AGU100 ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2017WR022466

Improved Rainfall-Runoff Calibration for Drying Climate: Choice of Objective Function

Keirnan Fowler[1], Murray Peel[1], Andrew Western[1], and Lu Zhang[2]

[1]Department of Infrastructure Engineering, University of Melbourne, Melbourne, VIC, Australia, [2]Land and Water, CSIRO, Canberra, ACT, Australia

Key Points:
- "Least squares" approaches should not be used to calibrate models for a drying climate
- Sum-of-absolute-error calibration approaches tend to select more robust parameter sets
- Equally weighting each year in the calibration data tends to make calibration more robust

Supporting Information:
- Figure S1
- Data Set S2

Correspondence to:
K. Fowler,
fowler.k@unimelb.edu.au

Citation:
Fowler, K., Peel, M., Western, A., & Zhang, L. (2018). Improved rainfall-runoff calibration for drying climate: Choice of objective function. *Water Resources Research*, *54*, 3392–3408. https://doi.org/10.1029/2017WR022466

Received 22 DEC 2017
Accepted 2 APR 2018
Accepted article online 6 APR 2018
Published online 13 MAY 2018

Abstract It has been widely shown that rainfall-runoff models often provide poor and biased simulations after a change in climate, but evidence suggests existing models may be capable of better simulations if calibration strategies are improved. Common practice is to use "least squares"-type objective functions, which focus on hydrological behavior during high flows. However, simulation of a drying climate may require a more balanced consideration of other parts of the flow regime, including mid-low flows and drier years in the calibration period, as a closer analogue of future conditions. Here we systematically test eight objective functions over 86 catchments and five conceptual model structures in southern and eastern Australia. We focus on performance when evaluated over multiyear droughts. The results show significant improvements are possible compared to least squares calibration. In particular, the Refined Index of Agreement (based on sum of absolute error, not sum of squared error) and a new objective function called the Split KGE (which gives equal weight to each year in the calibration series) give significantly better split-sample results than least squares approaches. This improvement held for all five model structures, regardless of basin characteristics such as slope, vegetation, and across a range of climatic conditions (e.g., mean precipitation between 500 and 1,500 mm/yr). We recommend future studies to avoid least squares approaches (e.g., optimizing NSE or KGE with no prior transformation on streamflow) and adopt these alternative methods, wherever simulations in a drying climate are required.

Plain Language Summary Rainfall-runoff models are useful tools in water resource planning under climate change. They are commonly used to quantify the impact of changes in climatic variables, such as rainfall, on water availability for human consumption or environmental needs. Many parts of the world are projected to be substantially drier, possibly with threatened water resources. Given the importance of water, reliable tools for understanding future water availability are vital for society. However, literature would suggest that the current generation of rainfall-runoff models is not reliable when applied in changing climatic conditions, underestimating the sensitivity of runoff to a given change in precipitation. Many hydrologists have assumed deficiencies in the underlying model equations are to blame. However, this paper demonstrates significant improvement without changing model equations, by using a different "objective function." The objective function defines how the model is "tuned" to observations of river discharge, and this article identifies objective functions that tend to make model simulations more robust when applied in a drying climate. Using these objective functions can improve the accuracy and plausibility of future water availability estimates made for climate change impact studies.

1. Introduction

Rainfall-runoff models have potential to be useful tools in planning for future climate variability. They are often used when translating projected climatic shifts (e.g., in rainfall or temperature) into projected changes in water availability (e.g., Bergström et al., 2001; Bosshard et al., 2013; Chiew & McMahon, 2002; Chiew et al., 2009; Christensen et al., 2004; Faramarzi et al., 2013; Forzieri et al., 2014; Fowler et al., 2007; Hagemann et al., 2013; Vaze et al., 2011). They are particularly important in regions where future projections lie beyond the scope of historical conditions, and/or where available data captures a relatively small portion of historic variability (cf. Cook et al., 2016; Gallant et al., 2011). Applying a rainfall-runoff model in such circumstances assumes that the model remains fit for purpose despite being applied under a changed climatic regime

© 2018. American Geophysical Union. All Rights Reserved.

compared to the calibration data (cf. Clarke, 2007; Ehret et al., 2014; Vaze et al., 2010). In reality, climate change may cause shifts in dominant processes which may invalidate models parameterized on the past (Beck, 2002; Peterson et al., 2009; Saft et al., 2015). Also, forcing data may be difficult to estimate because of complex interactions, for example, between future vegetation and future PET (e.g., Curtis & Wang, 1998; Rodriguez-Iturbe et al., 1999; cf. Donohue et al., 2010; Seiller & Anctil, 2015). Thus, hydrological projections in changing climate are subject to many challenges.

Even when judged over the historic record, rainfall-runoff models often provide poor simulations with high bias when applied in changed climatic conditions. For example, Vaze et al. (2010) tested four conceptual rainfall-runoff model structures in 61 catchments in south east Australia and noted a reduction in model performance and increase in bias, when evaluated in periods when rainfall was different to the calibration period. This was particularly the case if the change was from wetter to drier (cf. Saft et al., 2016a, 2016b). Coron et al. (2012) tested three conceptual model structures over 216 Australian catchments and reported that "calibration over a wetter (drier) climate than the validation climate leads to an overestimation (underestimation) of the mean simulated runoff" (p. 1). These findings are not limited to single regions or model types, with similar results in Europe (Coron et al., 2014; Merz et al., 2011; Wilby, 2005), Africa (Refsgaard & Knudsen, 1996), and North America (Singh et al., 2011); and similar problems with physically based models (Refsgaard & Knudsen, 1996) and simple relationships between annual rainfall and runoff (Saft et al., 2015).

These problems have prompted a variety of responses from hydrologists. Li et al. (2012, p. 1239) recommend to minimize the degree of extrapolation in climate, stating that "if a hydrological model is set up to simulate runoff for a wet climate scenario then it should be calibrated on a wet segment of the historic record, and similarly a dry segment should be used for a dry climate scenario" (see also Broderick et al., 2016). Similarly, authors such as Singh et al. (2011) and Vaze et al. (2010) recommended limits of acceptable change in climatic conditions within which models are more likely to provide acceptable results. Other authors have investigated patterns between parameter values and climatic conditions (Brigode et al., 2013; de Vos et al., 2010; Merz et al., 2011; Wilby, 2005), such as Merz et al. (2011) who conducted separate calibrations of the HBV model on multiple 5 year segments of the climatic record. They noted that the parameters varied systematically with climate, particularly those governing snow melt and the nonlinearity of runoff generation. Many researchers affirm the need for model improvements to better simulate runoff under changing climatic conditions (e.g., Coron et al., 2014; Merz et al., 2011; Petheram et al., 2011) and some studies have produced new model structures in this vein (e.g., Hughes et al., 2013; Ramchurn, 2012).

A complementary approach is to focus on improving calibration methods (Brigode et al., 2013; Thirel et al., 2015). Common practice is to optimize models to "least squares" metrics such as the Nash Sutcliffe Efficiency (NSE; Nash & Sutcliffe, 1970) or close variants like the Kling Gupta Efficiency (KGE; Gupta et al., 2009). The literature consensus that rainfall-runoff models are unreliable in changing climates is largely based on least squares methods or similar, sometimes applied with a penalty for model bias (e.g., Coron et al., 2014; Li et al., 2012; Saft et al., 2016a; Silberstein et al., 2013; Vaze et al., 2010). However, recent evidence suggests that least squares objective functions often do not choose parameter sets that are robust to changes in climate, even when such parameter sets are available within a model structure (Fowler et al., 2016). Least squares methods tend to choose parameter sets that match hydrological behaviors during high-flow periods (Freer et al., 1996; Gan et al., 1997; Krause & Boyle, 2005; Legates & McCabe, 1999). This causes sensitivity to data errors during times of high flow (Berthet et al., 2010), potentially causing selection of parameter sets that do not simulate natural processes to the best ability of the model. Even if data errors are negligible, models may be unable to match all aspects of the flow regime with the same parameter set (the "smooth" trade-offs of Gharari et al. [2013]; cf. Efstratiadis & Koutsoyiannis, 2010). Preparing such a model for a drier climate than the calibration period may require more emphasis on hydrologic behaviors at times of mid-to-low flow, or during years that are drier than average within the calibration period, since these are the closest available analogue of the future climate (Li et al., 2012), and thus may contain the most relevant information.

The literature contains many examples of alternative calibration methods, some of which could be helpful in changing climatic conditions. Applying transformations prior to least squares-type calculations may stabilize error variance and thus improve parameter inference (Engeland et al., 2005). As already mentioned, transformations may also place more emphasis on mid-to-low flow (Chiew et al., 1993, 1995; de Vos et al., 2010; Freer et al., 1996; Pushpalatha et al., 2012). Some authors assert a more balanced consideration of

"average" model performance can be attained through absolute-error, rather than squared-error, approaches (Willmott, 1982; Willmott & Matsuura, 2005). Hartmann and Bárdossy (2005) and Shamir et al. (2005) demonstrated that robustness could be improved by evaluating catchment response on a variety of time scales (e.g., annual) rather than using objective functions formulated on the daily time step only. For example, this might cause calibration to consider whether year-to-year variability is matched, in addition to day-to-day variability. Bárdossy and Singh (2008) investigated the utility of data depth (Tukey, 1975), a geometric property among parameter sets in an ensemble. Parameter sets with high depth were relatively less sensitive to data errors and more transferable to different time periods. Zhang et al. (2008) among others demonstrate the use of meta-objective functions which consider various aspects of the flow regime separately, then combine the results together into a single "meta" objective function, and this extra information may improve parameter inference. Gharari et al. (2013) extend this logic to include multiple subperiods within the calibration period, using a Pareto-based approach to search for parameter sets that provided the best overall compromise over all periods and objectives considered (see also "limits of acceptability" approaches for similar logic applied on a time step-by-time step basis: Beven [2006] and Liu et al. [2009]). In split-sample testing, the above methods generally outperformed least squares-based calibration methods.

