University Library



A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Fowler, KJA;Peel, MC;Western, AW;Zhang, L;Peterson, TJ

Title:
Simulating runoff under changing climatic conditions: Revisiting an apparent deficiency of conceptual rainfall-runoff models

Date:
2016-03-01

Citation:
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 Resources Research, 52 (3), pp.1820-1846. https://doi.org/10.1002/2015WR018068.

Persistent Link:
https://hdl.handle.net/11343/118330

AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2015WR018068

Key Points:
- Models may be more capable under changing climatic conditions than previously thought
- Common calibration methods often fail to identify parameter sets that are robust
- Caution is needed when interpreting the results of split sample testing

Supporting Information:
- Supporting Information S1

Correspondence to:
K. J. A. Fowler,
kfowler@student.unimelb.edu.au

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

Received 4 SEP 2015
Accepted 22 JAN 2016
Accepted article online 2 FEB 2016
Published online 12 MAR 2016

© 2016. American Geophysical Union. All Rights Reserved.

Simulating runoff under changing climatic conditions: Revisiting an apparent deficiency of conceptual rainfall-runoff models

Keirnan J. A. Fowler[1], Murray C. Peel[1], Andrew W. Western[1], Lu Zhang[2], and Tim J. Peterson[1]

[1]Department of Infrastructure Engineering, University of Melbourne, Melbourne, Victoria, Australia, [2]CSIRO Land and Water Flagship, Canberra, Australian Capital Territory, Australia

Abstract Hydrologic models have potential to be useful tools in planning for future climate variability. However, recent literature suggests that the current generation of conceptual rainfall runoff models tend to underestimate the sensitivity of runoff to a given change in rainfall, leading to poor performance when evaluated over multiyear droughts. This research revisited this conclusion, investigating whether the observed poor performance could be due to insufficient model calibration and evaluation techniques. We applied an approach based on Pareto optimality to explore trade-offs between model performance in different climatic conditions. Five conceptual rainfall runoff model structures were tested in 86 catchments in Australia, for a total of 430 Pareto analyses. The Pareto results were then compared with results from a commonly used model calibration and evaluation method, the Differential Split Sample Test. We found that the latter often missed potentially promising parameter sets within a given model structure, giving a false negative impression of the capabilities of the model. This suggests that models may be more capable under changing climatic conditions than previously thought. Of the 282[347] cases of apparent model failure under the split sample test using the lower [higher] of two model performance criteria trialed, 155[120] were false negatives. We discuss potential causes of remaining model failures, including the role of data errors. Although the Pareto approach proved useful, our aim was not to suggest an alternative calibration strategy, but to critically assess existing methods of model calibration and evaluation. We recommend caution when interpreting split sample results.

1. Introduction

Water resource planning is essential to ensure the ongoing security of water supply for domestic, agricultural, industrial and environmental needs. Long-term streamflow projections inform this planning and help to anticipate potential future shortfalls in surface water supply. Estimates of water availability should take into account both historical observations of river flow and also potential changes in environmental conditions such as climate or land use.

Hydrologic processes exhibit variability and cyclical behaviour on a variety of time scales, from familiar short term cycles (diurnal, event and seasonal) to multidecadal [*Hurst*, 1951]. Alongside the reality of climate variability is the potential for long term trends due to climate change [e.g., *Covey et al*., 2003; *Forster et al*., 2007]. A number of elements of the hydrologic cycle could be affected, including rainfall and evapotranspiration [*Meehl et al*., 2007; *Donohue et al*., 2010; *McVicar et al*., 2012]. Although the effects on precipitation are uncertain [*Covey et al*., 2003], many parts of the world, including southern Australia are likely to see reduced rainfall [*Chiew et al*., 2009] and catchments may be persistently drier in the future than the past.

Hydrologic models are useful tools in planning for future variability in climate. They allow hydrologists to estimate the impact that long-term changes in climatic variables, such as rainfall, might have on water availability for human consumption or environmental needs. In this research we focus on *conceptual* rainfall runoff models, which aim to represent mathematically the concepts underlying physical processes, without direct reference to physically based equations. Conceptual models generally have minimal data requirements, require minimal computing time, and often provide comparable simulations to more complex models [e.g., *Refsgaard and Knudsen*, 1996], so they are relatively popular in practice. As reviewed below, many studies have concluded that conceptual models are generally not suitable when climatic conditions change

(nevertheless they are often used in such conditions), and the intention of this paper is to revisit this conclusion. Before reviewing this literature in detail we describe the tests that are commonly used to support the conclusion, specifically the concept of split sample testing.

To increase the level of confidence in the predictive capability of a given model, *Klemeš* [1986] recommended a scheme known as the Split Sample Test, whereby a portion of historic recorded data is withheld from the calibration period, and used to check that the model can perform well over a period that it was not calibrated to–– hereafter referred to as an *evaluation* period rather than using the common terms *validation*, *verification* or *confirmation* [*Oreskes et al*., 1994; *Andréassian et al*., 2009]. In cases where a model will be applied in conditions different to the calibration period, *Klemeš* [1986] suggested that the calibration and evaluation periods be specifically chosen so as to reflect a similar contrast in conditions, a test known as the Differential Split Sample Test (DSST). In the context of a changing climate, whereby rainfall may be subject to long term trends, the DSST involves evaluating a model over a period that is significantly drier or wetter than the calibration period. More recently, variants of the DSST have been proposed, including the idea of using multiple calibration and evaluation periods via a sliding window in time [*Coron et al*., 2012, 2014].

