al.* [2003].

The experiments conducted in this study involved calibrating models over climatically distinctive subsets of the available observations. In order to specify calibration subsets, it was necessary to select an appropriate climate indicator. The Reconnaissance Drought Index (RDI) [Tsakiris and Vangelis, 2005] was chosen because it considers both precipitation and PET, and it has been shown that this will become increasingly important for classifying climate states in a warming world [Zarch et al., 2015]. Calculated on a monthly basis, the RDI is based on the ratio between precipitation and potential evaporation summed for the months of the water year so far (i.e. the calculation resets at the beginning of a new water year). It can be calculated over past periods using observations, or over future climate series by inputting data produced through perturbation methods or downscaling of General Circulation Model results [Maraun et al., 2010]. The RDI is defined as:

$$RDI^{i,k} = \frac{\sum_{j=1}^{j=k} P_{ij}}{\sum_{j=1}^{j=k} PET_{ij}} \quad (1)$$

Where $RDI^{i,k}$ is the RDI for the kth month of the ith water year. P_{ij} and PET_{ij} are the total monthly rainfall and potential evaporation for the jth month of the ith year. Normalized RDI (RDI_n) is usually calculated with reference to the average conditions for each month in order to identify drought conditions [Tsakiris and Vangelis, 2005], but for this study it was necessary to have a consistent measure of catchment wetness across time. It would not be suitable to have different RDI_n values for equivalent catchment wetness depending on time of year. Therefore, RDI was normalized as follows:

$$RDI_n^{i,k} = \frac{RDI^{i,k}}{mean(RDI)} - 1 \quad (2)$$

In Equation (2), each value is normalized based on the overall mean RDI for the catchment rather than the mean RDI calculated for the kth month, which is standard. The value of RDI_n increases with climatic wetness, with negative values indicating dry conditions and positive values indicating wet conditions.

For each catchment, subsets for model calibration were selected from the first 35 water years of observations (note the last five years were kept aside for use as unseen testing data). A monthly RDI_n series was calculated and the periods with the lowest and highest mean RDI$_n$, starting at the beginning of any month, were selected as the driest (dry1) and wettest (wet1) subsets. The second and third driest / wettest subsets were then identified as well (dry2, wet2, dry3 and wet3) ensuring there was no overlap (Figure 2a). This was repeated for subset periods of different lengths (discussed in detail in Section 0) to explore the trade-off between subset length and climate specificity. Shorter subsets allow for more climatic variation between periods (i.e. the driest period over one year will be more extreme than the driest period over two years). For longer subsets there may not be sufficient distinction between different climatic states. However, a sufficient amount of information is necessary to properly identify hydrologic parameters and ensure the catchment response is modeled realistically [Perrin et al., 2007; Sorooshian et al., 1983]. For this experiment, period lengths from 1 year to five years were tested.

Commented [A16]: This citaion is not found in Section Title.Please check if it is citation or not and proceed further

A limitation to this approach is that the conditions of preceding water years are not considered when defining subsets. In some catchments, soil moisture and groundwater stores can impact runoff behavior over multi-year periods [Saft et al., 2015]. However, given the large number and wide variety of catchments considered in this study, it was not feasible to define a universal period over which long-term antecedent conditions should be taken into account. Therefore, we do not account for conditions prior to the water year in which each subset begins, and this is a source of uncertainty for the methods we examine.

Note that short gaps left between defined periods may not be long enough to define subsequent periods without overlapping. Gaps shorter than the period length are therefore excluded from the process. This means that the length of data required to define six independent subsets can be up to twice the minimum necessary length (i.e. to define six two-year periods, at least 12 and at most 24 years of data will be required). If the maximum length of data necessary is not available, the timing of wet and dry periods will determine whether six independent subsets can be defined. In cases where six periods couldn't be defined without overlap, the catchment was excluded from the analysis for that length of subset (see Section 0). For this reason, two catchments were excluded from the four-year subset experiment and 98 catchments were excluded from the five-year analysis.

Commented [A17]: This citaion is not found in Section Title.Please check if it is citation or not and proceed further

For each catchment, the six climatically distinct subsets were used to calibrate six sets of GR4J parameters using the Shuffled Complex Evolution (SCE-UA) method [Duan et al., 1992]. Each of the six parameter sets was then used to simulate the full 35-year calibration period. This gave an ensemble of six series covering the full 35-year calibration period, each representing a possible mode of catchment response. Figure 2b shows an example of an ensemble, with flow on a log scale so that low flows are visible. The red series are generated with parameters calibrated over dry subsets, while the blue series are generated with parameters calibrated over wet subsets. These ensembles were used in the model weighting procedure described in Section 0.

Commented [A18]: This citaion is not found in Section Title.Please check if it is citation or not and proceed further

Conceptually, the ensemble shown in **Figure 2**b might align with our general expectations based on theoretical catchment behavior. The observations tend to be closer to the wet-calibrated series during periods of high flow and closer to the dry-calibrated series during periods of low flow. If the modeled ensembles follow this general pattern, it is likely that a series weighting technique will be effective in matching appropriate parameters with climate conditions.

Three experiments were conducted as part of this investigation. Starting with the parameter sets calibrated under wet and dry conditions for each catchment, an experiment was performed to investigate:

1 The general ability of hydrologic model parameters to transfer across distinct climate states.

A second experiment utilized the final five years of available data (kept aside for model testing) to explore:

2 The potential for improved simulation under specified future climate conditions through preferentially applying parameters calibrated under similar past conditions.

The results of the first two investigations led to a third experiment where the four GR4J parameters were defined dynamically as a function of climate conditions. The updated model structure is described in Section 0. This final experiment investigated:

Commented [A19]: This citaion is not found in Section Title.Please check if it is citation or not and proceed further

3 The potential for improved simulation through allowing the model parameters to vary based on climatic conditions.

Each experiment is discussed in a dedicated section (Sections 0, 0 and 0 respectively), with an outline of the overall process shown in **Figure 3**. Because the justification and methods for each experiment are influenced by the results of the experiment before, discussion of the specific steps taken are outlined in the individual sections.

Commented [A20]: This citaion is not found in Section Title.Please check if it is citation or not and proceed further

3 Parameter Transferability Between Periods

It has been shown that wet periods are generally easier to model hydrologically than dry periods, and that model performance degrades less when parameters are transferred between climatically similar periods than climatically contrasting periods [Vaze et al., 2010]. The concept of skillful parameter 'donors' and 'acceptors' was introduced by Smith et al. [2018] in the context of parameter transfer between different catchments. They found that some catchments acted as good donors (i.e. calibration parameters could be generalized to other catchments) while others acted as good acceptors (i.e. the catchment could be skillfully simulated with a broad range of parameter values). Here, we consider this concept when transferring parameters over time rather than space. This section describes testing undertaken with the 164 study catchments to understand:

- The calibration performance of GR4J under dry and wet conditions for the study catchments
- The performance degradation when parameters from a given period are 'donated' to climatically similar periods
- The performance degradation when parameters from similar periods are 'accepted' by a given period
- The performance degradation when parameters from a given period are 'donated' to climatically contrasting periods
- The performance degradation when parameters from contrasting periods are 'accepted' by a given period

The objective function used for calibration was the widely-used Nash-Sutcliffe Efficiency (NSE) [Nash and Sutcliffe, 1970]. The results presented are based on four-year climatic periods; similar results were obtained when periods were defined over other timeframes.

shows the distribution of calibration NSE values for dry and wet periods in 164 catchments (where dry1 is the driest period, dry2 is the second driest, etc.). All calibrations have high median performance (NSE ~0.8) indicating that four-year periods are generally adequate for model calibration. However, there are catchments where calibration performance is poor, particularly during the dry periods. Many catchments in Australia (including some used in this study) are intermittent and there may be many zero flow days in the drier periods, which can reduce the information available for calibration. It seems that the higher information content of the wetter periods is providing an advantage for calibration of the more difficult catchments.