However, few calibration methods have been systematically tested in regions of relatively high interannual variability in historic climate, nor with data from periods of sustained (>5 yr) change in hydroclimatic conditions. This is the motivation for the present study. We limit the scope to single-objective optimization, examining objective functions that, a priori based on existing literature, are each expected to improve simulations (relative to least squares methods) when models are applied in conditions that are drier than the calibration period. The aim is to systematically test these objective functions on a large set of catchments located in a region with high interannual variability associated with historic instances of persistent dry conditions. The results of such tests can help to guide future selection of calibration methods for studies examining rainfall-runoff model capabilities in changing climate and/or applying such models in climate change impact assessments.

2. Methods

The methods section is structured as follows. Section 2.1 describes the selection of objective functions for testing. Section 2.2 outlines the study catchments, and section 2.3 outlines the testing scheme based on split-sample testing. Sections 2.4 and 2.5 outline the rainfall-runoff model structures and the algorithm used to calibrate them, respectively, and section 2.6 outlines data sources.

2.1. Objective Function Selection

Based on the literature review in the previous section, numerous classes of objective function are expected to provide improved model calibration in changing climate: (i) greater sensitivity to model bias through a bias penalty; (ii) application of transforms to runoff values prior to least squares calculations; (iii) absolute-error approaches; (iv) meta-objective functions that consider different aspects of the flow regime and then combine these into a single-objective function; and (v) time-based meta-objective functions explicitly considering different subperiods of the calibration period.

This study tests at least one representative from each class, and for each class Table 1 summarizes why improvement may be expected over least squares methods, in the context of a drying climate. In general, the methods work by broadening the focus of the calibration so that there is less focus on high flows and a greater focus on other aspects of the flow regime, including mid-low flows and/or dry years. The exception is the bias penalty class, which simply corrects the tendency of least squares measures to ignore high bias in cases of high flow variability (Gupta et al., 2009). Note that equations for each objective function are provided in the supporting information, Table S2.

All objective functions are selected from previous studies, with one exception, the *Split KGE*, which is new for this study. The *Split KGE* is based on the principle of selecting parameter sets that are consistent in time (Gharari et al., 2013). Standard calibration may not select such parameter sets because of undue focus on some subperiods in the calibration data, such as wet years, at the expense of others. Gharari et al. (2013) avoided this problem by explicitly considering model performance in separate subperiods, and "limits of acceptability" approaches take this concept further by considering each time step individually (Beven, 2006;

Table 1
Objective Functions Tested in This Study and Reasons Why Improvement Is Expected

Class of obj. funct.	Description	Reason to expect improvement (relative to least squares, in drying climate)	Objective functions tested in this article
Least squares (common practice)	The error on each time step is squared, and the objective is to minimize the sum of squares for all time steps.	n/a	KGE (Kling Gupta Efficiency; Gupta et al., 2009)[a] NSE (Nash Sutcliffe Efficiency; Nash & Sutcliffe, 1970)
(i) Bias penalty on least squares	Similar to least squares except bias is penalized, sometimes using a nonlinear function.	Counters the tendency of least squares measures to ignore high bias in cases of high flow variability (Gupta et al., 2009).	NSE-bias (NSE with bias penalty; Viney et al., 2009)[b]
(ii) Use of transforms	Similar to the above, except prior to least squares calculations, simulated and observed runoff values are raised by an exponent (usually <1) or logged.	Smaller exponents increase emphasis on times of mid and low flow, which may contain information that is relevant to a drier climate. Transformations may stabilize the variance of errors, improving parameter inference (Engeland et al., 2005) and reducing sensitivity to gauging error during floods (Berthet et al., 2010).	NSE-sqrt (NSE on the square root of flows; Chiew et al., 1995) NSE-5th root (NSE on fifth root of flows, used for emphasis on low flow by Chiew et al. [1993])[c]
(iii) Absolute error approaches	The objective is to minimize the sum of absolute errors (not squared errors)	Not squaring the errors increases emphasis on times of mid- and low-flow, which may contain information that is relevant to a drier climate, and makes calibration less sensitive to gauging errors during floods (Berthet et al., 2010).	Index of Agreement (Refined Index of Agreement; Willmott et al., 2012)
(iv) Meta-objective functions	Multiple measures are calculated separately, each considering different aspects of the flow regime. Then they are combined together into a meta function.	Each measure considers different information in the calibration data. Consideration of a wider spread of information may improve process representation and parameter inference.	Zhang (Combined objective function from Zhang et al. [2008]). Equally weighted combination of metrics regarding high flows, low flows, timing, and bias.
(v) Time-based meta-objective functions	Multiple measures, each considering different subperiods of the calibration period, are calculated separately, then combined together into a meta function.	Ensures some attention is given to dry years in the calibration period, which may be a more suitable analogue for drier climate. Such years may be largely ignored by least squares. Consideration of a wider spread of information may improve process representation and parameter inference.	Split KGE (new for this study, but cf. Gharari et al. [2013], Beven [2006], and Liu et al. [2009]). KGE is calculated separately for each year in the calibration period. The Split KGE is the average of the yearly values.

Note. Equations for each objective function are provided in the supporting information Table S2.
[a]Technically this is not a least squares metric, but the formulation is very similar as per Gupta et al. (2009). [b]Many bias penalty functions exist; we adopt the formulation of Viney et al. (2009), namely NSE-bias = NSE – 5 $|\ln(1 + \text{bias})|^{2.5}$. [c]Similar to NSE of logged values, NSE fifth root emphasizes low flows, while avoiding taking the log of zero on cease-to-flow days.

Liu et al., 2009). Although it is not possible to conduct Pareto-based or limits of acceptability methods in the context of single-objective optimization, the *Split KGE* uses similar logic by splitting the calibration period into subperiods of 1 year duration, calculating a global performance measure (the KGE) for each subperiod, and taking the average value over all subperiods as the meta-objective function value.

The list includes objective functions used by many studies contributing to the literature consensus that rainfall-runoff models are unreliable in changing climates. For example, *NSE-bias* was adopted by Vaze et al. (2010), Silberstein et al. (2013), and Saft et al. (2016a); *NSE* by Wilby (2005) and Li et al. (2012); and *KGE* by Coron et al. (2014). Thus, results here can be directly related to these studies.

In order to compare "like-with-like," results are reported according to a common "reference" metric regardless of which objective function is used. The reference metric used in the body of this article is the KGE. Our initial preference would have been the NSE as it is widely used and thus easily interpretable by most readers; also, its components (mean, variability, correlation/timing: Gupta et al. [2009]) are important attributes in the water resources context, which is often the focus for climate change impact assessments. However, as mentioned, the NSE sometimes has a high score despite significant bias, due to interaction among components (Gupta et al., 2009). Thus, we adopt the KGE, a metric with the same components but free of

unhelpful interactions. Given the importance of assessing simulations via a range of metrics, in the supporting information we also evaluate the results according to metrics proposed by Thirel et al. (2015): NSE, bias, linear correlation, relative variability, and $NSE_{low\ flow}$ (Pushpalatha et al., 2012).

2.2. Study Catchments

This study is conducted in 86 catchments from southern and eastern Australia (Figure 1; cf. Fowler et al., 2016). The study catchments are from a range of temperate climates within a region that has relatively high hydroclimatic variability on an annual scale (Peel et al., 2004a). This region has experienced persistent droughts during recorded history, and is projected to get hotter and drier in the future (Chiew et al., 2009; Trenberth, 2011; Whetton et al., 2016). Thus, it provides an excellent case study of the need for hydrologic models that can operate reliably in changing climatic conditions. In the south east of Australia, a key historic event is the Millennium Drought (1997–2010) which effected large areas both coastal and inland (Potter et al., 2010; van Dijk et al., 2013; Verdon-Kidd & Kiem, 2009). Judged by runoff reductions, this drought was severe, with the return period estimated as 300 years by Potter et al. (2010) and 1,500 years by Gallant et al. (2011). In many catchments, the Millennium Drought caused runoff reductions of greater than 80% for periods of up to 13 years (cf. van Dijk et al., 2013; Figure 1d). These reductions had significant impacts on Australian society, including cessation of irrigation in some areas causing changes in rural communities, revision of water allocation arrangements to include water trading and provision for environmental flows, and installation of alternative sources such as desalination in the cities of Melbourne, Sydney, and Brisbane (Aghakouchak et al., 2014; van Dijk et al., 2013).

A small number of catchments are included from the south west of Australia. In this region, rainfall, streamflow, and groundwater stores have been in decline since the 1970s (Hughes et al., 2012; Petrone et al., 2010). Rain-bearing synoptic troughs have decreased in frequency, while stable high-pressure systems have become more common, consistent with the effects of climate change (Hope et al., 2006). In response, streamflow has declined by up to 75 percent (Petrone et al., 2010) and catchment average groundwater levels have dropped significantly (Hughes et al., 2012), with implications for municipal and irrigation supply (Yesertener, 2005).