Studies that have applied the DSST to assess the capabilities of models over a changing climate have generally reported unfavourable results. Model predictive ability following a change in climate does not appear to improve with more complex models, as demonstrated by *Refsgaard and Knudsen* [1996] who tested three models of varying complexity on three catchments in Zimbabwe. Furthermore, a number of studies have identified significant bias following application of the DSST. *Hartmann and Bárdossy* [2005] applied a lumped conceptual model to a 2000 km^2 catchment in Germany, calibrating successively to "wet," "dry," "warm" and "cold" years. They found that models calibrated to the wet periods systematically overestimated flow during dry periods unless the objective function explicitly included performance measures calculated over longer (e.g., annual) time steps. *Coron et al*. [2012] applied three conceptual models to 216 catchments and reported that "calibration over a wetter (drier) climate than the validation climate leads to an overestimation (underestimation) of the mean simulated runoff" (ibid. p1). *Chiew et al*. [2009] applied two conceptual models to provide climate change projections based on downscaled GCM outputs across south east Australia. Testing model performance over various periods with different climatic characteristics, they reported reductions in Nash Sutcliffe Efficiency (NSE) value of 0.1–0.3 compared to the calibration period, and long term bias of 30–40% in some cases. The recent workshop entitled *Testing simulation and forecasting models in non-stationary conditions* [*Thirel et al*., 2015a], held under the auspices of the International Association of Hydrological Sciences (IAHS), further confirmed––for a wide range of models and catchments––that hydrological models tend to perform poorly if applied under changing climatic conditions [*Thirel et al*., 2015b, and citations therein].

Some researchers have sought to quantify acceptable changes in climatic variables such as rainfall, such that a calibrated model still provides acceptable results. *Vaze et al*. [2010] tested four rainfall-runoff models in 61 catchments in South East Australia, and reported that the calibrated parameter sets generally gave acceptable simulations provided rainfall changes were not too large––no more than 15% less or 20% greater than rainfall over the calibration period. Similarly, *Singh et al*. [2011] identified an acceptable change of 10% drier or 20% wetter for five catchments across the continental USA.

Other studies have phrased the problem in terms of the nonstationarity of model parameters across different climatic conditions. *Merz et al*. [2011] applied the HBV model to 273 catchments in Austria and found that parameters relating to snow melt and the nonlinearity of runoff generation tended to change with time, showing significant correlation with climatic variables such as temperature. *Coron et al*. [2014] similarly observed problems with parameter robustness in twenty mountainous catchments in southern France. Some studies have observed that even if a rainfall-runoff model may appear to perform poorly in the DSST, it is usually possible to find a parameter set that can match a given period, even if it is unusually dry or wet, provided that the model is directly calibrated to that period exclusively. This observation led to *Li et al*. [2012] recommending 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" (ibid., p. 1239). Similar sentiments were expressed by *Vaze et al*. [2010]. However, this solution is limited to providing predictions that are within the range of climatic conditions experienced in the past [cf. *Refsgaard et al*., 2014]. *Choi and Beven* [2007] tested a hydrologic model in a

South Korean catchment and evaluated it over a variety of climatic conditions. Despite good performance according to classical performance measures on the time series as a whole, no parameter set tested was considered behavioural over all 15 of their categories of climatic conditions.

Despite these problems, some studies have had success searching for robust parameter sets, that is, parameter sets that can replicate streamflow over a wide variety of climatic conditions. *Hartmann and Bárdossy* [2005] formulated a number of objective functions based on least squares calculations at different time steps (e.g., daily, annual, decadal). Methods that combined both annual and daily objective functions into a single "meta-objective" were shown to reduce the error in annual flows from 30% to 10%. *Shamir et al*. [2005] applied similar multitimescale logic but based their analysis on flow statistics (signatures) rather than least squares measures. The result was an ensemble of parameter sets that performed well on all time scales considered; the identifiability of parameters in the Sacramento model was also improved. *Bárdossy and Singh* [2008] introduced the statistical concept of data depth to hydrological modeling. A parameter set has greater depth if it is located closer to the centre of a cloud of well performing sets. They found that parameter sets with greater data depth were more robust in split sample tests and less sensitive to random errors in input data.

Although in the above discussion we have used the term *model* quite loosely, henceforth we adopt the terminology outlined in *Andréassian et al*. [2009] where *model structure* refers to a set of equations representing a catchment whereas the term *model* refers to a model structure populated with a particular set of model parameters. A number of studies have concluded that a particular model structure is unsuitable for modeling under a changing climate [e.g., *Vaze et al*., 2010]. Others have suggested that a given model structure needs changing to do so [*Merz et al*., 2011] or have gone further and actually produced a model structure specifically designed to simulate under changing climatic conditions [e.g., *Ramchurn*, 2012; *Hughes et al*., 2013]. However, given the success of the studies mentioned above in finding more robust parameter sets under changing climates, perhaps the greater part of the problem lies with calibration and evaluation techniques rather than model structures. We suggest that a conclusion of model structure invalidity actually requires a much higher standard of proof than the tests of model evaluation suggested by *Klemeš* [1986]. To conclude that a model structure is invalid is to assert that no suitable parameter combinations exist [e.g., *Vogel and Sankarasubramanian*, 2003]; whereas *Klemeš* [1986] methodology seeks only to test the suitability of a chosen parameter set(s).

This research sought to investigate the apparent deficiency of a range of conceptual rainfall runoff model structures, across a large sample of catchments. The key research question was, *Are current conceptual rainfall runoff model structures deficient in their ability to simulate streamflow responses to long term changes in climate?* As described above, some existing literature portrays rainfall runoff models as suffering from poor performance if applied in climatic conditions different to those against which they were calibrated. The hypothesis tested here is that *the poor performance is due to poor or insufficient model calibration and evaluation techniques rather than deficient model structures*.

To conclude this section, we wish to clarify our intended meaning when using words such as *deficient*. *Gupta et al*. [2012] among others note that different hydrologists have different perspectives when defining model adequacy, contrasting the "physical science" viewpoint (where adequacy means consistency with the physical system) with the "engineering" viewpoint (where adequacy means that the model can emulate system input-output behaviour). In the context of rainfall runoff models, a physical science viewpoint would insist that a model can realistically represent the dominant physical processes occurring in a river catchment, whereas an engineering viewpoint would focus on whether the model streamflow outputs match with observations. We affirm the physical science viewpoint and the need to advance hydrologic science by developing more physically realistic models. However, the general nature of our research question requires testing a large variety of case studies (86 catchments, 5 model structures, see section 2 and cf. *Gupta et al*. [2014]), which renders detailed consideration of physical processes in each individual case difficult. Therefore, we use the word "deficiency" in a sense consistent with the engineering viewpoint, and this study investigates the ability of models to provide an empirical match with observed streamflow data, in catchments subject to long-term changes in climate.