When parameters from climatically similar periods were transferred (i.e. a dry calibrated parameter set to another dry period and similarly for wet calibrated parameters), the performance of wet and dry periods contrasted more strongly (**Figure 5**a and **Figure 5**b). Wet periods both donated and accepted parameters between themselves substantially more effectively than dry periods. Interestingly, of the three dry periods, the most extreme (dry1) was the best donor but the worst acceptor. This indicates that the transferability of parameters is more dependent on the nature of the period being modeled than the calibration period. Specifically, the driest period (hence most difficult to simulate) experienced the most performance degradation when calibration parameters from other periods were applied, but its own calibration parameters still appear to have been reasonably specified (otherwise dry1 would have been a very poor donor).

The same finding was reinforced when parameters were transferred between climatically contrasting periods (i.e. a dry calibrated parameter set to a wet period and a wet calibrated parameter set to a dry period) (**Figure 5**c and **Figure 5**d). Wet periods, being easy to model and with well-specified parameters, appeared to be poor donors because they were donating parameters to dry periods, being difficult to model. Despite the fact that dry period parameters may not be as well specified (evidenced by lower calibration NSEs, **Figure 1**), the dry parameters appeared to be better donors because the target wet periods were easier to model accurately and hence good acceptors (**Figure 5**c). The key finding here is that the conditions of the accepting period are more important than the conditions of the donating period. Of the three dry subsets (with dry1 being the driest and dry3 the least dry), more extreme dry periods experienced more performance degradation when wet-period calibration parameters were applied (**Figure 5**d). This is in line with previous findings [Coron et al., 2012; Vaze et al., 2010] and suggests that areas where climate change leads to drier conditions could be particularly difficult to model accurately using only historical data for model set-up.

For both the dry and wet periods, parameter acceptance was better when the donor period was climatically similar than when it was contrasting (Figure 5b and **Figure 5**d). This is both intuitive and in line with previous findings [Vaze et al., 2010]. It also indicates that, when modeling a future period under specified climatic conditions, parameters calibrated during similar past climatic periods will be most informative. This observation underpins the model weighting scheme that aims to improve simulation of an unseen validation period, detailed in the following section.

4 Weighting Modeled Series With Wet/Dry Calibration Parameters

Given that parameters can be most readily transferred between climatically similar periods (Section 0), it follows that simulation of an unseen future period could be best represented by parameters calibrated to a similar past period. This suggests that, to simulate drier future conditions (for example), it may be advantageous to only use dry historical periods in model calibration. However, this would require discarding much of the recorded data that could still contain relevant information about catchment function. In this section, we propose a model weighting scheme that uses both wet and dry calibration parameters to generate a number of series representing possible catchment behavior (as described in Section 0). The series are then weighted based on climatic similarity to a target simulation period. The subsections describing this process are structured as follows:

Commented [A21]: This citaion is not found in Section Title.Please check if it is citation or not and proceed further

Commented [A22]: This citaion is not found in Section Title.Please check if it is citation or not and proceed further

- Section 4.1 describes a testing framework that was implemented to inform decisions that could impact the model weighting strategy. Here, the length of distinct subsets for GR4J calibration is examined, but the same framework could be used to aid other decisions (e.g. choice of climate indicator).
- Section 4.2**.** describes the method for applying weights to the ensemble of series. This includes a strategy for defining how strongly the calibration climate similarity to the target period should impact series weights.
- Section 4.3. presents the application of the weighting strategy over an unseen validation period and reports on its performance.

4.1 Calibration testing framework

One important trade-off that needs to be considered in developing a model weighting method is the length of climatically distinct subsets for calibration. Naturally, shorter calibration subsets will show greater distinction between wet and dry conditions. However, if the subsets are too short, parameters may be poorly defined and therefore less useful for simulating catchment behavior outside the calibration period. It was necessary to develop a framework that could predict the likely performance of the method with calibration subsets of different lengths, without referring to the final validation data (which needs to remain unseen for testing the weighting strategy). This framework is described below for an example in which the subsets are one year long:

1. Calibrate the model over the three driest (dry1, dry2, dry3) and three wettest (wet1, wet2, wet3) one-year subsets of the calibration period (water years 1974-2009).

2. Produce six modeled series that cover the full calibration period using each of the six parameter sets obtained in (1).
3. For each series, calculate the difference between the modeled flows and the observations for every day of the calibration period, then take the one-year moving average of this difference. This gives a measure of how different the series is from the observations at a one-year timescale, calculated starting at the first day, second day, etc. through to one year before the final day. The inverse of this measure quantifies similarity of the series to the observations (i.e. how well the series matches the observations for every possible one-year period). The similarity values are then normalized so that their sum is one.
4. For every possible one-year period (i.e. starting at the first day, second day, etc. through to one year before the final day), multiply the series by the similarity value for that period calculated in (3). This gives a theoretical 'optimized weighting' based on true distance from the observed flow.
5. Calculate the correlation between the optimized weights and the average RDI_n over the corresponding one-year period. If there is a strong correlation between the true climate conditions and the weights assigned (i.e. during a dry one-year period the series calibrated over dry subsets are favored), this indicates that the climatic conditions are a good predictor of appropriate model parameters and the model weighting scheme is likely to perform well under validation.
6. Repeat this for all 164 catchments and calculate the average correlations. Negative correlations are expected between the dry series weights and RDI_n, and positive correlations are expected between the wet series weights and RDI_n.

This framework was applied to test the prediction strength of RDI_n in determining appropriate model parameters for one, two, three, four and five year subsets. In each case, six subsets were defined (hence six different series produced in step (2)) and the similarity was calculated at a timescale matching the subset length (e.g. a two-year moving average was used in step (3) when the subsets were two years long). The correlations obtained are shown in **Table 1**, averaged over all 164 catchments. Note that the issue of defining non-overlapping subsets (discussed in Section 2) affected some catchments in the four and five-year subset tests, so some catchments were discarded. This is based on the timing of the three wettest and three driest periods only and not on other climatic features, so the specific catchments removed should be somewhat random and not lead to overall bias in the results. This was confirmed by rerunning with only catchments for which six independent subsets could be defined in every case (not shown here), and this gave similar overall results. Based on the average correlations between climate conditions and optimized weights (see step 5), it seems that the best calibration subsets were defined over four-year periods.

4.2 Defining the weighting algorithm

The weighting strategy should preference information from climatically similar calibration periods without entirely discarding information from climatically different calibration periods. Equation (3) was developed to define calibration period (cal.per) similarity (sim) to the target period (tar.per):

$$sim[cal.per] = \left(\frac{1}{\left| \left(RDI_n[cal.per] - RDI_n[tar.per] \right) \right|} \right)^{\omega} \quad (3)$$

The weighting parameter (ω) determines how strongly the difference between calibration subset RDI_n and the target period RDI_n influences a series' weight. If ω equals zero, the series are simply averaged. A relatively large value of ω indicates heavy weighting on the most climatically similar calibration periods, with the remaining series having limited influence. Positive but smaller values of ω represent a preference for series calibrated under more similar conditions, but with substantial information also extracted from series calibrated under contrasting conditions. We aim to define a generalized value of ω across all study catchments in order to understand the typical importance of climate in determining model parameters. This strategy avoids excessive dependence on the particular conditions under which ω is calibrated, since a large number of sites will represent a variety of conditions without requiring lengthy data records. Additionally, if a widely applicable value of ω can be determined, the strategy can be applied to new catchments without recalibration.