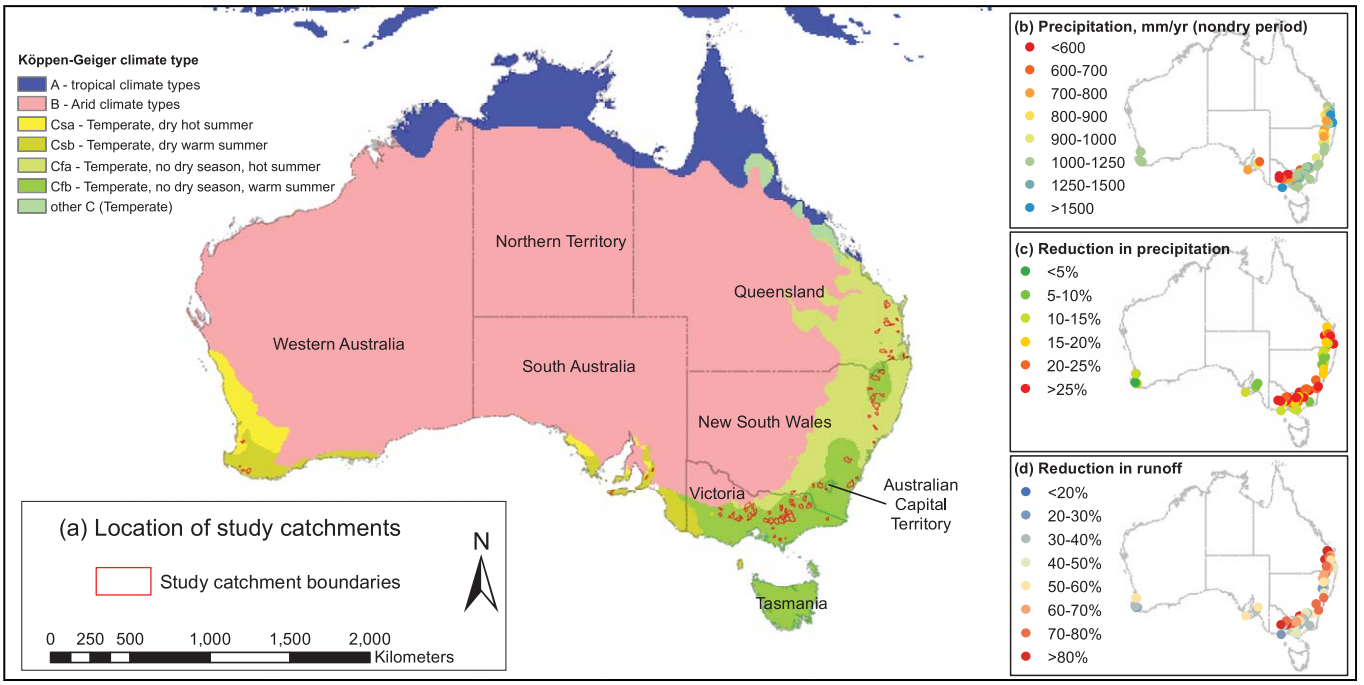

Figure 1. (a) Map of study catchments and Köppen-Geiger climate types in Australia (after Peel et al., 2007). (b) Mean annual precipitation over the nondry period, for all study catchments. (c) Reduction in precipitation in the dry (evaluation) period compared to the nondry (calibration) period. As per the text, the dry (evaluation) period is the driest run of 7 years in the historic record, and the nondry (calibration) period is the remainder of the historic record. (d) As with Figure 1c, but for runoff.

The hydroclimate of southern and eastern Australia is generally mild (Jones et al., 2009). Except in isolated mountain pockets in the south east, snow is rare as temperatures rarely drop below freezing. Average daily maximum temperatures during summer months are generally 30°C or less, and average annual rainfall is mostly between 600 and 1,500 mm/yr. Some parts of the study area are subject to dry summers, particularly in the south west of Australia and south east South Australia (Figure 1; Peel et al., 2007). However, in the south east and east, precipitation is more evenly spread year round.

The 86 study catchments are from a wider set of "Hydrologic Reference Stations" (Turner, 2012) defined by Australia's Bureau of Meteorology as a set of catchments "with minimal water resource development and land use disturbances" (p. 6) such as regulation from large reservoirs and broad-scale land use changes. The selection of catchments was subject to a variety of data quality checks, as described by Fowler et al. (2016). The study catchments vary in size from 4.4 to 1,106 km^2 (median ~200 km^2). Forest cover is generally high, with tree cover exceeding 90% in over half of the catchments. Catchment elevation ranges from sea level to 2,000 m, although most catchments do not exceed 1,500 m.

2.3. Testing Scheme

Objective functions are tested using the Differential Split-Sample Test (DSST). As outlined by Klemeš (1986), this involves evaluating model performance over an independent period with conditions that are different to the calibration data. This independent period is often called a "validation" period; in this article the term "evaluation period" is used (cf. Oreskes et al., 1994). Periods are defined based on climatic conditions, using the same period definition as Fowler et al. (2016), as follows. The seven driest consecutive years on record are used for model evaluation and are referred to as the "dry (evaluation) period." This period is defined individually for each catchment based on streamflow. The remainder of the available time series is the calibration period, referred to as the "nondry (calibration) period." In addition, a warm-up period of two years is used, chosen for each catchment separately as the two years immediately prior to the start of streamflow gauging.

As mentioned in Fowler et al. (2016), given that the Millennium Drought duration is considered to be 1997–2009 (Chiew et al., 2014), we considered adopting a length of 13 years, or alternatively a round figure such as 10 years. However, in some places, the drought was punctuated by an average or wet year midway through an otherwise dry spell (e.g., the year 2000 in the state of Victoria). Although such a sequence does not invalidate the drought as a whole from being used as an analogue for climate change (i.e., even areas subject to pronounced future drying may have occasional wet years), it was felt that such a year could dominate the calculation of performance metrics across the evaluation period as a whole. To avoid this, the adopted evaluation period duration is seven continuous years. The limits of using historic drought events as analogues of future change are discussed further in section 4.4. Figure 1d shows that reductions in flow for the dry (evaluation) period, compared to the nondry (calibration) period, exceed 80% in some cases (minimum: 24%; median: 56%; maximum 99.8%).

2.4. Rainfall-Runoff Model Structures

The same set of rainfall-runoff model structures are used as Fowler et al. (2016). The intention is to test a variety of model structures chosen to reflect common usage in the study area (particularly for water resource studies) and breadth of design of conceptual rainfall-runoff models, leading to selection of five model structures: GR4J, SIMHYD, IHACRES, GR4JMOD and SACRAMENTO. Table 2 provides references and other details for these model structures, and supporting information Table S1 provides the parameter ranges. GR4JMOD is an eight-parameter variant of GR4J, with changes intended to facilitate simulation under changing climates (Hughes et al., 2013). Note that Hughes et al.'s (2013) module to account for changes in Leaf Area Index is not adopted here. The modeling framework is implemented in a hybrid Matlab-Fortran system where the rainfall-runoff models are run in Fortran 90 (using the code of the original authors where available—Table 2).

2.5. Optimization Algorithm

Optimization is undertaken using the evolutionary algorithm CMA-ES (Hansen et al., 2003). This algorithm compares favorably with commonly used optimizers (Arsenault et al., 2014) and has been used across various fields (Hansen, 2006) including hydrology (Fowler et al., 2016; Peterson & Western, 2014). The output of CMA-ES is a single parameter set purported to correspond to the highest possible value of the objective

Table 2
Details of the Conceptual Rainfall-Runoff Model Structures Tested in This Study (After Fowler et al., 2016)

Name	Original authors	Number of free parameters[a]	Source of Fortran code
GR4J	Perrin et al. (2003)	4	Checked against code provided by authors
SIMHYD	Chiew et al. (2002)	7	Provided by authors
IHACRES	Jakeman and Hornberger (1993); Ye et al. (1997)	8	Based on original papers and Andrews (2013)
GR4JMOD	Hughes et al. (2013)	8	GR4J (see above), with changes implemented based on Hughes et al.'s (2013) paper
SACRAMENTO	Burnash et al. (1973)	16	Website of the National Oceanic and Atmospheric Administration (NOAA)[b]

[a]Note that IHACRES parameter PET_{ref} was set to zero. [b]http://www.nws.noaa.gov/iao/sacsma/fland1.f, accessed 30/03/2015.

function. For more information on the reliability of this algorithm in hydrology, refer to the comparisons of Arsenault et al. (2014) and Fowler et al. (2016, Figure S4). The software version used is CMA-ES v3.60, sourced from www.lri.fr/~hansen/cmaesintro.html (accessed 20 May 2015).

For each of the eight objective functions separately, the CMA-ES optimizer is applied in each of the 86 study catchments, for each of the five model structures, giving a total of 3440 CMA-ES runs. Consistency of results is checked using the same method as in Fowler et al. (2016); namely, by running CMA-ES three separate times and if the optimum objective function value is not the same within 1%, the number of restarts (the only user-defined parameter in CMA-ES) is increased by one (starting from zero restarts). The process is then repeated until consistency is achieved.

2.6. Data Sources

A lumped modeling approach is adopted, with the two main inputs being rainfall and potential evapotranspiration (PET), each derived as a time series on a daily time step. Rainfall is derived from the observation-based gridded product of Jones et al. (2009; www.bom.gov.au/jsp/awap/) and PET is from the wet environment areal evapotranspiration of Morton (1983), from the gridded estimates of Jeffrey et al. (2001; www.longpaddock.qld.gov.au/silo/). Catchment boundaries are derived from Shuttle Radar Topography Mission (SRTM) data, using the version by Gallant et al. (2011) on a grid size of 1 s (~30 m). ESRI's ArcHydro toolbox is used to define flow pathways, using the D8 method. Catchment boundaries are available from the lead author upon request. Streamflow data for the Hydrologic Reference Stations are publically available from www.bom.gov.au/hrs (accessed 2 January 2014). Quality codes were inspected and periods of relatively low quality are excluded from the analysis. Since quality code systems are different for each state of Australia, the details of this checking depended on location.

3. Results

Figure 2 shows differential split-sample test results for each objective function. As mentioned, to compare "like-with-like," all values plotted are KGE values, regardless of which objective function is being tested. Boxplots show distributions of values across the 86 study catchments, with evaluation (dry) period results shown in bold boxplots, and nondry (calibration) period results in faded boxplots. Equivalent plots using different reference metrics (NSE, bias, Linear Correlation, variability and $NSE_{low\ flow}$) are given in the supporting information, Figures S1–S5. The parameter values for the optimal solutions are also included in the supporting information, along with a table listing the mean, median, and percentage of negative values for each boxplot in Figure 2.