2. Method

2.1. Rationale

To explain the methodology, let us consider a simple hypothetical case study. A rainfall runoff model structure A is applied to a catchment B using a calibration method C. Let us assume that method C is a single

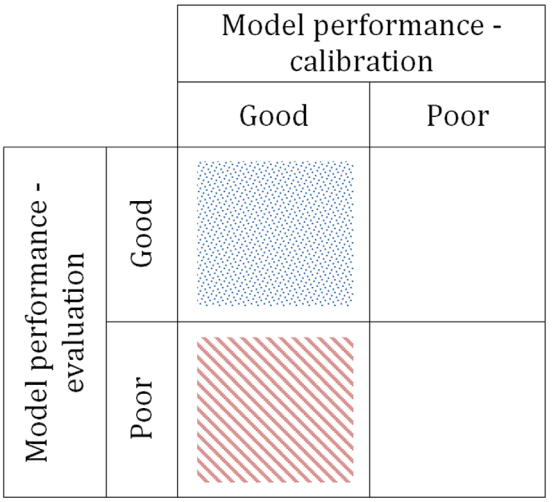

Figure 1. Explanatory diagram for possible outcomes of a Differential Split Sample Test, using idealized categories. The results of such a test can be considered to lie in a two dimensional objective function space.

objective optimization, such as would commonly be used within a DSST, optimizing to a single objective function that varies between 0 (poor) and 1 (good). In order to conduct a DSST, the observed data are split so as to reserve a period for independent evaluation. The evaluation period is much drier, on average, than the calibration period. The result is a very good score over the calibration period (say, 0.9) but a very poor score over the drier evaluation period (say, 0.2). Since the purpose of the exercise is to identify a model that performs well in evaluation [*Klemeš*, 1986], it is tempting to conclude that model structure A has failed for catchment B, or that poor data quality is degrading performance. However, consider Figure 1. The parameter set identified in optimization lies in the red hashed region, whereas a solution in the blue dotted region is desired. However, the fact that the parameter set identified as optimal over the calibration period is in the red region, does not imply that no parameter set exists in the blue region. For example, it may be that the model structure itself is capable of simulating well in evaluation, but the relevant parameters remained poorly identified in this particular calibration exercise. Some other parameter set which has slightly lower (but still good) performance over the calibration period may exist that also performs adequately in evaluation. This latter question remains untested in this hypothetical case, and is the subject of this paper. Note also that a parameter set in the red region may result from a calibration procedure caught in a local optimum [see e.g., *Arsenault et al.*, 2013].

Continuing the hypothetical case study, the periods are now switched, and the period that was the evaluation period now becomes the calibration period, and vice versa. The calibration is re-run and the result is that the score over the dry period is much increased (say, to 0.8) at the cost of some performance over the nondry period (say, to 0.5). In summary, in this hypothetical we have conducted two separate calibrations to two independent periods, and in each case we have obtained a parameter set that performs well over its training data, but poorly over the evaluation data. Figure 2a considers this as a two dimensional plot. Since

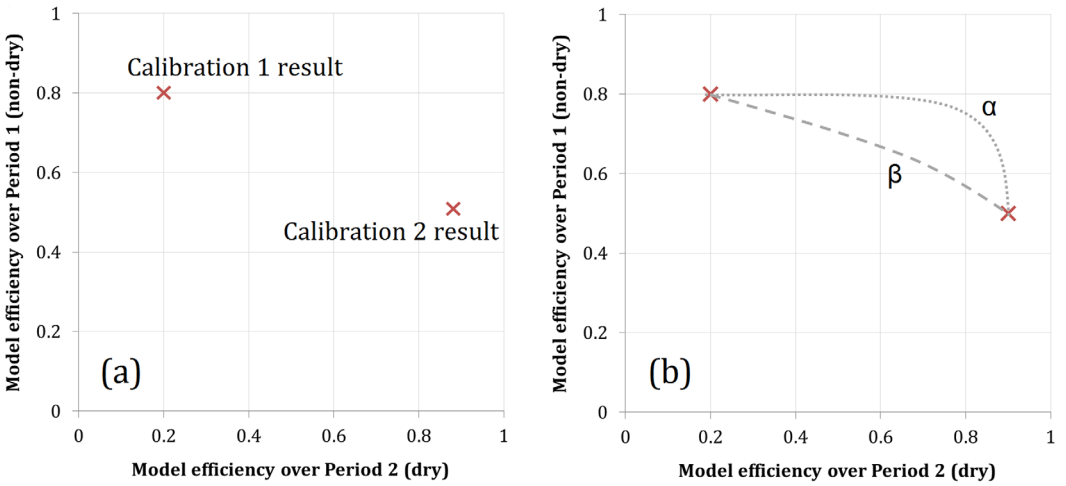

Figure 2. (a) Results of two hypothetical calibrations, plotted in two dimensional objective space. (b) Two Pareto Fronts joining the two points from Figure 2a. Each front is composed of multiple parameter sets. Obtaining curve α would demonstrate that a model structure has relatively greater potential for simulation under changing climatic conditions than would curve β.

both periods have had a turn at being calibration and evaluation periods, we dispense with this language altogether, and name them simply *Period 1* and *Period 2*. For clarity we use the descriptions "nondry" and "dry," respectively. If parameter sets which are robust to changes in climatic conditions exist, they will have high values on *both* the x axis and the y axis. Whether or not our rainfall runoff model structure A meets this condition depends upon the shape of the line that joins the two points and describes the trade-off between one objective and the other–– the Pareto Front [*Pareto*, 1927]. Under this scheme, a model structure with Pareto Front α in Figure 2b is more likely to produce robust simulations under changing climate than a model structure with Pareto Front β. Each Pareto Front is composed of numerous parameter sets, and Pareto Front α indicates robust simulations are possible because it contains parameter set(s) that have good performance on both the x axis and the y axis (e.g., [0.8, 0.75]). Such robust parameter set(s) are akin (but not identical) to the *hydrologic optimum* of *Andreassian et al*. [2012], which they define as the parameter set "that ideally would permit representing the catchment under all possible calibration periods encompassing climate forcings of interest, i.e. one allowing extrapolation." [ibid. p2206] In contrast, the two endpoints of the curve are the *mathematically optimum* parameter set(s) obtained via optimization to each of the single objectives in turn.