As noted in Section 0, the last five years of data were kept aside from the initial selection of calibration subsets (see **Error! Reference source not found.**a). The first two water years of this unseen data starting in 2009 (herein referred to as the ‘ω calibration period’) were used to calibrate ω, with the last two water years saved for validation. The third water year was discarded to minimize dependence between the two two-year periods. Two stations (G0050115 and G0060005) recorded zero flow over one of these two-year periods so they were removed from both this step and the application step described in Section 0. Calibration of ω was undertaken through the following steps:

Commented [A23]: This citaion is not found in Section Title.Please check if it is citation or not and proceed further

Commented [A24]: This citaion is not found in Section Title.Please check if it is citation or not and proceed further

1 Select a ω value based on the SCE-UA algorithm (initially random).

2. For the first catchment, calculate the similarity between each calibration subset RDI$_n$ and the ω calibration period RDI$_n$ (the target in this case) as per Equation (3). The greater the value of ω, the more strongly RDI$_n$ difference will influence this measure.

3. Standardize the similarity measures to sum to one (as in a weighted average procedure), giving fractional weights for each series based on climatic similarity to the ω calibration period. Equation (4) shows an example calculation for the series calibrated over the driest subset, dry1.

$$weight\left[dry1\right]=\frac{sim\left[dry1\right]}{sim\left[dry1\right]+sim\left[dry2\right]+sim\left[dry3\right]+sim\left[wet1\right]+sim\left[wet2\right]+sim\left[wet3\right]} \quad (4)$$

4. Take the six series modelled over the ω calibration period (each with parameters from one of the calibration subsets) and multiply the by the calculated weight, then sum them to give an overall weighted prediction.

5. Compare the weighted prediction with the observed flow to calculate C_{2M}, a bounded formulation that has been proposed as an alternative to NSE for averaging results across large samples of catchments, described by Mathevet et al. [2006] and reproduced in Equation (5).

$$C_{2M} = \frac{NSE}{2 - NSE} \quad (5)$$

6 Repeat steps (2) to (6) for all catchments and calculate the average C_{2M}.

These steps were repeated using the SCE-UA method until the ω value that optimized average C_{2M} was identified, with ω allowed to vary between -10 and 10. Note that C_{2M} was used as the objective function in this process instead of NSE to keep the possible range between -1 and 1, preventing poorly performing catchments from having a disproportionate impact on the average calculated in step 6 and hence the calibrated value for ω. Aggregated performance results (e.g. mean performance across all catchments) are herein reported in C_{2M} for the same reason. Time series calibrations (Sections 0, 0 and 0) were undertaken using the more widely-applied NSE criterion, but since

Commented [A25]: This citaion is not found in Section Title.Please check if it is citation or not and proceed further

there is a monotonic relationship between NSE and C_{2M}, use of either function would give the same parameter sets in these cases.

The following parameters were obtained for the five GR4J calibration subset lengths (**Error! Reference source not found.**):

In all five cases, values above zero but below 10 (the user specified upper limit for calibration) were selected, indicating that the weighting scheme tends to preference series calibrated under similar climatic conditions, but also extracts information from series representing contrasting conditions. For the four-year calibration subsets, the value of ω was particularly small (0.07), which means that all six series tend to be weighted relatively evenly (for most catchments, proportional weightings tended to vary between about 0.14 and 0.20 for the six different series). In contrast, the series associated with shorter calibration subsets gave optimal performance with heavier weighting on more similar conditions. This could indicate that calibration under more extreme conditions (i.e. the driest or wettest single year in the entire record) does not provide much useful information for the model when simulating more average conditions.

The exercise was repeated excluding the catchments for which independent four and five-year calibration subsets could not be defined (so all five tests used the same subset of catchments). In this test, larger values of ω were obtained in every case (**Error! Reference source not found.**). This highlights that the ω parameter is dependent on the combination of catchments used. Therefore, ω calibration was repeated using different catchment groupings, based on four-year subsets (since these performed best in the testing framework described in Section 0). The groupings were defined based on hydrological, physical and climatic characteristics of the watersheds, with thresholds set such that the number of catchments in each group was fairly consistent.

Commented [A26]: This citaion is not found in Section Title.Please check if it is citation or not and proceed further

The calibrated values for ω are shown in **Error! Reference source not found.**. Consistent trends in ω across a grouping are highlighted in shading (e.g. consistently increasing ω relative to the proportion of rainy days), where the largest ω value is shown in the darkest shade. While there is potential for overlap in groups (for example, catchments with a high proportion of zero flow days may also have a small area), criteria sets that could be expected to give overlaps were identified and checked. If there were less than 20 catchments that were different between the two groups, one of the criteria sets was discarded. As such, the catchment groupings shown in **Error! Reference source not found.** are all distinctly different.

For some catchment groupings, no discernible pattern exists in the calibrated values of ω. However, in other cases there does seem to be a trend. The value for ω tended to increase with decreasing dryness (indicated by zero rain days) and increasing latitude. This suggests that wet catchments with strong rainfall seasonality (as in the more northern parts of Australia) may benefit particularly from the use of the model weighting technique, since the ω parameter indicates fairly strong preference for series calculated under similar conditions to the ω calibration period. It is also interesting that the value for ω tended to increase with increasing flow variance (relative to average flow). This suggests that catchments with more variable flow regimes have higher parameter dependence on climate, indicating that a single set of parameters may not appropriately capture the different catchment states. Negative parameter values indicate that the group of catchments is not benefiting from the weighting procedure as intended – rather, the best results are obtained from series calibrated under the conditions that deviate most from the ω calibration period (i.e. in the extreme, an average period would be modeled by taking the mean of the series associated with the driest and wettest calibration periods). A preference for negative ω values in certain catchments was associated with three main patterns:

1. All six models of the catchment performed very poorly in the ω calibration period. In these cases, it was not meaningful to weight the series and unexpected values of ω often optimized results.

2. There was no meaningful relationship between calibration subset conditions and modeled flow behavior. For example, dry-calibrated series sometimes matched high flows more accurately than wet-calibrated series. This was fairly common in the results and suggests that the GR4J models may not be reflecting catchment response realistically.
3. One series gave very bad results. The ω calibration algorithm would tend to 'distance' this series by choosing a value of ω that minimized its contribution. If the series had a similar calibration RDI_n to the ω calibration period, this would be a large negative value of ω.

The overall efficacy of the weighting method for the full set of catchments, as well as groupings based on different properties, is investigated in Section 0.

Commented [A27]: This citaion is not found in Section Title.Please check if it is citation or not and proceed further

4.3 Application of the weighting strategy

For each calibration time period and in each catchment, the six series (Section 0) were weighted and aggregated over the test period (two water years starting in 2012). The steps are outlined as follows:

Commented [A28]: This citaion is not found in Section Title.Please check if it is citation or not and proceed further

1. Take the appropriate ω value (depending on calibration subset length) from **Error! Reference source not found.**.
2. For the first catchment, calculate RDI_n similarity between each calibration subset and the unseen test period (Equation (3)).
3. Standardize similarity values to sum to one, giving weights for each associated series (Equation (4)).

4. Multiply the test-period modelled series associated with each calibration subset by the appropriate weight and sum to give a weighted prediction.