The top panel shows results when KGE is used as the objective function. As expected, the calibration KGEs (faded boxplots) are uniformly high but the evaluation KGEs (bold boxplots) are varied and, judged across the whole sample, quite poor. Likewise, evaluation scores for NSE and NSE-bias are poor (similar to KGE

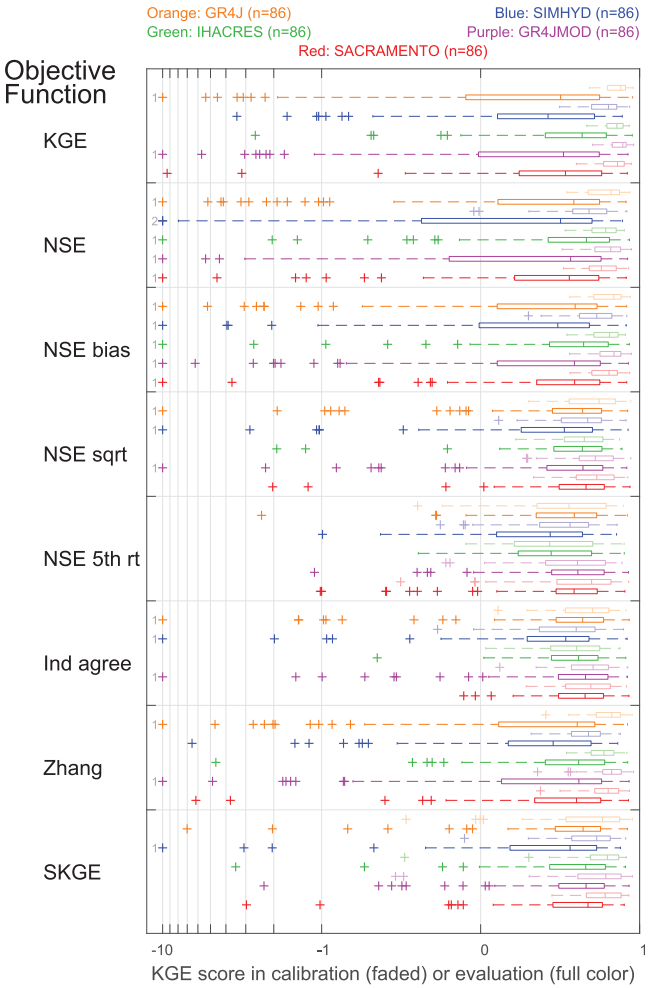

Figure 2. Boxplots of split-sample results over all catchments (n = 86) for different objective functions. To compare "like-with-like," all values plotted are KGE values. Bold colors denote $KGE_{dry\ (evaluation)\ period}$ when the model was parameterized by optimizing the objective function over the nondry (calibration) period. Faded colors denote the $KGE_{nondry\ (calibration)\ period}$ for the same parameterization. Boxplot color denotes model structure. The whiskers extend a maximum of 1.5 times the interquartile range. Values beyond the whiskers are marked as outliers and are denoted as +. Boxplots containing outliers <−10 are marked with a cross at −10 accompanied by a gray number indicating how many.

results), while calibration scores are not as high for NSE and NSE-bias as they are for KGE (as expected, since we have optimized to a different metric to the one displayed on the *x* axis). NSE-bias shows only limited improvements over NSE, for example, for the SIMHYD and GR4JMOD structures. These three objective functions account for the majority of large-sample DSST studies (e.g., Coron et al., 2014; Li et al., 2012; Saft et al., 2016a; Silberstein et al., 2013; Vaze et al., 2010), and the poor results here are consistent with these studies.

However, other objective functions show significantly improved results. In particular, NSE-sqrt and Index of Agreement each provide significantly higher KGE scores in evaluation. Note that, although similar, these two objective functions are not identical; differences arise because minimizing the difference between absolutes is not the same as minimizing the sum of squares for values subject to prior transform by square root (i.e., $|a-b| \neq \left(\sqrt{a}-\sqrt{b}\right)^2$ for most a and b). For the Refined Index of Agreement, the mean (median) $KGE_{dry\ (evaluation)\ period}$ across all 430 case studies is 0.519 (0.635), an improvement from 0.323 (0.531) when KGE is used as objective function. Further, the number of instances of $KGE_{dry\ (evaluation)\ period} < 0$ is reduced from 20.2% (KGE as objective function) to 7.0% (Refined Index of Agreement as objective function). Supporting information Figures S1–S5 indicate that the Index of Agreement and $NSE_{square\ root}$ perform well across a variety of evaluation metrics, with less biased simulations for all model structures except IHACRES, and better low flow replication for all model structures except GR4J (relative to using KGE as objective function). A comparison of simulation bias (supporting information Figure S2) indicates that the Refined Index of Agreement provides slightly less biased simulations in evaluation than NSE-sqrt.

Another well-performing objective function is the Split KGE. Recall that this is the same as the KGE except that no year can have more influence than any other year (cf. Table 1). This relatively simple change significantly improves the split-sample results: the mean (median) $KGE_{dry\ (evaluation)\ period}$ across all 430 case studies is 0.516 (0.650), an improvement from 0.323 (0.531) when KGE is used as objective function. Further, the number of instances of $KGE_{dry\ (evaluation)\ period} < 0$ is reduced from 20.2% (KGE as objective function) to 8.4% (Split KGE as objective function). As with the Refined Index of Agreement, the Split KGE generally performs well against other evaluation metrics (supporting information Figures S1–S5), providing simulations of relatively low bias and the closest replication of flow variability in evaluation of any of the metrics tested.

Figure 3 focusses on the two best performing objective functions (Refined Index of Agreement and Split KGE), comparing these with KGE as objective function. The scatter plots show that improvement is not uniform across all catchments, and for cases where KGE performs well as an objective function (*x* axis > 0.5) there is little advantage in swapping to the alternate objective functions. In contrast, improvements are significant in cases where using KGE as objective function gives poor evaluation performance (*x* axis < 0.5 in Figure 3). This is true across all model structures (Figures 3a-i and 3b-i).

Figure 3 also shows the spread of results for various catchment characteristics (smaller plots). For this set of catchments, high rainfall (long-term average > 1,500 mm/yr) locations show negligible benefit from the alternative objective functions; using KGE as objective function generally leads to good results in these catchments (Figures 3a-ii and 3b-ii). In contrast, catchments less than 1,500 mm/yr show significantly improved results from using Index of Agreement or Split KGE compared to KGE. For this catchment sample, using Index of Agreement or Split KGE as objective function provides improved split-sample results regardless of how steep the catchment is (Figures 3a-iii and 3b-iii) and regardless of how forested (Figures 3a-iv and 3b-iv). We also compare results based on how severe the drought was (Figures 3a-v and 3b-v), as measured by flow

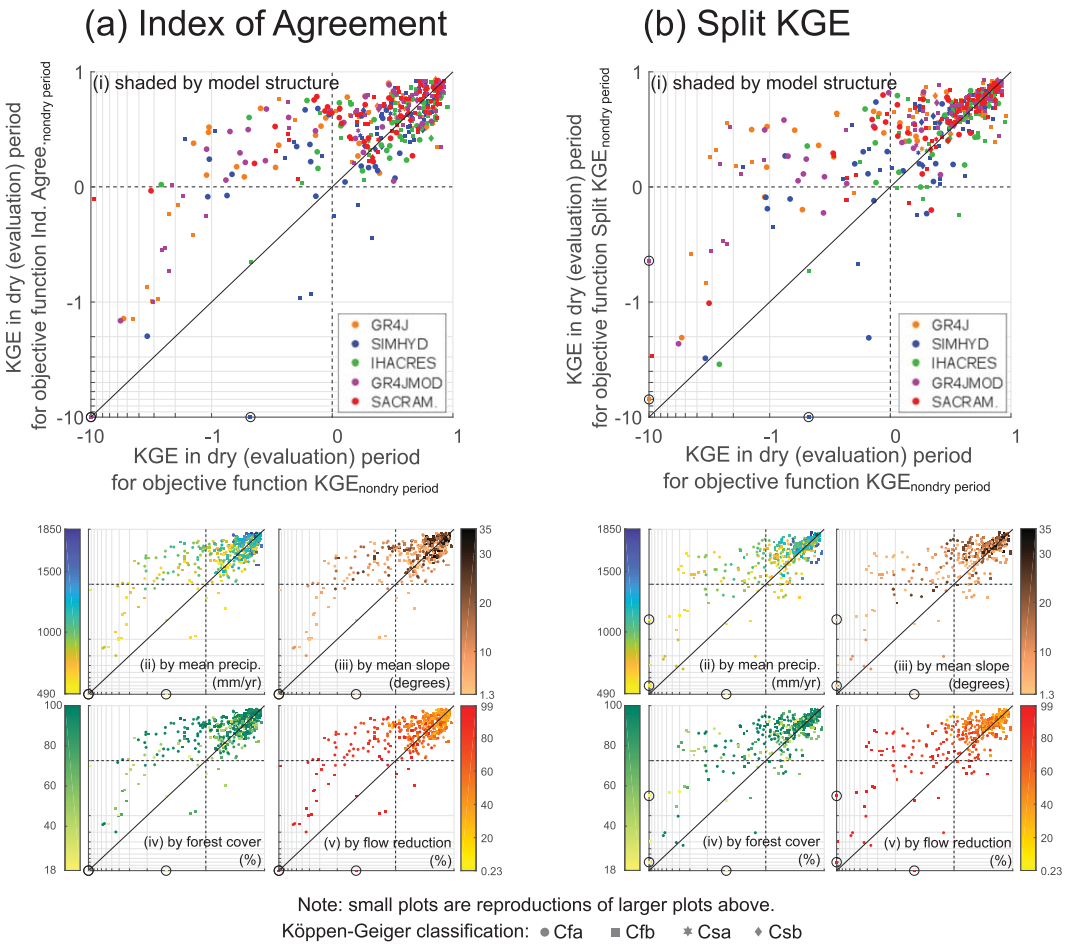

Figure 3. Comparison between objective functions for all 86 catchments with results for all five model structures shown together, thus n = 430. The smaller plots are reproductions of the larger plot above. Both *x* and *y* ordinates are KGE values in the dry (evaluation) period, but for different calibration objective functions. (a) Objective functions: Index of Agreement$_{nondry}$ (*y* axis) versus KGE$_{nondry}$ (*x* axis). Colors differentiate results by (i) model structure, (ii–iv) catchment characteristics, and (v) drought severity (flow reduction in dry period cf. nondry period). (b) Same as Figure 3a but with Split KGE instead of Index of Agreement. KGE values <−10 are shifted to −10 and marked with a circle. Color bar ranges indicate the minimum and maximum of the sample.

reductions in the dry (evaluation) period relative to the nondry (calibration) period. Note that the greatest reductions in flow (as a percentage) tended to occur in the driest catchments. In general, swapping the objective function from KGE to either Index of Agreement or Split KGE provides improved split-sample results regardless of drought severity, but benefits are relatively greater in locations where the drought was more severe.

In summary, the Refined Index of Agreement and the Split KGE were the two objective functions that provided the best performance. They performed significantly better in split-sample testing than the NSE, KGE or NSE with bias penalty, regardless of model structure, catchment slope, or vegetation. For all except the highest rainfall locations tested (i.e., mean annual rainfall > 1,500 mm/yr), results indicated significant benefits in adopting these objective functions over least squares-type functions.