Based on the above, the search for parameter sets that are robust to changing climatic conditions can be informed if we know the shape of a model structure's Pareto Front. In this study we therefore applied a multiobjective optimizer to define the Pareto Front. Note that the method is intended only to critically assess existing methods of model calibration and evaluation; in this paper we are not suggesting that this method should be adopted for general use in rainfall-runoff model calibration.

The remainder of this section is organized as follows: we first present the method for identifying the Pareto Front, called AMALGAM [*Vrugt and Robinson*, 2007]; we then present the rainfall runoff model structures to be tested; the catchment case studies; input data; the objective functions used; and methods for checking the results of the AMALGAM algorithm.

2.2. Multicriteria Analysis and Pareto Search Method

Multicriteria analysis has been used in hydrology for some time in various contexts [*Efstratiadis and Koutsoyiannis*, 2010]. Early examples included optimizing treatment and monitoring of groundwater contamination [*Cieniawski et al*., 1995; *Ritzel et al*., 1994] and the incorporation of multiresponse data in hydrologic modeling [*Seibert*, 2000; *Madsen*, 2003]. Some authors adopted multiobjective approaches to improve identifiability of highly parameterized distributed models [e.g., *Muleta and Nicklow*, 2005; *Bekele and Nicklow*, 2007; *Wöhling et al*., 2013]. Other studies have used multicriteria approaches to integrate different data types into model calibration, including "soft" information (such as local or expert knowledge [*Seibert and McDonnell*, 2002]) and regionalized information [*Kim and Lee*, 2014]. *Gupta et al*. [1998] suggested the potential for multiobjective calibration of rainfall runoff models using different aspects of the same observed time series of flow (e.g., high flow versus low flow metrics), and a number of studies have adopted this approach [e.g., *Booij and Krol*, 2010; *Kollat et al*., 2012]. The use of hydrologic signatures in model calibration can be seen as a variant on the multicriteria approach, although methodological approaches vary [e.g., *Shamir et al*., 2005; *Yadav et al*., 2007; *Bárdossy*, 2007; *Winsemius et al*., 2009; *Vrugt and Sadegh*, 2013]. *Gharari et al*. [2013] noted that in addition to trade-offs between different metrics in the same time period, there are also trade-offs between model performance during one period and performance during another. They defined Pareto Fronts on both of these levels, and then designed a meta-Pareto analysis to choose parameter sets that provided the best overall compromise on the objective functions considered, over all periods considered. A key difference with the current study is that they were proposing a new model calibration approach, whereas in the current study we are using Pareto analysis to critically assess existing methods of model calibration and evaluation.

A number of algorithms to search for Pareto fronts are available in the hydrologic literature with notable early contributions being the development of the hydrology-specific multiobjective calibration algorithms MOCOM-UA [*Yapo et al*., 1998] and MOSCEM [*Vrugt et al*., 2003]. However, the concept is used in many fields and numerous algorithms from outside the field of hydrology are potentially applicable (e.g., *Storn and Price*, 1997; *Deb et al*., 2002]. Algorithms are generally evolutionary rather than gradient-based, and this led *Vrugt and Robinson* [2007] to suggest a hybrid approach whereby the evolutionary process is conducted not only between different model parameter sets, but also between different search algorithms. The

Table 1. Details of the Five Conceptual Rainfall Runoff Model Structures Tested in This Study

Name	Original Authors	Number of Free Parameters	Comments are Model Code
GR4J	*Perrin et al.* [2003]	4	Checked against code provided by authors
SIMHYD	*Chiew et al.* [2002]	7	Code provided by authors
IHACRES[a]	*Jakeman and Hornberger* [1993]; *Ye* et al. [1997]	8	Code 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[b]	*Burnash et al.* [1973]	13	Based on code from the website of the National Oceanic and Atmospheric Administration (NOAA)[c]

[a]Note that IHACRES parameter PET_{ref} was set to zero.
[b]Note that three SACRAMENTO parameters were fixed (RSERV, SIDE and RIVA) as per *Duan et al.* [1994].
[c]http://www.nws.noaa.gov/iao/sacsma/fland1.f, accessed 30 March 2015.

resulting Pareto search meta-algorithm, called AMALGAM (A MultiAlgorithm, Genetically Adaptive Multiobjective method), calls upon four commonly used methods for multiobjective searches (NSGA-II––*Deb et al.* [2002]; Particle Swarm Optimization (PSO)––*Haario et al.* [2001]; Adaptive Metropolis Search (AMS)––*Kennedy et al.* [2001]; and Differential Evolution (DE)––*Storn and Price* [1997]). These search algorithms are run simultaneously during an AMALGAM run, and the evolution of the population of parameter sets is directed by a combination of the search algorithms, with the influence of each in proportion to its performance at that point in the search. *Vrugt and Robinson* [2007] reported efficiency gains of up to a factor of 10 in some multiobjective problems. For this research, we adopted AMALGAM to search for Pareto Fronts.