The results of the weighting strategy application were benchmarked against a traditional calibration approach, where data over the full calibration period (35 years) was used to calibrate one set of parameters under the assumption of stationarity. The performance indicators were:

- The mean difference in C_{2M} between the weighting strategy and the traditional calibration approach.
- The median difference in C_{2M} between the weighting strategy and the traditional calibration approach.
- The percentage of catchments for which the weighting strategy improved prediction.

Based on all three indicators, the best performance was achieved with four-year calibration subsets (**Error! Reference source not found.**). This was predicted by the testing framework described in Section 0, indicating that the framework is appropriate for assessing decisions that influence the weighting strategy.

Commented [A29]: This citaion is not found in Section Title.Please check if it is citation or not and proceed further

The results presented here suggest that extracting past information in a targeted way based on climatic conditions may improve simulation of a climatically distinct future period. When the optimum length of calibration subset (four years) was used in the weighting strategy, predictions improved in 66% of catchments, although the mean and median improvements in C_{2M} were fairly small. It is possible that the strategy would provide more benefit in a climate change assessment, where the testing period would likely be more different to the average past conditions, since model weighting is unlikely to provide significant benefit if the testing period is not climatically distinct from the full calibration period. The four-year-subset results for catchments where the test period conditions deviated from the average are shown in **Error! Reference source not found.**. For wet validation periods ($RDI_n > 0.1$), the weighting scheme improved results for 74% of catchments, compared to 66% over the full set of catchments, and the median improvement was also higher (**Error! Reference source not found.**). For dry validation periods ($RDI_n < -0.1$), the mean and median NSE improvements were both higher than for the full set of catchments. This indicates that the scheme tends to give the greatest improvements under dry conditions, but that improvements are more consistent under wet conditions. As expected, model weighting appears to be more advantageous under conditions that deviate notably from the average climate at a catchment.

The process was repeated for the different catchment groupings outlined in Section **Error! Reference source not found.**, using the ω parameter values given in **Error! Reference source not found.**. The model test results are shown in **Error! Reference source not found.**. Consistent trends in performance across a criterion (i.e. all three indicators moving in the same direction) are highlighted in shading, where the greatest improvement is shown in the darkest shade.

Unexpectedly, the results indicate that the calibrated value of ω does not necessarily predict the validation performance of the model weighting strategy under unseen conditions. In the cases of both flow variance and latitude, the groupings with negative and very small ω values (respectively) in **Error! Reference source not found.** had the best validation performance. Performance improved especially strongly as runoff ratio decreased, perhaps indicating that dry, intermittent catchments (generally considered difficult to model) may benefit particularly from model weighting. However, overall it seems that the strategy is highly dependent on the specific set of catchments used, and it is not necessarily straightforward to predict how ω is likely to vary under different conditions. It follows that a more flexible approach where parameter changes are specifically calibrated for each catchment might be worth investigating. This finding provided the motivation for the work described in Section 0.

Commented [A30]: This citaion is not found in Section Title.Please check if it is citation or not and proceed further

5 Climate-Dependent Dynamic Model Parameters

The results discussed in Section 0, particularly those shown in **Error! Reference source not found.** and **Error! Reference source not found.**, demonstrate that it is not straightforward to generalize the weighting strategy across different

Commented [A31]: This citaion is not found in Section Title.Please check if it is citation or not and proceed further

catchments. It is unlikely that a universal weighting strategy could be defined such that all catchments experienced improved performance consistently under future conditions. Therefore, for the third experiment, a more flexible approach was adopted in which all four GR4J parameters were allowed to vary based on RDI_n. This dynamic version of the model was calibrated individually for each catchment, meaning that model parameters' dependence on climate did not need to be generalized across multiple catchments with different sensitivities.

In the dynamic version of GR4J, each of the four standard GR4J parameters (X_1, X_2, X_3 and X_4) is defined as a 3-parameter sigmoidal function. This form was selected because it is substantially more flexible than a simple linear function, but still requires that the parameter change somewhat consistently with RDI_n (i.e. the value cannot increase with RDI_n up to a certain point, then decrease). This increases the likelihood of parameter shifts being physically interpretable and decreases the extent of equifinality expected in the flexible model. The functions defining the four GR4J parameters are as follows (Equation (6)):

$$\begin{aligned}
X_1 &= t_{0\_1} + t_{1\_1}\frac{e^{a_1 RDI}}{e^{a_1 RDI}+1} \\
X_2 &= t_{0\_2} + t_{1\_2}\frac{e^{a_2 RDI}}{e^{a_2 RDI}+1} \\
X_3 &= t_{0\_3} + t_{1\_3}\frac{e^{a_3 RDI}}{e^{a_3 RDI}+1} \\
X_4 &= t_{0\_4} + t_{1\_4}\frac{e^{a_4 RDI}}{e^{a_4 RDI}+1}
\end{aligned} \qquad (6)$$

RDI_n is defined on a monthly basis, so the parameters vary at a monthly timestep based on the overall conditions of that month. The sigmoidal function parameters (e.g. t_{0_1}, t_{1_1} and a_1 to define X_1) are calibrated to give a function specifying the four original parameters, meaning that the number of calibration parameters is 12 for the dynamic model. Naturally, this leads to a substantial reduction in calibration efficiency, which could only be justified if the method significantly outperformed both traditional calibration and the model weighting strategy. The original static model is one possible realization of the dynamic model (i.e. if t_{1_1}, t_{1_2}, t_{1_3} and t_{1_4} all calibrate to zero), so the option of one or more parameters remaining static is inherently available if this provides the best calibration.

Because catchment behavior under very wet or dry climate conditions could be more extreme than overall catchment behavior, we decided to add flexibility by allowing greater variation in parameter values than is generally applied for GR4J (**Error! Reference source not found.**). For comparison, the standard four-parameter version of GR4J was also run with the expanded parameter ranges.

The models were calibrated based on NSE over the same 35 year period (water years 1974-2009) used in the model weighting strategy calibration step. Again, the final five water years were kept aside for validation. In this case, no second calibration period was required (as in ω calibration for the experiment described in Section 0), so the full five year period was used for validation. The 12 calibrated parameters were used to define the four parameters in the standard version of GR4J dynamically with respect to RDI_n. An example of how the parameters may vary with climate is shown in **Error! Reference source not found.**.

Commented [A32]: This citaion is not found in Section Title.Please check if it is citation or not and proceed further

The static and dynamic models were compared over the calibration and validation periods. As would be expected from a more flexible model, the dynamic model outperformed the static model over the calibration period in 91% of cases, indicating that the dynamic calibration was mostly successful. Theoretically, the dynamic model should always give equal or better calibrations than the static model, since the static model is nested within the dynamic model. In cases where the dynamic calibration NSE was lower, it indicates that the calibration algorithm located a local peak in the objective function rather than the true optimum, but this is difficult to avoid entirely for a model with a large number of parameters. The dynamic model also outperformed the static model over the validation period in 59% of catchments. Results aggregated across catchments are reported in terms of C_{2M}, as discussed in Section **Error! Reference source not found.**, because this avoids excessive influence from individual catchments with very poor performance. The mean and median C_{2M} values were higher for the dynamic model in both the calibration and validation periods (**Error! Reference source not found.**). Interestingly, the variance in validation performance was substantially smaller for the dynamic model, indicating that the performance was more consistent across the 164 catchments. This suggests that the dynamic model may be particularly useful when simulating catchments that the static model represents poorly.