4. Discussion

4.1. Discussion of Successful Objective Functions

The results provide a strong case for improvement over commonly used objective functions such as the NSE or KGE. We now review the reasons underlying the success of the objective functions Refined Index of Agreement and Split KGE, with reference to selected case studies. As discussed in section 1, through the

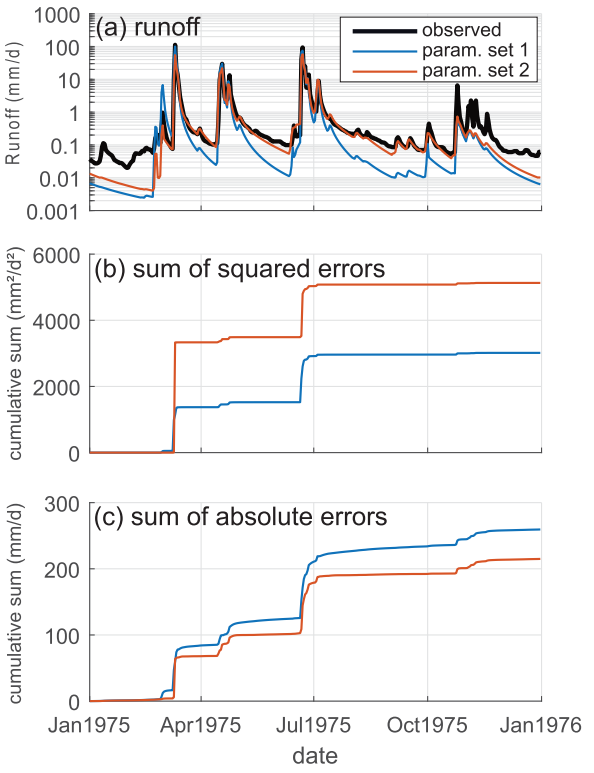

Figure 4. (a) Observed and GR4J simulated flow during part of a calibration period, for two parameter sets: parameterization by optimizing KGE_{nondry} (parameter set 1) and parameterization by optimizing Index of $Agreement_{nondry}$ (parameter set 2). The example is Currambene Creek at Falls Creek (216004, catchment area 93.5 km^2, mean annual rainfall 1,130 mm/yr, runoff ratio 0.20). (b) Cumulative evolution of the sum of squared errors over the same period. (c) Cumulative evolution of the sum of absolute errors over the same period. Sum of squared errors is very sensitive on days of high flow, but effectively ignores other days, as indicated by the near-zero gradient in between floods. The sum of absolute errors is still sensitive to floods but also considers lower flows.

squaring of errors on each time step, least squares methods cause the influence of large errors to become larger, which tends to emphasize high flow days because model error and measurement error usually increase for higher values of flow (heteroscedasticity; cf. Criss & Winston, 2008; Krause & Boyle, 2005; Legates & McCabe, 1999). Figure 4 shows graphically the consequences of this tendency. In the calculation of sum of squares over the year 1975 (Figure 4b), only the two highest flow peaks influence the calculations, making the cumulative time series appear as a sharp step function with near-zero gradient in between flood peaks, and causing the blue parameter set to be considered superior to the red parameter set. Sum of absolute errors (as in the Index of Agreement) still considers the high flow days (Figure 4c), but the step function is more rounded, and the inter-peak periods have a more significant gradient, indicating that the mid-low flows on either side of the peaks are considered with greater weight. This means that the red parameter set is considered superior, acknowledging its closer tracking of recessions and subsidiary peaks. We suggest that this more balanced consideration of the hydrograph is important for preparing a model for drier climatic conditions, as information on physical processes relevant to catchment drying may be contained in mid- and low-flow periods, in addition to high flow periods.

Next, consider the Split KGE, which (as per Table 1) is calculated by splitting the time series into individual years, calculating the KGE value for each year in isolation, and then taking the average of these annual values. Like in Figure 4, the effect is to "even out" the influence of parts of the calibration time series, but the process happens on a time scale of years (rather than days for the Refined Index of Agreement). A given wet year may have a very high influence on the standard KGE score, but with the Split KGE the influence is limited to 1/N (where N is the number of years in the calibration period). Conversely, dry years may be ignored by the standard KGE—despite potentially containing the most relevant information for a drying climate—but with the Split KGE their influence is guaranteed to be 1/N. Figure 5 shows a practical example. The observed flows in each year are provided (top) to indicate wet and dry years. The simulated flows are not directly plotted, but the daily KGE value for each year in isolation is given, for parameter set 1 (chosen when the standard KGE is the objective function) and for parameter set 3 (chosen when the Split KGE is the objective function). The color coding shows that using the standard KGE as the objective function provides good performance during most wet years but largely ignores poor performance during the dry years, particularly toward the end of the calibration period. The Split KGE, in contrast, must consider each year as equally weighted, leading to more accurate simulations of most dry years in the calibration period, and subsequently better performance in the dry (evaluation) period. The application of the Split KGE can thus be seen as a sacrifice in the (overall) KGE value during the calibration period, in exchange for better consideration of average and dry calibration years, less "overfitting" to streamflow data in high flow years, and improved evaluation performance.

It is noted that averaging of KGE values can be problematic because they have no lower bound—that is, KGE can attain very negative values, resulting in highly skewed distributions (see Mathevet et al. [2006] for a discussion of the NSE in this regard). Therefore, so long as very negative scores such as −10 or −100 exist in one or more years, Split-KGE optimization will focus on these years almost exclusively since they dominate the calculation of averages. In cases where the very low scores result from data errors and it is impossible to raise these numbers further, this could distort the calibration over the other years in the data set. To avoid these problems, an alternative and future research topic would be to apply the "split" logic to a metric with bounded formulation (e.g., that by Mathevet et al. [2006] for sum of squared errors, or that by Willmott et al. [2012] for sum of absolute errors, each of which ensure all values lie between −1 and +1; note that

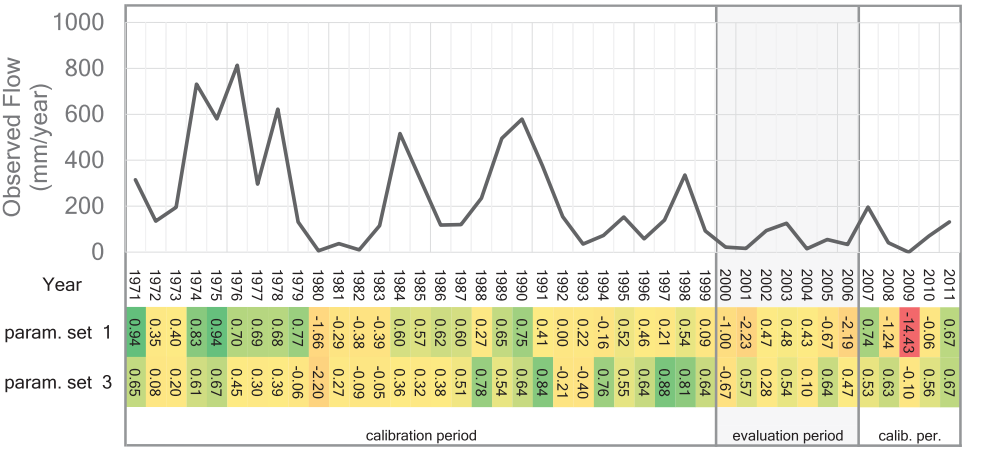

Figure 5. Annual observed flow (top) with tabulated KGE values for 41 separate calendar years (years are also shaded by KGE value, to ease interpretation), for two parameter sets: parameter set 1 is the same as Figure 4 (optimizing KGE_{nondry}); parameter set 3 is from parameterization by optimizing Split KGE_{nondry}. The case study is the same as in Figure 4 (GR4J in 216004). Optimization of standard KGE gives superior values during wet years while largely ignoring poor performance in the drier years of the calibration period, whereas the Split KGE considers all years equally. In terms of conventional KGE values, parameter set 1 attained $KGE_{nondry\ (calibration)\ period} = 0.85$ and $KGE_{dry\ (evaluation)\ period} = -0.06$, and parameter set 3 attained $KGE_{nondry\ (calibration)\ period} = 0.59$ and $KGE_{dry\ (evaluation)\ period} = 0.61$.

the latter is the Revised Index of Agreement tested in this paper). Adopting percentile values such as the median is discouraged because their insensitivity to extreme values is counter to the intention of the Split KGE (i.e., bad years should not be ignored!). For the present study, the favorable results for Split KGE suggest that these issues are not salient in the data set used.

4.2. Distinguishing Calibration Objective From Modeling Objective

Based on the above arguments, it appears logical to swap objective function from least squares to either the Index of Agreement or the Split KGE, or others that use similar logic, wherever simulations of a drying climate are required. To some readers, this might be confusing, since the objective function is often chosen to correspond closely with whatever the model is trying to achieve—for example, in a water resource study, the KGE might be an appropriate choice, because the mean, variability, and timing of flows (all KGE components) are each important to water resource modeling. Given such a purpose, most modelers would consider it logical to adopt the same metric as the objective function, under the (false) assertion that the best chance of maximizing a certain metric in the evaluation period is to optimize the same metric in the calibration period. In reality, robust simulation performance in drier or wetter conditions than the calibration data depends on fidelity of process representation. Thus, calibration objective function(s) should be chosen to extract information relevant to these processes from the calibration data. This discussion suggests that it is useful to distinguish the calibration objective function(s) from the modeling objective(s), a principle that is supported by the empirical results of this study. Put another way, if the modeling objective is to maximize KGE when evaluating a model over a drought, Figure 2 suggests that the KGE is a poor choice of objective function to achieve this goal.

4.3. Reliability of Numerical Optima

Single-objective optimization in hydrology often suffers from the problem of equifinality (e.g., Beven, 2006), where many parameter sets may achieve near-optimal scores. This is a problem, particularly if these parameter sets diverge in behavior during the evaluation period. For example, in Figure 5, only one parameter set is shown for each objective function; possibly, the second- or third-best parameter set may exhibit different behavior. Due to the large scope of this paper (in particular, the large numbers of catchments, model structures, and objective functions) it is difficult to systematically analyze robustness while maintaining brevity. However, this manuscript is part of a wider study that includes consideration of ensembles of parameter sets, not just mathematically optimal solutions, described in Fowler (2017, Chapter 5). The analysis of Fowler (2017) supports the robustness of the recommended objective functions because it demonstrated tendencies in behavior across large ensembles of parameter sets that are consistent with the results shown here.