2.3. Rainfall-Runoff Model Structures

The intention of this study is to test a variety of model structures chosen to reflect common usage in the study area and, where possible, breadth of design of conceptual rainfall runoff models. Since this study is focused in Australia three model structures that are commonly used in Australia were selected: GR4J [*Perrin et al.*, 2003]; SIMHYD [*Chiew et al.*, 2002]; and IHACRES [*Jakeman and Hornberger*, 1993; *Ye et al.*, 1997]. We adopt the version of IHACRES used in similar studies in Australia [e.g., *Vaze et al.*, 2010] which incorporates the two parallel storages of IHACRES "Classic" [*Jakeman and Hornberger*, 1993; see also *Jakeman et al.*, 1990] with the option for a threshold of runoff production proposed by *Ye et al.* [1997]. These three model structures, GR4J, SIMHYD and IHACRES are the result of three different ways of formulating conceptual rainfall runoff models, as follows: (1) SIMHYD is an attempt to represent physical processes in conceptual equations, so that it has separate components for such processes as interception, infiltration excess overland flow, interflow/saturation excess flow and baseflow [*Porter and McMahon*, 1975; *Chiew and McMahon*, 1994; *Chiew et al.*, 2002]; (2) IHACRES has much less emphasis on physical processes, having been derived from mathematical analysis of the number of parameters that could reasonably be inferred from typical calibration data [*Jakeman et al.*, 1990; *Jakeman and Hornberger*, 1993]; and (3) GR4J has a similarly low emphasis on physical processes but was derived using an empirical approach that tested a large number of candidate structures and used a rejection method based on the empirical match with calibration data [*Perrin et al.*, 2001, 2003]. We consider that these three approaches to model formulation cover the majority of conceptual rainfall runoff models currently in the literature.

In addition, two further model structures were included. GR4JMOD [*Hughes et al.*, 2013] was chosen as a case study for improvement of rainfall runoff models. *Hughes et al.* [2013] started with the GR4J model [*Perrin et al.*, 2003] and tested a number of changes designed to better simulate environments with long-term (i.e., multiyear) catchment storage. Their changes allowed the soil moisture to deplete below the level required for runoff production, effectively increasing catchment "memory." They also added exponents to increase nonlinearity of runoff production and actual evapotranspiration. Note that *Hughes et al.*'s [2013] module to account for changes in Leaf Area Index was not adopted here. Lastly, one model structure has been selected because it is widely used in the literature and in practice in the USA, namely SACRAMENTO [*Burnash et al.*, 1973].

These model structures are summarized in Table 1. Model complexity varied, with the number of conceptual storages ranging from two to four, and the number of free parameters ranging from four to thirteen.

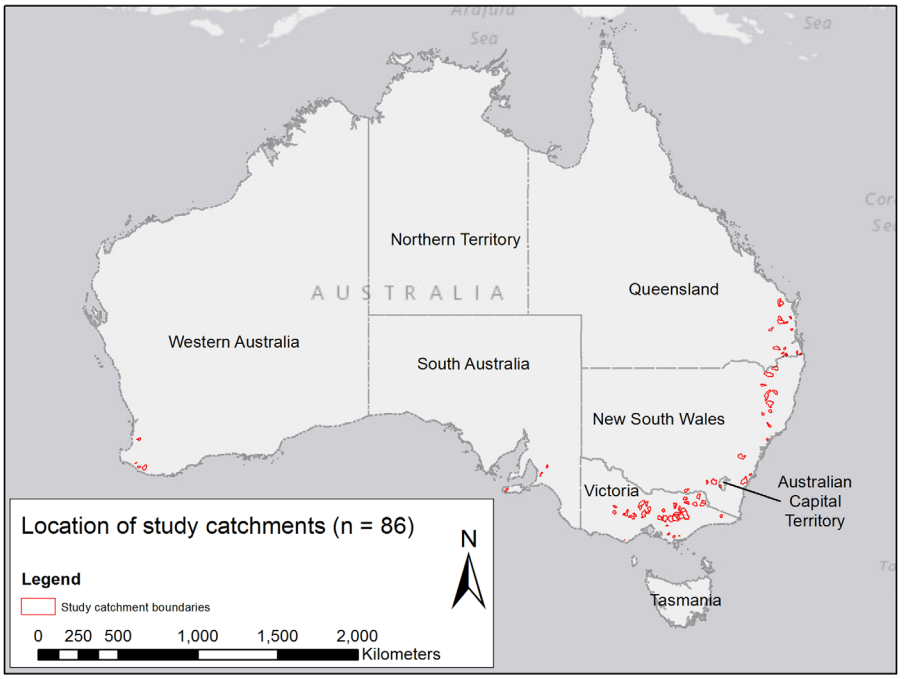

Figure 3. Study catchments used in this analysis.

All models take the same inputs, namely, rainfall and potential evapotranspiration (PET––note the adopted version of IHACRES used PET rather than temperature). A lumped modeling approach was taken, whereby a single time series was derived for rainfall and PET in each catchment (section 2.5). The modeling framework was implemented in a hybrid Matlab-Fortran system whereby the rainfall runoff models were run in Fortran 95 (checking against the code of the original authors where available–Table 1) which was called by the AMALGAM code in MATLAB provided by *Vrugt and Robinson* [2007].

2.4. Study Area

This study was conducted in 86 catchments in southern and eastern Australia (Figure 3). This region is well-suited to studying hydrological responses to long-term shifts in climate, because the variability of annual flows is relatively high on a global scale [*Peel et al.*, 2001] and there have been a number of dry periods lasting several years or even decades on which to test model simulations. For example, the reduction in rainfall since the 1970s in the south west corner of Australia relative to the 1960s [e.g., *Petrone et al.*, 2010] has led local water authorities to run their long-term planning simulations using post-1975 data only. The south-east of the country experienced a severe and prolonged drought throughout much of the 2000s, known as the Millenium Drought [*Potter et al.*, 2010]. River flows during the Millenium Drought, even given the low rainfall, were unexpectedly low in some areas [*Potter and Chiew*, 2011; *Chiew et al.*, 2014; *Saft et al.*, 2015]. These droughts had numerous impacts on Australian society, including installation of alternative water sources such as desalination in most major cities, the cessation of irrigation in some areas causing changes in rural communities, and revision of water allocation arrangements to include water trading and provision for environmental flows [see e.g., *Aghakouchak et al.*, 2014].