Further investigation revealed that the dynamic model offered far greater performance improvements for dry validation periods. Most of the catchments tested were relatively wet over the water years from 2009 to 2014, but 29 catchments had average RDI_n values less than zero for the validation period. For these catchments, the dynamic model outperformed the static model in 76% of cases (versus 56% of the models where average validation period RDI_n was greater than zero). It is well established that existing conceptual models achieve better validation results over wet periods than dry periods [Vaze et al., 2010], so the dynamic model may not offer a significant advantage during wet periods. When only the 29 simulations with dry validation periods were considered (**Error! Reference source not found.**), the validation performance improvement was more substantial than for the full suite of models (**Error! Reference source not found.**).

These results suggest that there may be promise in updating existing conceptual models by using climate-dependent parameter values, especially for drier future climate scenarios. There was no clear difference in performance between catchments with different areas or latitudes. Intuitively, it seems that the utility of climate-varying model parameters should depend on how they relate to physical processes in the catchment. This can be difficult to ascertain in conceptual models, but different conceptual models inherently account for physical processes in widely varying ways. Therefore, future work should test the potential of climate-varying parameters in other conceptual models.

In this study, we found that dynamic calibration resulted in widely varying effects in different catchments. For some catchments, the optimal values for three parameters remained static with climate and just one parameter varied notably. This suggests that, for some catchments, it may be possible to reduce the total number of calibration parameters by representing three GR4J parameters as static and one as dynamic. However, for other catchments several parameters varied together to give the optimal dynamic model. This is evident in the results for station 304497 shown in **Error! Reference source not found.**, where X_1 remains stable with climatic conditions but X_2, X_3 and X_4 all vary substantially. There were some patterns identified across catchments (e.g.

X_2 tended to increase with increasing RDI_n), but these were not sufficiently consistent to inform potential model structural changes.

While the results presented here show potential for model improvement through accounting for parameter dependence on climate, they certainly don't solve the problem of model nonstationarity entirely. Efficacy varied across catchments, and poor performance often seemed to be associated with models where the parameters did not seem to represent catchment behavior as expected (for example, when wet-period calibration parameters performed best under dry testing conditions). This implies that some minimum level of reasonable process representation may be required to reliably relate model parameters to climate conditions and ultimately to improve model performance under nonstationarity. Deb et al. [2019] recently showed that better representing surface water / ground water interactions improved hydrologic simulation under change. These sorts of processes are notoriously difficult to model in some highly intermittent Australian catchments with complex spatially and temporally varying runoff mechanisms [Dean et al., 2016], so it is likely to be some time before hydrologists can rely entirely on physical process representation rather than calibrated conceptual parameters. However, lumped conceptual models like GR4J make no attempt to represent any of these physical complexities and may not be suitable for continued use in a changing climate.

5 Conclusion

Our results, along with much prior work, indicate that transferring conceptual hydrologic models to unseen future climate states will result in performance degradation. The experiments presented in this paper show that increased reliability may be achievable through preferential use of climatically similar calibration data, while still extracting some information from contrasting periods as well. However, there is high variability in performance improvements between catchments. In many cases, this seems to relate to the conceptual model failing to reflect realistic catchment behavior changes under different climate states. It also seems that the importance of climate conditions in determining model parameters cannot be easily generalized across numerous catchments. An alternative approach, defining climate-dependent dynamic parameters, was also tested. The dynamic model gave improved validation performance, especially under dry conditions. In combination, these experiments show that the performance of conceptual models under altered climate conditions can be improved through both targeted use of calibration data and model structural adjustments. Since previous studies have demonstrated parameter dependence on climate for a wide range of hydrologic models [Guo et al., 2018; Merz et al., 2011; Vaze et al., 2010], it is likely that models other than GR4J could also see some prediction improvement through application of the methods outlined here. However, these sorts of strategies are unlikely to offer a 'silver bullet' that reliably improves model predictions in all catchments, and so it is important to continue improving representation of physical dynamics that contribute to nonstationarity in conceptual hydrologic models.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Professor Upmanu Lall from Columbia University, New York, for helpful advice and discussions that improved this work. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. Clare Stephens receives an Australian Government Research Training Program Scholarship. Dr Marshall was supported through an Australian Research Council Future Fellowship FT120100269. The streamflow data used for this study can be obtained from the BoM website at http://www.bom.gov.au/water/hrs/. The rainfall and temperature data is accessible online through CSIRO and BoM's Australian Water Availability Project.

References

Commented [A33]: Please note that we matched the supplied reference content against the Crossref.org database and provided required missing information in the output. Kindly check the output of all the references.

Anshuman, A., Kunnath-Poovakka, A., and Eldho, T. I. (2018), Performance evaluation of conceptual rainfall-runoff models GR4J and AWBM, *ISH Journal of Hydraulic Engineering*, 1-10, doi: 10.1080/09715010.2018.1556124.

Anshuman et al., 2018 A. Anshuman A. Kunnath-Poovakka T.I. Eldho Performance evaluation of conceptual rainfall-runoff models GR4J and AWBM ISH Journal of Hydraulic Engineering 1–10 2018 10.1080/09715010.2018.1556124

Brigode, P., Oudin, L., and Perrin, C. (2013), Hydrological model parameter instability: A source of additional uncertainty in estimating the hydrological impacts of climate change?, *Journal of Hydrology*, *476*, 410-425, doi: 10.1016/j.jhydrol.2012.11.012.

Brigode et al., 2013 P. Brigode L. Oudin C. Perrin Hydrological model parameter instability: A source of additional uncertainty in estimating the hydrological impacts of climate change? Journal of Hydrology 476 2013 410 425 10.1016/j.jhydrol.2012.11.012

Coron, L., Andréassian, V., Perrin, C., Bourqui, M., and 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, doi: 10.5194/hess-18-727-2014.

Coron et al., 2014 L. Coron V. Andréassian C. Perrin M. Bourqui F. Hendrickx 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 2014 727 746 10.5194/hess-18-727-2014

Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., and Hendrickx, F. (2012), Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments, *Water Resources Research*, *48*(5), doi: 10.1029/2011WR011721.

Coron et al., 2012 L. Coron V. Andréassian C. Perrin J. Lerat J. Vaze M. Bourqui F. Hendrickx Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments Water Resources Research 48 5 2012 10.1029/2011WR011721

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

Dakhlaoui et al., 2017 H. Dakhlaoui D. Ruelland Y. Tramblay Z. Bargaoui Evaluating the robustness of conceptual rainfall-runoff models under climate variability in northern Tunisia Journal of Hydrology 550 2017 201 217 10.1016/j.jhydrol.2017.04.032

Dean, J. F., Camporese, M., Webb, J. A., Grover, S. P., Dresel, P. E., and Daly, E. (2016), Water balance complexities in ephemeral catchments with different land uses: Insights from monitoring and distributed hydrologic modeling, *Water Resources Research*, *52*(6), 4713-4729, doi: 10.1002/2016WR018663.

Dean et al., 2016 J.F. Dean M. Camporese J.A. Webb S.P. Grover P.E. Dresel E. Daly Water balance complexities in ephemeral catchments with different land uses: Insights from monitoring and distributed hydrologic modeling Water Resources Research 52 6 2016 4713 4729 10.1002/2016WR018663

Deb, P., Kiem, A. S., and Willgoose, G. (2019), A linked surface water-groundwater modelling approach to more realistically simulate rainfall-runoff non-stationarity in semi-arid regions, *Journal of Hydrology*, *575*, 273-291, doi: https://doi.org/10.1016/j.jhydrol.2019.05.039.