4.4. Suitability for Climate Change Impact Studies

Even though the split-sample tests above showed significantly improved results over historic dry periods, this is no guarantee that models thus calibrated will be adequate for future climates, for various reasons. Firstly, not all models that obtain good scores in numerical metrics necessarily provide a good numerical match with evaluation data. Simulations from models may be deficient in ways not captured by the criteria, and/or the criteria may be poorly chosen and thus not reflect the modeling purpose (see previous section). A variety of checks and visualizations (Bennett et al., 2013; Thirel et al., 2015) can assist here, in addition to choosing criteria that are matched to the context as closely as possible, and/or using multiple criteria to more fully characterize the quality of simulation. Second, even a model that provides a near-exact match in simulations across all time steps may not do so for the right reasons. Different model components may combine to produce the same outcome, and it can be difficult to tell which (if any) demonstrates fidelity with dominant processes (e.g., Beven, 2006). This is particularly the case if calibration is based on streamflow only (Clark et al., 2011).

Third, even models that match historic data for the right reasons may not be adequate to simulate under projected change because the future may be so different from the past as to change the mechanisms that govern the rainfall-runoff relationship (Beck, 2002; Peterson et al., 2009; Saft et al., 2015, 2016b). The droughts used as case studies in this article, while severe, are generally within the envelope of prior historical variability (Gallant et al., 2011; Potter et al., 2010)—that is, such hydroclimatic conditions have occurred before in these locations. In contrast, climate change may cause a permanent change to conditions that have never been seen before, and thus new processes may become dominant. Changing climate is commonly thought of as a change in the mean over many years, but it likely also entails differences in the severity and duration of extreme events (e.g., Forzieri et al., 2014). Living components of the system (e.g., vegetation) may respond unexpectedly due to complex feedbacks (Curtis & Wang, 1998; Rodriguez-Iturbe et al., 1999). In some systems, changes in external forcing may cause a transition from one stable state to another (Peterson et al., 2009; cf. D'Odorico & Porporato, 2004). Thus, difficulties in characterizing future processes is a profound source of uncertainty in providing future runoff projections.

Lastly, even if all the above challenges are adequately addressed, it is important to ensure future projections are driven by appropriate climatic boundary conditions, the selection of which is subject to considerable uncertainty. A good example is the potential evapotranspiration input used in rainfall-runoff modeling (Seiller & Anctil, 2015; Guo et al., 2017). Many formulations of PET do not consider numerous factors in their formulation, such as wind, which can cause problems in representing historic behavior (Donohue et al., 2010). Approximating future boundary conditions adds additional difficulties, since PET depends on vegetative factors (e.g., albedo and canopy resistance) in addition to climatic factors, all of which are likely to change under future climate, possibly as part of complex feedbacks (Rodriguez-Iturbe et al., 1999). Thus, ignoring vegetation-climate interactions when estimating PET in a nonstationary climate is potentially a large source of bias in runoff projections.

Thus, although the DSST (i.e., the test used in this manuscript) may be "the best possible evaluation method" (Refsgaard et al., 2014), the adequacy of models that pass the DSST is far from guaranteed and the quality of future projections depends on many factors.

4.5. Limitations and Further Research

This study found that the Index of Agreement and Split KGE objective functions provided improved split-sample results, and that this did not depend strongly on catchment attributes (with the possible exception of mean precipitation, as discussed above). Given the relatively large number of catchments tested (86), this provides some confidence that the findings are generally applicable across the temperate parts of Australia. Given no nontemperate catchments were included, the applicability of findings to other climate types is untested, and we recommend that application to other climate types (or continents other than Australia) be preceded by split-sample testing to confirm the local relevance of the findings of this paper.

Furthermore, no testing of a wetting climate was undertaken. In the split-sample testing undertaken here, the evaluation period was always drier than the calibration period. It is recommended that future research uses similar methods to identify suitable objective functions for use in a wetting climate.

Even using the recommended objective functions, model performance problems persist for many catchments (Figure 1), and the cause of these problems is a potential future research topic. It is possible that

rainfall errors play a role in some catchments, since rainfall may become more localized with drying climate. However, using the same "nondry" and "dry" periods as in this study, Fowler et al. (2016) showed that robust performance is often possible with a different choice of parameter set, as demonstrated by their multiobjective approach that explicitly defined separate objectives for each period. Their results would not be possible if the errors were purely random, as would be expected if due to more localized precipitation. This may imply systematic structural errors are common, as suggested by, for example, Petheram et al. (2011, p. 3622) and Fowler (2017, p. 33). As stated, few studies provide systematic testing of calibration methods under changing climatic boundary conditions. This article has aimed to fill this gap for single-objective optimization, but it is recommended that future studies extend this to other calibration paradigms, including ensemble methods, that is, calibration methods that explicitly consider many parameter sets, rather than the single "optimum." Such methods have numerous advantages compared to single-objective optimization, including the ability to estimate uncertainty in predictions, and to more fully characterize the ability of a model structure by analyzing a broader pool of parameter sets. Ensemble methods useful in the context of changing climate could include data depth methods (Bárdossy & Singh, 2008), limits of acceptability methods (Beven, 2006; Blazkova & Beven, 2009; Liu et al., 2009), Pareto methods (Gharari et al., 2013) and Approximate Bayesian Computation (Fowler, 2017; Nott et al., 2012; Vrugt & Sadegh, 2013). In addition, methods that explicitly separate error sources (e.g., Kavetski et al., 2006a, 2006b; Renard et al., 2010) have rarely been applied under changing conditions, so that their potential for improving the robustness of predictions is untested.

5. Conclusions

Although the literature contains many calibration methods that may be relevant to rainfall-runoff modeling in changing climatic conditions, relatively few studies systematically compare these methods in regions with relatively high interannual variability in historic climate. This study has aimed to fill this gap for single-objective optimization methods. Eight objective functions were tested in 86 catchments in southern and eastern Australia, with a focus on performance when evaluated over historic droughts. Two classes of method proved particularly effective:

i. **Sum-of-absolute-error methods** such as the Refined Index of Agreement of Willmott et al. (2012). The practice of calculating the NSE on the square root of flows (termed NSE_{sqrt} in this study) is similar to a sum-of-absolute-error approach, and NSE_{sqrt} provided similar evaluation results to the Refined Index of Agreement but with slightly more bias.
ii. **Methods which weight each year in the calibration series equally** such as the Split KGE, which limits the influence of wet years in calibration data and ensures dry years are not ignored.

These results suggest that there is information in calibration data that is not fully exploited by common "least squares" calibration methods applied on untransformed streamflow. The success of the above methods was consistent across the five rainfall-runoff models tested, across 86 temperate catchments regardless of slope, degree of forestation, or severity of drought. However, the relative benefit of these methods was less for the higher rainfall catchments tested (>1,500 mm/yr). The testing was limited to the case of climate drying rather than wetting, and future research is recommended to confirm the generality of results on continents other than Australia, for other model structures and for wetting rather than drying climate.

We recommend future studies avoid "least squares" approaches (e.g., optimizing the NSE, RMSE or KGE on untransformed streamflow) and adopt these alternative methods, wherever simulations of a drying climate are required. Whereas some studies previously assumed that the poor performance of models under changing climate was due to the model structures themselves, this study demonstrated that improvements are possible without changing the model structures. This should encourage future modelers to employ a range of methods to extract information from data, rather than relying on commonly used methods of calibration.

Acknowledgments
The authors gratefully acknowledge the support of the Australian Government in carrying out this work. Specifically, Keirnan Fowler's work was supported by a Research Training Program Scholarship, and Murray Peel is the recipient of an Australian Research Council Future Fellowship (FT120100130). The assistance and patience of Tim Peterson, Lev Lafayette, and Daniel Tosello is gratefully acknowledged, as is the kind support, guidance, and hospitality extended by Thorsten Wagener, Jim Freer, and Ross Woods of the University of Bristol. The comments of Wouter Knoben and Sina Khatami, along with three anonymous reviewers and the Associate Editor, are gratefully acknowledged in shaping the manuscript. Streamflow data used in this project were from the Australian Bureau of Meteorology's (BOM) Hydrologic Reference Station project website (Turner, 2012), www.bom.gov.au/hrs. Rainfall data were from the Australian Water Availability Project (AWAP) project (Jones et al., 2009), www.bom.gov.au/jsp/awap/. Potential evapotranspiration data were from the SILO project (Jeffrey et al., 2001), www.longpaddock.qld.gov.au/silo/. The authors acknowledge the assistance of the three anonymous reviewers and the associate editor, whose feedback greatly improved the quality of the article. In addition, the feedback on early versions of this document provided by Sina Khatami and Wouter Knoben is gratefully acknowledged.

References

Aghakouchak, A., Feldman, D., Stewardson, M. J., Saphores, J. D., Grant, S., & Sanders, B. (2014). Australia's drought: Lessons for California. *Science*, *343*, 1430–1431.

Andrews, F. (2013). R code repository for HYDROMAD. Retrieved from http://hydromad.catchment.org/

Arsenault, R., Poulin, A., Cote, P., & Brissette, F. (2014). Comparison of stochastic optimization algorithms in hydrological model calibration. *Journal of Hydrologic Engineering*, *19*(7), 1374–1384. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000938

Bárdossy, A., & Singh, S. K. (2008). Robust estimation of hydrological model parameters. *Hydrology and Earth System Sciences*, *12*(6), 1273–1283. https://doi.org/10.5194/hess-12-1273-2008

Beck, M. B. (2002). *Environmental foresight and models: a manifesto*. Oxford, UK: Gulf Professional Publishing.