The 86 study catchments were chosen 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" (ibid, p6) such as regulation from large reservoirs and broadscale land use changes. Of the 154 Hydrologic Reference Stations that lie within southern Australia, (broadly defined as south of the Tropic of Capricorn), the list was refined according to:

1. *Data quality checking* including inspection of quality flags, missing data, plotted daily data, inspection of double mass curves for flow and rainfall, and plotting long term climatic averages on axes similar to

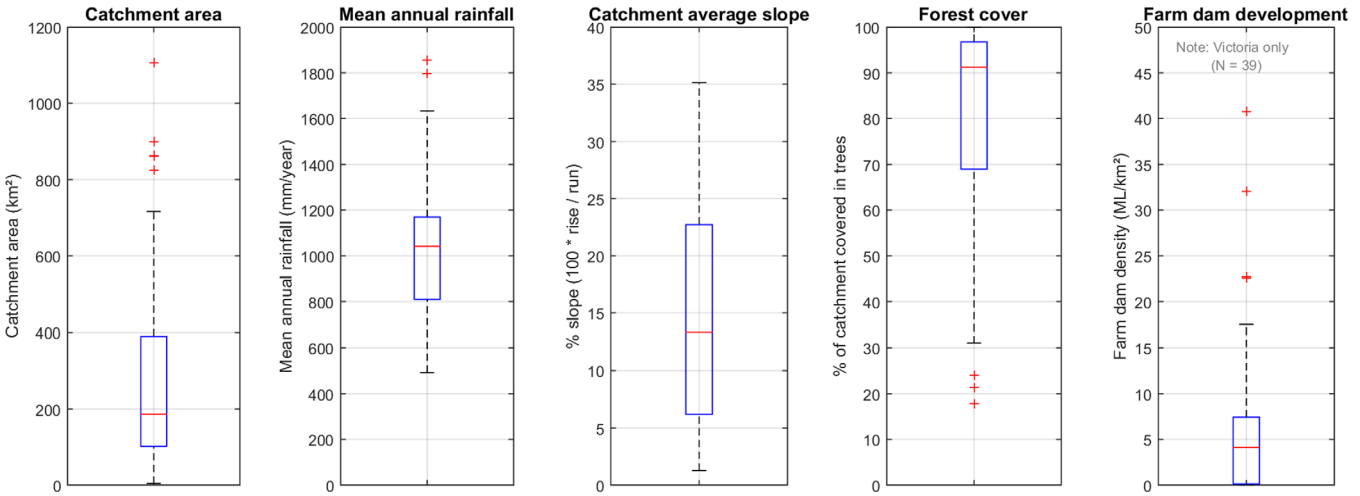

Figure 4. Catchment properties for the 86 study catchments. The whiskers extend a maximum of 1.5 times the interquantile range. Values beyond the whisker are marked as outliers and are denoted as +. Catchment average slope was derived based on analysis of a DEM (section 2.5) and represents the spatial average of cell-by-cell slope values. Forest cover was from *Lymburner et al.* [2011] and is the sum of the four landuses in the "tree" category. Farm dam development is based on the dam locations and estimated volumes published by the *Department of Environment, Land, Water and Planning* [2015a,b]. For catchment area and rainfall data, see section 2.5.

those used by *Budyko* [1971]––specifically, Actual Evapotranspiration versus Potential Evapotranspiration, both normalized by rainfall [see also *Zhang et al.*, 2001].

2. *Rain gauge coverage*: Catchments were checked for coverage of rainfall gauges, and catchments with relatively low coverage were flagged.
3. *Spatial rainfall contrasts*: As mentioned above, a spatially lumped modeling approach was adopted, meaning that a single rainfall time series was used over a catchment (namely, the spatial average). Catchments with high spatial contrasts in rainfall are more difficult to simulate using a lumped approach, because the average rainfall is generally less representative of the rainfall extremes within the catchment. While a certain degree of rainfall contrast is usually inevitable due to topographic differences, the catchments with relatively higher contrast were flagged. Rainfall contrasts were assessed using the gridded rainfall data as described in the next section.

The final data set of 86 catchments (Figure 3) was chosen so as to exclude those catchments with the clearest data issues, the lowest rain gauge coverage, and/or the highest spatial rainfall contrast, while aiming to preserve both the majority of catchments, and the spatial and climatic coverage inherent in the original data set.

The set of 86 catchments vary in size from 4.4 to 1106 km^2, with 49 of the catchments between 100 and 500 km^2 (see Figure 4). All of the catchments are in the temperate climate zone, falling within Group C of the Köppen-Geiger climate classification scheme [*Peel et al.*, 2007]. This means that the average maximum temperature of the hottest month is greater than 10°C, and the average maximum temperature of the coldest month is between 0°C and 18°C. Mean annual rainfall is generally less than 1200 mm/yr, while catchment average slope is generally less than 25% (Figure 4). Forest cover is generally high, with tree cover exceeding 90% in over half of the catchments. Catchment elevation ranges from sea level to 2000 m AHD, although most catchments do not exceed 1500 m AHD. Winter snowfall occurs in some catchments, but the snowpack is generally not sufficient to significantly affect hydrology. The development of small private waterbodies (referred to as "farm ponds" in the USA and "farm dams" in Australia) was also assessed where available. Over half of the catchments had an estimated farm dam storage of less than 5 ML/km^2, which can be considered quite low [*Nathan and Lowe*, 2012], although three catchments had more than 20 ML/km^2. These physical properties will be related to model performance later in the paper.

2.5. Input Data

The two main inputs to the rainfall runoff models were rainfall and potential evapotranspiration (PET), each derived as a timseries on a daily time step. Rainfall was derived from the interpolated gridded product of *Jones et al.* [2009] which is available as a set of daily grids at a resolution of approximately 5 km, based on

gauged rainfall data and including land elevation as a spatial co-variate. For each day to be simulated, the spatial average across the catchment was derived from the daily grids from *Jones et al.* [2009]. PET estimates were derived using the Wet Environment method from *Morton* [1983]. Given the relatively low spatial variability of potential evapotranspiration, this was extracted for the catchment centroid only, from the gridded data sets produced by *Jeffrey et al.* [2001].