Deb et al., 2019 P. Deb A.S. Kiem G. Willgoose A linked surface water-groundwater modelling approach to more realistically simulate rainfall-runoff non-stationarity in semi-arid regions Journal of Hydrology 575 2019 273 291 10.1016/j.jhydrol.2019.05.039

Duan, Q., Sorooshian, S., and Gupta, V. (1992), Effective and efficient global optimization for conceptual rainfall-runoff models, *Water Resources Research*, *28*(4), 1015-1031, doi: 10.1029/91WR02985.

Duan et al., 1992 Q. Duan S. Sorooshian V. Gupta Effective and efficient global optimization for conceptual rainfall-runoff models Water Resources Research 28 4 1992 1015 1031 10.1029/91WR02985

Fowler, K. J. A., Peel, M. C., Western, A. W., Zhang, L., and Peterson, T. J. (2016), Simulating runoff under changing climatic conditions: Revisiting an apparent deficiency of conceptual rainfall-runoff models, *Water Resources Research*, doi: 10.1002/2015WR018068.

Fowler et al., 2016 K.J.A. Fowler M.C. Peel A.W. Western L. Zhang T.J. Peterson Simulating runoff under changing climatic conditions: Revisiting an apparent deficiency of conceptual rainfall-runoff models Water Resources Research 2016 10.1002/2015WR018068

Grigg, A. H., and Hughes, J. D. (2018), Nonstationarity driven by multidecadal change in catchment groundwater storage: A test of modifications to a common rainfall–run-off model, *Hydrological Processes*, *32*(24), 3675-3688, doi: 10.1002/hyp.13282.

Grigg and Hughes, 2018 A.H. Grigg J.D. Hughes Nonstationarity driven by multidecadal change in catchment groundwater storage: A test of modifications to a common rainfall–run-off model Hydrological Processes 32 24 2018 3675 3688 10.1002/hyp.13282

Guo, D., Westra, S., and Maier, H. R. (2016), An R package for modelling actual, potential and reference evapotranspiration, *Environmental Modelling & Software*, *78*, 216-224, doi: http://dx.doi.org/10.1016/j.envsoft.2015.12.019.

Guo et al., 2016 D. Guo S. Westra H.R. Maier An R package for modelling actual, potential and reference evapotranspiration Environmental Modelling & Software 78 2016 216 224 10.1016/j.envsoft.2015.12.019

Guo, D., Westra, S., and Maier, H. R. (2017), Impact of evapotranspiration process representation on runoff projections from conceptual rainfall-runoff models, *Water Resources Research*, *53*(1), 435-454, doi: 10.1002/2016WR019627.

Guo et al., 2017 D. Guo S. Westra H.R. Maier Impact of evapotranspiration process representation on runoff projections from conceptual rainfall-runoff models Water Resources Research 53 1 2017 435 454 10.1002/2016WR019627

Guo, D., Johnson, F., and Marshall, L. (2018), Assessing the Potential Robustness of Conceptual Rainfall-Runoff Models Under a Changing Climate, *Water Resources Research*, <xocs:firstpage xmlns:xocs=""/>, doi:

10.1029/2018WR022636.

Guo et al., 2018 Guo, D., Johnson, F., and Marshall, L. (2018), Assessing the Potential Robustness of Conceptual Rainfall-Runoff Models Under a Changing Climate, *Water Resources Research*, <xocs:firstpage xmlns:xocs=""/>, doi: 10.1029/2018WR022636.

Humphrey, G. B., Gibbs, M. S., Dandy, G. C., and Maier, H. R. (2016), A hybrid approach to monthly streamflow forecasting: Integrating hydrological model outputs into a Bayesian artificial neural network, *Journal of Hydrology*, *540*, 623-640, doi: https://doi.org/10.1016/j.jhydrol.2016.06.026.

Humphrey et al., 2016 G.B. Humphrey M.S. Gibbs G.C. Dandy H.R. Maier A hybrid approach to monthly streamflow forecasting: Integrating hydrological model outputs into a Bayesian artificial neural network Journal of Hydrology 540 2016 623 640 10.1016/j.jhydrol.2016.06.026

IPCC (2013), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, 1535 pp., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

IPCC, 2013 IPCC (2013), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, 1535 pp., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Maraun, D., et al. (2010), Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user, *Reviews of Geophysics*, *48*(3), doi: 10.1029/2009RG000314.

Maraun et al., 2010 D. Maraun et al. Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user Reviews of Geophysics 48 3 2010 10.1029/2009RG000314

Mathevet, T., Michel, C., Andréassian, V., and Perrin, C. (2006), A bounded version of the Nash-Sutcliffe criterion for better model assessment on large sets of basins, in *Large Sample Basin Experiments for Hydrological Model Parameterization: Results of the Model Parameter Experiment – MOPEX*, edited by V. Andréassian, A. Hall, N. Chahinian and J. Schaake, IAHS Publ.

Mathevet et al., 2006 Mathevet, T., Michel, C., Andréassian, V., and Perrin, C. (2006), A bounded version of the Nash-Sutcliffe criterion for better model assessment on large sets of basins, in *Large Sample Basin Experiments for Hydrological Model Parameterization: Results of the Model Parameter Experiment – MOPEX*, edited by V. Andréassian, A. Hall, N. Chahinian and J. Schaake, IAHS Publ.

McGuinness, J. L., and Bordne, E. F. (1972), A comparison of lysimeter-derived potential evapotranspiration with computed values, *Rep.*, Agricultural Research Service, U.S. Dept. of Agriculture, Washington DC.

McGuinness and Bordne, 1972 McGuinness, J. L., and Bordne, E. F. (1972), A comparison of lysimeter-derived potential evapotranspiration with computed values, *Rep.*, Agricultural Research Service, U.S. Dept. of Agriculture, Washington DC.

Merz, R., Parajka, J., and Bloschl, G. (2011), Time stability of catchment model parameters: Implications for climate impact analyses, *Water Resources Research*, *47*, doi: 10.1029/2010WR009505.

Merz et al., 2011 R. Merz J. Parajka G. Bloschl Time stability of catchment model parameters: Implications for climate impact analyses Water Resources Research 47 2011 10.1029/2010WR009505

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., and Stouffer, R. J. (2008), Stationarity Is Dead: Whither Water Management?, *Science*, *319*(5863), 573-574.

Milly et al., 2008 P.C.D. Milly J. Betancourt M. Falkenmark R.M. Hirsch Z.W. Kundzewicz D.P. Lettenmaier R.J. Stouffer Stationarity Is Dead: Whither Water Management? Science 319 5863 2008 573 574

Nash, J. E., and Sutcliffe, J. V. (1970), River flow forecasting through conceptual models part I — A discussion of principles, *Journal of Hydrology*, *10*(3), 282-290, doi: 10.1016/0022-1694(70)90255-6.

Nash and Sutcliffe, 1970 J.E. Nash J.V. Sutcliffe River flow forecasting through conceptual models part I — A discussion of principles Journal of Hydrology 10 3 1970 282 290 10.1016/0022-1694(70)90255-6

Osuch, M., Romanowicz, R. J., and Booij, M. J. (2015), The influence of parametric uncertainty on the relationships between HBV model parameters and climatic characteristics, *Hydrological Sciences Journal*, *60*(7-8), 1299-1316, doi: 10.1080/02626667.2014.967694.

Osuch et al., 2015 M. Osuch R.J. Romanowicz M.J. Booij The influence of parametric uncertainty on the relationships between HBV model parameters and climatic characteristics Hydrological Sciences Journal 60 7–8 2015 1299 1316 10.1080/02626667.2014.967694

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., and 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), 290-306, doi: https://doi.org/10.1016/j.jhydrol.2004.08.026.