Bennett, N. D., Croke, B. F. W., Guariso, G., Guillaume, J. H. A., Hamilton, S. H., Jakeman, A. J., et al. (2013). Characterising performance of environmental models. *Environmental Modelling and Software*, *40*, 1–20. https://doi.org/10.1016/j.envsoft.2012.09.011

Bergström, S., Carlsson, B., Gardelin, M., Lindström, G., Petterson, A., & Rummukainen, M. (2001). Climate change impacts on runoff in Sweden—Assessments by global climate models, dynamical downscalling and hydrological modelling. *Climate Research*, *16*(2), 101–112. https://doi.org/10.3354/cr016101

Berthet, L., Andréassian, V., Perrin, C., & Loumagne, C. (2010). How significant are quadratic criteria? Part 2. On the relative contribution of large flood events to the value of a quadratic criterion. *Hydrological Sciences Journal*, *55*(6), 1063–1073. https://doi.org/10.1080/02626667.2010.505891

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, *320*(1–2), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Blazkova, S., & Beven, K. (2009). A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic. *Water Resources Research*, *45*, W00B16. https://doi.org/10.1029/2007WR006726

Bosshard, T., Carambia, M., Goergen, K., Kotlarski, S., Krahe, P., Zappa, M., & Schär, C. (2013). Quantifying uncertainty sources in an ensemble of hydrological climate-impact projections. *Water Resources Research*, *49*, 1523–1536. https://doi.org/10.1029/2011WR011533

Brigode, P., Oudin, L., & 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. https://doi.org/10.1016/j.jhydrol.2012.11.012

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

Burnash, R. J. C., Ferral, R. L., & McGuire, R. A. (1973). *A generalized streamflow simulation system—Conceptual modelling for digital computers* (Technical Report). Sacramento: U.S. Department of Commerce, National Weather Service and State of California, Department of Water Resources.

Chiew, F. H. S., & McMahon, T. A. (2002). Modelling the impacts of climate change on Australian streamflow. *Hydrological Processes*, *16*(6), 1235–1245. https://doi.org/10.1002/hyp.1059

Chiew, F. H. S., Peel, M. C., & Western, A. W. (2002). Application and testing of the simple rainfall runoff model Simhyd. In V. P. Singh & D. K. Frevert (Eds.), *Mathematical models of small watershed hydrology and applications* (pp. 335–367). Littleton, CO: Water Resources Publications.

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. *Stochastic Environmental Research and Risk Assessment*, *28*(1), 3–15. https://doi.org/10.1007/s00477-013-0755-5

Chiew, F. H. S., Stewardson, M. J., & McMahon, T. A. (1993). Comparison of six rainfall-runoff modelling approaches. *Journal of Hydrology*, *147*(1–4), 1–36. https://doi.org/10.1016/0022-1694(93)90073-I

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 Resources Research*, *45*, W10414. https://doi.org/10.1029/2008WR007338

Chiew, F. H. S., Whetton, P., McMahon, T., & Pittock, A. B. (1995). Simulation of the impacts of climate change on runoff and soil moisture in Australian catchments. *Journal of Hydrology*, *167*(1–4), 121–147. https://doi.org/10.1016/0022-1694(94)02649-V

Christensen, N. S., Wood, A. W., Voisin, N., Lettenmaier, D. P., & Palmer, R. N. (2004). The effects of climate change on the hydrology and water resources of the Colorado River Basin. *Climatic Change*, *62*(1), 337–363. https://doi.org/10.1023/B:CLIM.0000013684.13621.1f

Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, *47*(9), 1–16. https://doi.org/10.1029/2010WR009827

Clarke, R. T. (2007). Hydrological prediction in a non-stationary world. *Hydrology and Earth System Sciences*, *11*(1), 408–414. https://doi.org/10.5194/hess-11-408-2007

Cook, B. I., Anchukaitis, K. J., Touchan, R., Meko, D. M., & Cook, E. R. (2016). Spatiotemporal drought variability in the Mediterranean over the last 900 years. *Journal of Geophysical Research: Atmospheres*, *121*(5), 2060–2074. https://doi.org/10.1002/2015JD023929

Coron, L., Andréassian, V., Perrin, C., Bourqui, M., & Hendrickx, F. (2014). On the lack of robustness of hydrologic models regarding water balance simulation: A diagnostic approach applied to three models of increasing complexity on 20 mountainous catchments. *Hydrology and Earth System Sciences*, *18*(2), 727–746. https://doi.org/10.5194/hess-18-727-2014

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

Criss, R. E., & Winston, W. E. (2008). Do Nash values have value? Discussion and alternate proposals. *Hydrological Processes*, *22*(14), 2723–2725. https://doi.org/10.1002/hyp.7072

Curtis, P. S., & Wang, X. (1998). A meta-analysis of elevated CO2 effects on woody plant mass, form, and physiology. *Oecologia*, *113*(3), 299–313. https://doi.org/10.1007/s004420050381

de Vos, N. J., Rientjes, T. H. M., & Gupta, H. V. (2010). Diagnostic evaluation of conceptual rainfall-runoff models using temporal clustering. *Hydrological Processes*, *24*(20), 2840–2850. https://doi.org/10.1002/hyp.7698

D'Odorico, P., & Porporato, A. (2004). Preferential states in soil moisture and climate dynamics, *Proceedings of the National Academy of Science of the United States of America*, 101, 8848–8851. https://doi.org/10.1073/pnas.0401428101

Donohue, R. J., McVicar, T. R., & Roderick, M. L. (2010). Assessing the ability of potential evaporation formulations to capture the dynamics in evaporative demand within a changing climate. *Journal of Hydrology*, *386*(1–4), 186–197. https://doi.org/10.1016/j.jhydrol.2010.03.020

Efstratiadis, A., & Koutsoyiannis, D. (2010). One decade of multi-objective calibration approaches in hydrological modelling: A review. *Hydrological Sciences Journal*, *55*(1), 58–78. https://doi.org/10.1080/02626660903526292

Ehret, U., Gupta, H. V., Sivapalan, M., Weijs, S. V., Schymanski, S. J., Blöschl, G., et al. (2014). Advancing catchment hydrology to deal with predictions under change. *Hydrology and Earth System Sciences*, *18*(2), 649–671. https://doi.org/10.5194/hess-18-649-2014

Engeland, K., Xu, C.-Y., & Gottschalk, L. (2005). Assessing uncertainties in a conceptual water balance model using Bayesian methodology/ Estimation bayésienne des incertitudes au sein d'une modélisation conceptuelle de bilan hydrologique. *Hydrological Sciences Journal*, *50*(1), 37–41. https://doi.org/10.1623/hysj.50.1.45.56334

Faramarzi, M., Abbaspour, K. C., Ashraf Vaghefi, S., Farzaneh, M. R., Zehnder, A. J. B., Srinivasan, R., & Yang, H. (2013). Modeling impacts of climate change on freshwater availability in Africa. *Journal of Hydrology*, *480*, 85–101. https://doi.org/10.1016/j.jhydrol.2012.12.016

Forzieri, G., Feyen, L., Rojas, R., Flörke, M., Wimmer, F., & Bianchi, A. (2014). Ensemble projections of future streamflow droughts in Europe. *Hydrology and Earth System Sciences*, *18*(1), 85–108. https://doi.org/10.5194/hess-18-85-2014

Fowler, H. J., Kilsby, C. G., & Stunell, J. (2007). Modelling the impacts of projected future climate change on water resources in north-west England. *Hydrology and Earth System Sciences*, *11*(3), 1115–1126. https://doi.org/10.5194/hess-11-1115-2007

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 Resource Research*, *52*, 1820–1846. https://doi.org/10.1002/2015WR018068

Fowler, K. J. A. (2017). *Towards improved rainfall-runoff modelling under changing climatic conditions* (Ph.D. thesis). Melbourne, Australia: University of Melbourne, Department of Infrastructure Engineering. Retrieved from http://hdl.handle.net/11343/208776

Freer, J., Beven, K., & Ambroise, B. (1996). Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach. *Water Resources Research*, *32*(7), 2161–2173. https://doi.org/10.1029/96WR03723

Gallant, J., Dowling, T., Read, A. M., Wilson, N., Tickle, P., & Inskeep, C. (2011). *1 second SRTM derived products user guide* (Technical Report). Canberra, Australia: Geoscience Australia.

Gan, T. Y., Dlamini, E. M., Biftu, G. F., Thian, Y. G., Dlamini, E. M., & Biftu, G. F. (1997). Effects of model complexity and structure, data quality, and objective functions on hydrologic modeling. *Journal of Hydrology*, *192*(1–4), 81–103. https://doi.org/10.1016/S0022-1694(96)03114-9

Gharari, S., Hrachowitz, M., Fenicia, F., & Savenije, H. H. G. (2013). An approach to identify time consistent model parameters: Sub-period calibration. *Hydrology and Earth System Sciences*, *17*(1), 149–161. https://doi.org/10.5194/hess-17-149-2013

Guo, D., Westra, S., & Maier, H. R. (2017). Sensitivity of potential evapotranspiration to changes in climate variables for dierent Australian climatic zones. *Hydrology and Earth System Sciences*, *21*(4), 2107–2126. https://doi.org/10.5194/hess-21-2107-2017.209

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

Hagemann, S., Chen, C., Clark, D. B., Folwell, S., Gosling, S. N., Haddeland, I., et al. (2013). Climate change impact on available water resources obtained using multiple global climate and hydrology models. *Earth System Dynamics*, *4*(1), 129–144. https://doi.org/10.5194/esd-4-129-2013

Hansen, N. (2006). The CMA evolution strategy: A comparing review. *Studies in Fuzziness and Soft Computing*, *192*(2006), 75–102. https://doi.org/10.1007/11007937-4

Hansen, N., Müller, S. D., & Koumoutsakos, P. (2003). Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evolutionary Computation*, *11*(1), 1–18. https://doi.org/10.1162/106365603321828970

Hartmann, G., & Bárdossy, A. (2005). Investigation of the transferability of hydrological models and a method to improve model calibration. *Advances in Geosciences*, *5*, 83–87. https://doi.org/10.5194/adgeo-5-83-2005

Hope, P. K., Drosdowsky, W., & Nicholls, N. (2006). Shifts in the synoptic systems influencing southwest Western Australia. *Climate Dynamics*, *26*, 751–764. https://doi.org/10.1007/s00382-006-0115-y

Hughes, J., Silberstein, R., Grigg, A., & Mountain, B. (2013). Extending rainfall-runoff models for use in environments with long-term catchment storage and forest cover changes. In J. Piantadosi, R. S. Anderssen, & J. Boland (Eds.), *MODSIM2013, 20th International Congress on modelling and simulation* (pp. 1–6). Australia: Modelling and Simulation Society of Australia and New Zealand.