In the case of both rainfall and PET, the catchment boundary was required in order to extract information from the gridded data sets. Catchment boundaries were derived using flow analysis on Shuttle Radar Topography Mission (SRTM) data on a grid size of 1 s (approximately thirty metres). The post processed version by *Gallant et al.* [2011] was used for the flow analysis, which was done in ESRI's ArcHydro toolbox using the D8 method to define flow pathways.

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 with quality issues were excluded from the analysis. Since quality code systems are different for each state of Australia, the details of this checking depended on location.

2.6. Defining Dry Periods and Wet Periods

As described in section 2.1, the intention of the Pareto analysis is to search for parameter sets within a given model structure that provide a favourable trade-off between performance in dry climatic conditions and performance in wet climatic conditions. There were two separate tasks in order to develop this logic into a working system: first, to define "dry periods" and "wet periods" more precisely (this section), and second to choose a single objective function as an indicator of model performance over a given period (described in the next section).

To define dry periods, it would be possible to simply select the driest X% of years (or months), regardless of where those years may fall in the historic record. The results would be a set of years that are not concurrent. However, one of the key aspects of the recent droughts in Australia was not only their severity but also their length and persistence; the persistent dry conditions have been shown to be associated with lower than expected streamflow response [*Petrone et al.*, 2010; *Potter and Chiew*, 2011; *Hughes et al.*, 2012, 2013; *Potter et al.*, 2013; *Saft et al.*, 2015]. Therefore, we focused on *sequences* of dry years in this research, with the intention of examining multiyear droughts. While a number of studies have proposed methods of defining drought [see e.g., *Mishra and Singh*, 2010, 2011, and citations therein] there is no single accepted method for doing so. In this study we opt for a relatively simple definition, where we define the "dry period" to be the driest consecutive set of years of a given length in the historic record. Given that the Millennium Drought is generally considered to have lasted from 1997 to 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). It was felt that such a year could dominate the calculation of performance metrics relative to the drier years that are the topic of interest. Therefore, it was decided to use a shorter period, specifically seven years, instead. Thus, the dry period for this paper (also called "Period 2") is defined as the driest set of seven consecutive years in the historic record. This is defined according to streamflow, not rainfall.

While it is possible to define a "wet period" in a similar way, (i.e., by identifying the wettest series of concurrent years in the historic series), we have elected to adopt a method similar to that described in the hypothetical in section 2.1. We defined "Period 1" as all years in the historic record, apart from Period 2––that is, Period 1 is the complement of Period 2. This definition meant that Period 1 contains the majority of the historic data. Since the intention of this paper was to provide a critique of the single-objective calibration approach (i.e., single objective calibration to the nondry period and subsequent evaluation to the dry period), it was logical to provide as much calibration data as possible to this approach, such that the method under scrutiny was given the best possible chance to succeed. In any case, when calibrating to objective functions such as the Kling Gupta Efficiency used in this paper [*Gupta et al.*, 2009] (see next section), the wetter periods tend to be matched preferentially since the components of the KGE (linear correlation, error in mean and error in standard deviation) tend to be more strongly influenced by larger flow values. Thus, performance in Period 1 is an acceptable surrogate for performance over the wettest years in a given time series. For convenience, Period 1 will be referred to as the "nondry" period.

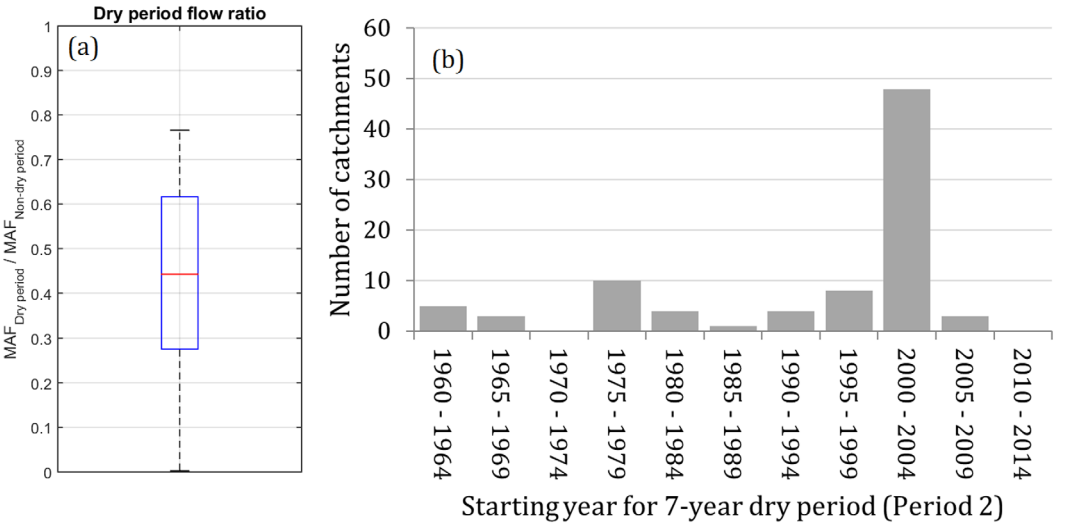

Figure 5. (a) Boxplot of mean annual flows in Period 2 (dry period) expressed as a ratio to Period 1 (nondry period). The whiskers extend a maximum of 1.5 times the interquantile range. Values beyond the whisker are marked as outliers and are denoted as +. (b) Histogram showing the most common starting years for the 7 year Period 2.

We acknowledge that, in a given case study, Period 1 will usually contain some years that are relatively dry. Period 1 may contain entire sequences of droughts that were not the most severe on record, plus portions of the most severe drought not captured within Period 2 in cases where drought duration exceeds 7 years. Conversely, Period 2 may contain years that were immediately prior to or following the drought of interest, in cases where the most severe drought is less than 7 years in duration. Nonetheless, these simple definitions were sufficient to examine differences in model performance between wet and dry periods, particularly for catchments where droughts tended to be longer and more severe.