Oudin et al., 2005 L. Oudin F. Hervieu C. Michel C. Perrin V. Andréassian F. Anctil C. Loumagne 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 2005 290 306 10.1016/j.jhydrol.2004.08.026

Pathiraja, S., Marshall, L., Sharma, A., and Moradkhani, H. (2016), Detecting non-stationary hydrologic model parameters in a paired catchment system using data assimilation, *Advances in Water Resources*, *94*, 103-119, doi: http://dx.doi.org/10.1016/j.advwatres.2016.04.021.

Pathiraja et al., 2016 S. Pathiraja L. Marshall A. Sharma H. Moradkhani Detecting non-stationary hydrologic model parameters in a paired catchment system using data assimilation Advances in Water Resources 94 2016 103 119 10.1016/j.advwatres.2016.04.021

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

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

Perrin, C., Oudin, L., Andreassian, V., Rojas-Serna, C., Michel, C., and Mathevet, T. (2007), Impact of limited streamflow data on the efficiency and the parameters of rainfall—runoff models, *Hydrological Sciences Journal*, *52*(1), 131-151, doi: 10.1623/hysj.52.1.131.

Perrin et al., 2007 C. Perrin L. Oudin V. Andreassian C. Rojas-Serna C. Michel T. Mathevet Impact of limited streamflow data on the efficiency and the parameters of rainfall—runoff models Hydrological Sciences Journal 52 1 2007 131 151 10.1623/hysj.52.1.131

Raupach, M., Briggs, P., Haverd, V., King, E., Paget, M., and Trudinger, C. (2009), Australian Water Availability Project (AWAP): CSIRO Marine and Atmospheric Research Component: Final Report for Phase 3. CAWCR Technical Report, *Rep.*, 67 pp.

Raupach et al., 2009 M. Raupach P. Briggs V. Haverd E. King M. Paget C. Trudinger Australian Water Availability Project (AWAP): CSIRO Marine and Atmospheric Research Component: Final Report for Phase 3 2009 CAWCR Technical Report Rep. 67

Raupach, M., Briggs, P., Haverd, V., King, E., Paget, M., and Trudinger, C. (2012), Australian Water Availability Project, edited by C. M. a. A. Research, Canberra, Australia.

Raupach et al., 2012 Raupach, M., Briggs, P., Haverd, V., King, E., Paget, M., and Trudinger, C. (2012), Australian Water Availability Project, edited by C. M. a. A. Research, Canberra, Australia.

Saft, M., Western, A. W., Zhang, L., Peel, M. C., and Potter, N. J. (2015), The influence of multiyear drought on the annual rainfall-runoff relationship: An Australian perspective, *Water Resources Research*, *51*(4), 2444-2463, doi: 10.1002/2014WR015348.

Saft et al., 2015 M. Saft A.W. Western L. Zhang M.C. Peel N.J. Potter The influence of multiyear drought on the annual rainfall-runoff relationship: An Australian perspective Water Resources Research 51 4 2015 2444 2463 10.1002/2014WR015348

Sleziak, P., Szolgay, J., Hlavčová, K., Duethmann, D., Parajka, J., and Danko, M. (2018), Factors controlling alterations in the performance of a runoff model in changing climate conditions, *Journal of Hydrology and Hydromechanics*, *66*(4), 381-392, doi: 10.2478/johh-2018-0031.

Sleziak et al., 2018 P. Sleziak J. Szolgay K. Hlavčová D. Duethmann J. Parajka M. Danko Factors controlling alterations in the performance of a runoff model in changing climate conditions Journal of Hydrology and Hydromechanics 66 4 2018 381 392 10.2478/johh-2018-0031

Smith, T., Marshall, L., and McGlynn, B. (2018), Typecasting catchments: Classification, directionality, and the pursuit of universality, *Advances in Water Resources*, *112*, 245-253, doi: https://doi.org/10.1016/j.advwatres.2017.12.020.

Smith et al., 2018 T. Smith L. Marshall B. McGlynn Typecasting catchments: Classification, directionality, and the pursuit of universality Advances in Water Resources 112 2018 245 253 10.1016/j.advwatres.2017.12.020

Sorooshian, S., Gupta, V. K., and Fulton, J. L. (1983), Evaluation of Maximum Likelihood Parameter estimation techniques for conceptual rainfall-runoff models: Influence of calibration data variability and length on model credibility, *Water Resources Research*, *19*(1), 251-259, doi: 10.1029/WR019i001p00251.

Sorooshian et al., 1983 S. Sorooshian V.K. Gupta J.L. Fulton Evaluation of Maximum Likelihood Parameter estimation techniques for conceptual rainfall-runoff models: Influence of calibration data variability and length on model credibility Water Resources Research 19 1 1983 251 259 10.1029/WR019i001p00251

Stephens, C. M., Johnson, F. M., and Marshall, L. A. (2018), Implications of future climate change for event-based hydrologic models, *Advances in Water Resources*, *119*, 95-110, doi: 10.1016/j.advwatres.2018.07.004.

Stephens et al., 2018 C.M. Stephens F.M. Johnson L.A. Marshall Implications of future climate change for event-based hydrologic models Advances in Water Resources 119 2018 95 110 10.1016/j.advwatres.2018.07.004

Team City (2015), GR4J - SRG, in *Rainfall Runoff Models SRG*, edited by D. Black, eWater.

City, 2015 Team City (2015), GR4J - SRG, in *Rainfall Runoff Models SRG*, edited by D. Black, eWater.

Tsakiris, G., and Vangelis, H. (2005), Establishing a drought index incorporating evapotranspiration, *European Water*, *9*(10), 3-11.

Tsakiris and Vangelis, 2005 G. Tsakiris H. Vangelis Establishing a drought index incorporating evapotranspiration European Water 9 10 2005 3 11

Vaze, J., Post, D. A., Chiew, F. H. S., Perraud, J. M., Viney, N. R., and 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, doi: 10.1016/j.jhydrol.2010.09.018.

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

Westra, S., Thyer, M., Leonard, M., Kavetski, D., and Lambert, M. (2014), A strategy for diagnosing and interpreting hydrological model nonstationarity, *Water Resources Research*, *50*(6), 5090-5113, doi: 10.1002/2013WR014719.

Westra et al., 2014 S. Westra M. Thyer M. Leonard D. Kavetski M. Lambert A strategy for diagnosing and interpreting hydrological model nonstationarity Water Resources Research 50 6 2014 5090 5113 10.1002/2013WR014719

Xiong, M., Liu, P., Cheng, L., Deng, C., Gui, Z., Zhang, X., and Liu, Y. (2019), Identifying time-varying hydrological model parameters to improve simulation efficiency by the ensemble Kalman filter: A joint assimilation of streamflow and actual evapotranspiration, *Journal of Hydrology*, *568*, 758-768, doi: https://doi.org/10.1016/j.jhydrol.2018.11.038.

Xiong et al., 2019 M. Xiong P. Liu L. Cheng C. Deng Z. Gui X. Zhang Y. Liu Identifying time-varying hydrological model parameters to improve simulation efficiency by the ensemble Kalman filter: A joint assimilation of streamflow and actual evapotranspiration Journal of Hydrology 568 2019 758 768 10.1016/j.jhydrol.2018.11.038

Zarch, M., Sivakumar, B., and Sharma, A. (2015), Droughts in a warming climate: A global assessment of Standardized precipitation index (SPI) and Reconnaissance drought index (RDI), *J. Hydrol.*, *526*, 183-195, doi: 10.1016/j.jhydrol.2014.09.071.