Hughes, J. D., Petrone, K. C., & Silberstein, R. P. (2012). Drought, groundwater storage and stream flow decline in southwestern Australia. *Geophysical Research Letters*, *39*, L03408. https://doi.org/10.1029/2011GL050797

Jakeman, A. J., & Hornberger, G. M. (1993). How much complexity is warranted in a rainfall-runoff model? *Water Resources Research*, *29*(8), 2637–2649. https://doi.org/10.1029/93WR00877

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. *Environmental Modelling and Software*, *16*(4), 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. *Australian Meteorological and Oceanographic Journal*, *58*(4), 233–248. https://doi.org/10.22499/2.5804.003

Kavetski, D., Kuczera, G., & Franks, S. W. (2006a). Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resources Research*, *42*(3), 1–9. https://doi.org/10.1029/2005WR004368

Kavetski, D., Kuczera, G., & Franks, S. W. (2006b). Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. *Water Resources Research*, *42*, W03408. https://doi.org/10.1029/2005WR004376

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

Krause, P., & Boyle, D. P. (2005). Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences*, *5*(89), 89–97. https://doi.org/10.5194/adgeo-5-89-2005

Legates, D. R., & McCabe, G. J. (1999). Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, *35*(1), 233–241. https://doi.org/10.1029/1998WR900018

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. *Hydrology and Earth System Sciences*, *16*(4), 1239–1254. https://doi.org/10.5194/hess-16-1239-2012

Liu, Y., Freer, J., Beven, K., & Matgen, P. (2009). Towards a limits of acceptability approach to the calibration of hydrological models: Extending observation error. *Journal of Hydrology*, *367*(1), 93–103.

Mathevet, T., Michel, C., Andréassian, V., & 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* (IAHS Red Books Ser., Vol. 307, pp. 211–219). London: IAHS.

Merz, R., Parajka, J., & Blöschl, G. (2011). Time stability of catchment model parameters: Implications for climate impact analyses. *Water Resources Research*, *47*, W02531. https://doi.org/10.1029/2010WR009505

Morton, F. I. (1983). Operational estimates of areal evapotranspiration and their significance to the science and practice of hydrology. *Journal of Hydrology*, *66*, 1–76.

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

Nott, D. J., Marshall, L., & Brown, J. (2012). Generalized likelihood uncertainty estimation (GLUE) and approximate Bayesian computation: What's the connection? *Water Resources Research*, *48*, W12602. https://doi.org/10.1029/2011WR011128

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, *263*(February), 641–646. https://doi.org/10.1126/science.263.5147.641

Peel, M. C., Finlayson, B. L., & McMahon, T. A. (2007). Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences*, *11*(5), 1633–1644. https://doi.org/10.5194/hess-11-1633-2007

Peel, M. C., McMahon, T. A., & Finlayson, B. L. (2004). Continental differences in the variability of annual runoff-update and reassessment. *Journal of Hydrology*, *295*(1–4), 185–197. https://doi.org/10.1016/j.jhydrol.2004.03.004

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

Peterson, T. J., Argent, R. M., Western, A. W., & Chiew, F. H. S. (2009). Multiple stable states in hydrological models: An ecohydrological investigation. *Water Resources Research*, *45*, W03406. https://doi.org/10.1029/2008WR006886

Peterson, T. J., & Western, A. W. (2014). Nonlinear time-series modeling of unconfined groundwater head. *Water Resources Research*, *50*, 8330–8355. https://doi.org/10.1002/2013WR014800

Petheram, C., Potter, N., Vaze, J., Chiew, F., & Zhang, L. (2011). Towards better understanding of changes in rainfall- runoff relationships during the recent drought in south-eastern Australia. Paper presented at 19th International Congress on Modelling and Simulation, Perth, Australia, December 12–16.

Petrone, K. C., Hughes, J. D., Van Niel, T. G., & Silberstein, R. P. (2010). Streamflow decline in southwestern Australia, 1950–2008. *Geophysical Research Letters*, *37*, L11401. https://doi.org/10.1029/2010GL043102

Potter, N. J., Chiew, F. H., & Frost, A. J. (2010). An assessment of the severity of recent reductions in rainfall and runoff in the Murray-Darling Basin. *Journal of Hydrology*, *381*(1–2), 52–64. https://doi.org/10.1016/j.jhydrol.2009.11.025

Pushpalatha, R., Perrin, C., Moine, N. L., & Andréassian, V. (2012). A review of efficiency criteria suitable for evaluating low-flow simulations. *Journal of Hydrology*, *420–421*, 171–182. https://doi.org/10.1016/j.jhydrol.2011.11.055

Ramchurn, A. (2012). Improved modelling of low flows and drought impacts in Australian catchments using new rainfall-runoff model SpringSIM. Paper presented at Australian Hydrology and Water Resources Symposium, 2012, Sydney, Australia, November 22.

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

Refsgaard, J. C., Madsen, H., Andréassian, V., Arnbjerg-Nielsen, K., Davidson, T. A., Drews, M., et al. (2014). A framework for testing the ability of models to project climate change and its impacts. *Climatic Change*, *122*(1–2), 271–282. https://doi.org/10.1007/s10584-013-0990-2

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., & Franks, S. W. (2010). Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research*, *46*, W05521. https://doi.org/10.1029/2009WR008328

Rodriguez-Iturbe, I., D'odorico, P., Porporato, A., & Ridolfi, L. (1999). On the spatial and temporal links between vegetation, climate, and soil moisture. *Water Resources Research*, *35*(12), 3709–3722. https://doi.org/10.1029/1999WR900255

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. *Geophysical Research Letters*, *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 Resources Research*, *52*, 9290–9305. https://doi.org/10.1002/2016WR019525

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

Seiller, G., & Anctil, F. (2015). How do potential evapotranspiration formulas influence hydrological projections? *Hydrological Sciences Journal*, *6667*(March), 151001115055,005. https://doi.org/10.1080/02626667.2015.1100302

Shamir, E., Imam, B., Gupta, H. V., & Sorooshian, S. (2005). Application of temporal streamflow descriptors in hydrologic model parameter estimation. *Water Resources Research*, *41*, W06021. https://doi.org/10.1029/2004WR003409

Silberstein, R. P., Aryal, S. K., Braccia, M., Durrant, J., & Silberstein, R. (2013). Rainfall-runoff model performance suggests a change in flow regime and possible lack of catchment resilience. In J. Piantadosi, R. S. Anderssen, and J. Boland (Eds.), *MODSIM2013, 20th International Congress on Modelling and Simulation* (pp. 1–6). Australia: Modelling and Simulation Society of Australia and New Zealand.

Singh, R., Wagener, T., Van Werkhoven, K., Mann, M. E., & Crane, R. (2011). A trading-space-for-time approach to probabilistic continuous streamflow predictions in a changing climate-accounting for changing watershed behavior. *Hydrology and Earth System Sciences*, *15*(11), 3591–3603. https://doi.org/10.5194/hess-15-3591-2011

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

Trenberth, K. E. (2011). Changes in precipitation with climate change. *Climate Research*, *47*(1/2), 123–138.

Tukey, J. (1975). Mathematics and the picturing of data. In *Proceedings of the international congress of mathematicians* (pp. 523–532). Reykjavik, Iceland: Springer.

Turner, M. (2012). *Hydrologic reference station selection guidelines* (Report). Melbourne, VIC: Bureau of Meteorology, Australia.

van Dijk, A. I. J. M., Beck, H. E., Crosbie, R. S., De Jeu, R. A. M., Liu, Y. Y., Podger, G. M., et al. (2013). The Millennium Drought in southeast Australia (2001–2009): Natural and human causes and implications for water resources, ecosystems, economy, and society. *Water Resources Research*, *49*, 1040–1057. https://doi.org/10.1002/wrcr.20123

Vaze, J., Chiew, F. H. S., Perraud, J. M., Viney, N., Post, D., Teng, J., et al. (2010). Rainfall-runoff modelling across southeast Australia: Datasets, models and results. *Australian Journal of Water Resources*, *14*(2), 101–116.

Vaze, J., Davidson, A., Teng, J., & Podger, G. (2011). Impact of climate change on water availability in the Macquarie-Castlereagh River Basin in Australia. *Hydrological Processes*, *25*(16), 2597–2612. https://doi.org/10.1002/hyp.8030

Verdon-Kidd, D. C., & Kiem, A. S. (2009). Nature and causes of protracted droughts in southeast Australia: Comparison between the Federation, WWII, and Big Dry droughts. *Geophysical Research Letters*, *36*, L22707. https://doi.org/10.1029/2009GL041067

Viney, N. R., Perraud, J. M., 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. Paper presented at 18th World IMACS/MODSIM Congress, Cairns, Australia, July 13–17. Retrieved from http://mssanz.org.au/modsim09

Vrugt, J. A., & Sadegh, M. (2013). Toward diagnostic model calibration and evaluation: Approximate Bayesian computation. *Water Resources Research*, *49*, 4335–4345. https://doi.org/10.1002/wrcr.20354

Whetton, P. H., Grose, M. R., & Hennessy, K. J. (2016). A short history of the future: Australian climate projections 1987–2015. *Climate Services*, *2–3*, 1–14. https://doi.org/10.1016/j.cliser.2016.06.001

Wilby, R. L. (2005). Uncertainty in water resource model parameters used for climate change impact assessment. *Hydrological Processes*, *19*(16), 3201–3219. https://doi.org/10.1002/hyp.5819

Willmott, C. J. (1982). Some comments on the evaluation of model performance. *Meteorological Society*, *12*(12), 1309–1313.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79–82. https://doi.org/10.3354/cr030079

Willmott, C. J., Robeson, S. M., & Matsuura, K. (2012). A refined index of model performance. *International Journal of Climatology*, *32*(13), 2088–2094. https://doi.org/10.1002/joc.2419

Ye, W., Bates, B. C., Viney, N. R., Sivapalan, M., & Jakeman, A. J. (1997). Performance of conceptual rainfall-runoff models in low yielding ephemeral catchments. *Water Resources Research*, *33*(1), 153–166.

Yesertener, C. (2005). Impacts of climate, land and water use on declining groundwater levels in the Gnangara Groundwater Mound, Perth, Australia. *Australian Journal of Water Resources*, *8*(2), 143–152.

Zhang, L., Potter, N., Hickel, K., Zhang, Y., & Shao, Q. (2008). Water balance modeling over variable time scales based on the Budyko framework—Model development and testing. *Journal of Hydrology*, *360*(1–4), 117–131. https://doi.org/10.1016/j.jhydrol.2008.07.021