Zarch et al., 2015 M. Zarch B. Sivakumar A. Sharma Droughts in a warming climate: A global assessment of Standardized precipitation index (SPI) and Reconnaissance drought index (RDI) J. Hydrol. 526 2015 183 195 10.1016/j.jhydrol.2014.09.071

Zhou, Y., Zhang, Y., Vaze, J., Lane, P., and Xu, S. (2015), Impact of bushfire and climate variability on streamflow from forested catchments in southeast Australia, *Hydrological Sciences Journal*, *60*(7-8), 1340-1360, doi:

10.1080/02626667.2014.961923.

Zhou et al., 2015 Y. Zhou Y. Zhang J. Vaze P. Lane S. Xu Impact of bushfire and climate variability on streamflow from forested catchments in southeast Australia Hydrological Sciences Journal 60 7–8 2015 1340 1360 10.1080/02626667.2014.961923

Figure 1. Streamflow gauge locations

Figure 2. Demonstration of concept: (a) targeted calibration to produce parameters representing catchment behavior under different climate conditions and (b) ensemble representing possible catchment flow series (observed flow in black).

Figure 3. Experiments and the key questions they aim to address

Figure 4. Calibration performance of models representing 164 catchments under different climatic conditions. Note the number of outliers presented above the x-axis for each case.

Figure 5. Mean reduction in NSE (compared to calibration NSE) when (a) parameters from the given period are donated to the two other climatically similar periods, (b) the given period accepts parameters from the two other climatically similar periods, (c) parameters from the given period are donated to the three climatically different periods and (d) the given period accepts parameters from the three climatically different periods. Note the number of outliers not shown due to plot limits presented above the x-axis for each case.

Figure 6. Example of dynamically defined GR4J parameters (station 304497)

Table 1. Correlations (indicating predictive strength) between climate indicator (RDI$_n$) and optimal weights

Subset length (years)	Mean dry series weight correlation with RDI$_n$	Mean wet series weight correlation with RDI$_n$
1	-0.25	0.33
2	-0.28	0.40
3	-0.35	0.46
4*	**-0.38**	**0.48**
5**	-0.35	0.43

*Two catchments excluded here due to timing of wet/dry periods necessitating subset overlap

** 98 catchments excluded here due to timing of wet/dry periods necessitating subset overlap

Table 2. Optimized values of ω based on GR4J calibration subsets with different time periods

Subset length (years)	ω	ω calibrated with reduced catchments
1	0.48	0.55**
2	0.27	0.56**
3	0.26	0.43**
4	*0.07	0.41**
5	n/a	0.46**

*Two catchments excluded here due to timing of wet/dry periods necessitating subset overlap for 4y subsets – note that the 3y subset length case was tested with these two catchments removed and results did not change notably, indicating that their impact is minimal

**97 catchments excluded due to timing of wet/dry periods necessitating subset overlap for 5y (note that the two catchments removed in the 4y case were also removed here). One station (G0050115) for which independent 5y subsets could not be defined was already removed in a previous step due to zero flow in a relevant period.

Table 3. Values for ω calibrated with catchments grouped according to climatic, hydrologic and physical characteristics

Catchment criteria	Number of catchments in group	ω
Zero flow days > 8%	51	0.09
1% < Zero flow days < 8%	45	0.29
Zero flow days < 1%	64	0.02
Zero rain days > 45%	55	-0.07
35% < Zero rain days < 45%	48	-0.06
Zero rain days < 35%	57	0.17

Runoff ratio > 0.3	48	-0.09
0.15 < Runoff ratio < 0.3	55	0.23
Runoff ratio < 0.15	57	0.17
[flow variance]/[average flow] > 4500	50	0.31
1000 < [flow variance]/[average flow] < 4500	60	0.22
[flow variance] / [average flow] < 1000	50	-0.55
Area > 600 km^2	49	0.44
200 km^2 < Area < 600 km^2	55	0.23
Area < 200km^2	56	-0.18
Latitude > -25°	35	0.29
-30^o < Latitude < -25°	27	0.20
Latitude < -30°	98	0.05

Table 4. Performance of the weighting strategy compared to traditional calibration for different calibration subset lengths

Subset length (years)	Mean C_{2M} improvement	Median C_{2M} improvement	Catchments that improved (%)
1	0.040	0.003	51.9
2	0.055	0.015	54.3
3	0.051	0.022	60.5
4*	**0.062**	**0.027**	**66.3**
5**	0.021	0.021	56.9

*Two catchments excluded here due to timing of wet/dry periods necessitating subset overlap for 4y subsets

**97 catchments excluded due to timing of wet/dry periods necessitating subset overlap for 5y

Table 5. Validation performance under non-average climate conditions

Conditions	Number of catchments	Mean C_{2M} improvement	Median C_{2M} improvement	Catchments that improved (%)
Wet (RDI_n >	46	0.042	0.045	73.9

0.1)				
Dry (RDI$_n$ < -0.1)	45	0.076	0.046	66.7

Table 6. Validation performance with catchments grouped according to climatic, hydrologic and physical characteristics

Catchment criteria	**Mean NSE improvement**	**Median NSE improvement**	**Catchments that improved (%)**
Zero flow days > 8%	0.086	0.051	66.7
1% < Zero flow days < 8%	0.065	0.033	62.2
Zero flow days < 1%	0.044	0.015	67.2
Zero rain days > 45%	0.077	0.062	67.3
35% < Zero rain days < 45%	0.082	0.047	64.6
Zero rain days < 35%	0.024	0.015	64.9
Runoff ratio > 0.3	0.012	0.007	56.3
0.15 < Runoff ratio < 0.3	0.056	0.028	65.5
Runoff ratio < 0.15	0.108	0.074	73.7
[flow variance]/[average flow] > 4500	0.007	0.010	56.0
1000 < [flow variance]/[average flow] < 4500	0.076	0.029	66.7
[flow variance] / [average flow] < 1000	0.106	0.037	70.0
Area > 600 km^2	0.032	0.017	59.2
200 km^2 < Area < 600 km^2	0.068	0.026	67.3
Area < 200km^2	0.072	0.016	62.5
Latitude > -25°	0.012	0.013	54.3
-30º < Latitude < -25°	0.028	0.016	59.3
Latitude < -30°	0.088	0.037	72.4

Table 7. Typical [*Team* City, 2015] and expanded parameter ranges for GR4J calibration

GR4J parameter	**Parameter description**	**Typical calibration range**	**Expanded calibration range**

Commented [A34]: AUTHOR: Ref(s). 'Team City, 2015' is/are cited in the text but not provided in the reference list. Please provide it/them in the reference list or delete these citations from the text.

X1	Capacity of the production soil store (mm)	1 – 1500	1 – 4000
X2	Water exchange coefficient (mm)	-10 – 5	-80 – 40
X3	Capacity of the routing store (mm)	1 – 500	1 – 1300
X4	Time parameter for unit hydrographs (days)	0.5 – 4	0.5 – 12

Table 8. Overall static and dynamic model statistics for calibration and validation

Statistic	**Static calibration**	**Dynamic calibration**	**Static validation**	**Dynamic validation**	**Validation performance difference**
Mean C_{2M}	0.602	0.634	0.496	0.524	0.028
Median C_{2M}	0.605	0.631	0.548	0.555	0.007
C_{2M} variance	0.018	0.014	0.062	0.040	-0.022

Table 9. Static and dynamic model statistics for validation during a dry period

Statistic	**Static validation**	**Dynamic validation**	**Validation performance difference**
Mean NSE	0.300	0.403	0.103
Median NSE	0.445	0.465	0.020
NSE variance	0.126	0.067	-0.059