Using these definitions, the flow series in each catchment was analyzed so as to define Period 1 (nondry) and Period 2 (dry). As expected, the mean annual flow tended to be significantly less over Period 2 than over Period 1 (Figure 5a); for example, in over a quarter of catchments the flow reduction exceeded 70%. Figure 5b shows the distribution of start years over the set of 86 catchments. The duration of the Millennium Drought is considered to have been 1997–2009 [*Chiew et al.*, 2014], so it is not surprising that the starting year of Period 2 is commonly within the range 2000–2004 (48 out of 86 catchments).

2.7. Objective Functions

When deciding which objective function to use, the NSE [*Nash and Sutcliffe*, 1970] was the first candidate considered because its common use in practice means that values of NSE can be interpreted by a relatively wide audience. However, we encountered problems using the NSE. In some cases the NSE value was quite high (i.e., ~0.8) but upon further investigation the simulations were significantly biased. *Gupta et al.* [2009] provide an explanation for this in their decomposition of the NSE into the linear correlation and terms related to the error in the mean (i.e., the bias) and error in the standard deviation. *Gupta et al.* [2009] noted that the bias term is normalized by the observed standard deviation, which means that in catchments with high flow variability (as in this study) the magnitude of the bias can be high without penalising the NSE score. One option was to add a bias weighting to the NSE, as applied by, for example, *Vaze et al.* [2010]. However, *Gupta et al.* [2009] noted a further problem with the treatment of the standard deviation σ in the NSE, regarding the ratio $\sigma_{simulated}/\sigma_{observed}$. Although this ratio should ideally have a value of unity, the mathematically optimum value for NSE occurs when the ratio is equal to the linear correlation. Given these problems, this study adopted the alternative objective function proposed by *Gupta et al.* [2009], called the Kling-Gupta Efficiency, or KGE. The KGE is a function of the same three components as the NSE (linear correlation; error in mean; error standard deviation) but the formulation removes the interactions between the components, providing a more robust measure of model performance. For those readers who are not familiar with KGE scores, in the supporting information we provide a table that relates the KGE to the more

familiar NSE objective function (supporting information Figure S3), and also highlights the problems noted by *Gupta et al.* [2009].

2.8. Results Checking

Since AMALGAM is an evolutionary algorithm, it is possible that calibration runs may proceed in different directions through the parameter space and have divergent end results [see e.g., *Arsenault et al.*, 2013; *Peterson and Western*, 2014]. To check the consistency of the AMALGAM results, we started with a relatively low number of function evaluations (10,000) and ran the algorithm three times, resulting in three different Pareto Fronts. These Pareto Fronts were checked for consistency both visually and using the numerical rule that the Euclidian Distance separating any two of the three curves could not exceed 0.01 at any point on the curves. If this numerical rule was violated, the number of function iterations was doubled and the analysis re-run and re-checked. Around one quarter of the case studies passed at the first iteration (i.e., 10,000 function evaluations). Case studies that failed the numerical test at 40,000 iterations were manually (visually) checked and accepted only if the differences were judged to be immaterial to the conclusions of this paper.

The presentation of results in the following section initially focuses on one objective (i.e., one period) at a time, before moving to consideration of the two objectives (i.e., performance over dry and nondry periods) simultaneously. Presentation of the AMALGAM results in this way implicitly assumes that AMALGAM is a sufficiently powerful search algorithm to find the optimum of a single objective. Another way of stating this assumption is that the endpoints of the Pareto Curves are assumed to be accurate. To test this assumption, the single-objective optimization algorithm CMA-ES [*Hansen et al.*, 2003] was applied. CMA-ES has been widely used across a number of fields and tested favourably in the context of hydrology compared to more common methods in hydrology such as Shuffled Complex Evolution [see *Duan et al.*, 1992; *Arsenault et al.*, 2013; *Peterson and Western*, 2014]. In the current study, CMA-ES was trialed in ten catchments, for each of the five model structures, for each of the two objectives (KGE over Period 1 and KGE over Period 2). This gave a total of one hundred CMA-ES case studies. Similarly to AMALGAM, CMA-ES was run three separate times and if the results were not consistent, the number of restarts (the only user-defined parameter in CMA-ES) was increased by one (starting from zero restarts) and the process was repeated.

For brevity, the CMA-ES results are not shown in the body of this paper but are provided in the supporting information (supporting information Figure S4). In summary, the results indicated that AMALGAM was a capable and reliable optimizer to a single objective. Optimization results (in terms of KGE scores) were within 0.005 in 76 of 100 cases. In the remaining 24 cases AMALGAM produced the best result in 15 and CMA-ES in 9. There were a few cases where AMALGAM results were significantly better than CMA-ES. In fact, ordering the case studies according to the absolute difference between the two results revealed that the top five cases (cases of greatest difference) were all cases where AMALGAM found a better solution than CMA-ES. We also note that, on average, the AMALGAM algorithm generally used less function evaluations than CMA-ES, although this varied based on the case study. Given these favourable results, we will now present the AMALGAM results with similar confidence as we would have in a dedicated single-objective optimizer.

3. Results

3.1. Performance When Optimizing to Each Objective in Isolation

As demonstrated in the previous section, although a tool for multiobjective problems, the AMALGAM algorithm can also be used to provide results of a single-objective optimization, by considering the endpoints of the Pareto curves only. In this section, we present single-objective Differential Split Sample Test results, extracted from the wider set of AMALGAM outputs.

In general, optimizing the rainfall runoff models to KGE over Period 1 (nondry) provided good KGE values over Period 1 (Figure 6a). The median KGE score across all 86 catchments was 0.8 or higher, regardless of which rainfall runoff model structure was chosen. For those readers who are not familiar with KGE scores, in the supporting information we provide a table that relates the KGE to the more familiar NSE objective function (supporting information Figure S3). The GR4J and GR4JMOD model structures appeared to perform best. However, when the same parameter sets were evaluated by simulating flows over the driest 7 consecutive years (Period 2), model performance was much lower (Figure 6c). The model structures with the highest calibration KGE scores (GR4J and GR4JMOD) showed negative evaluation KGE values in more than 25%

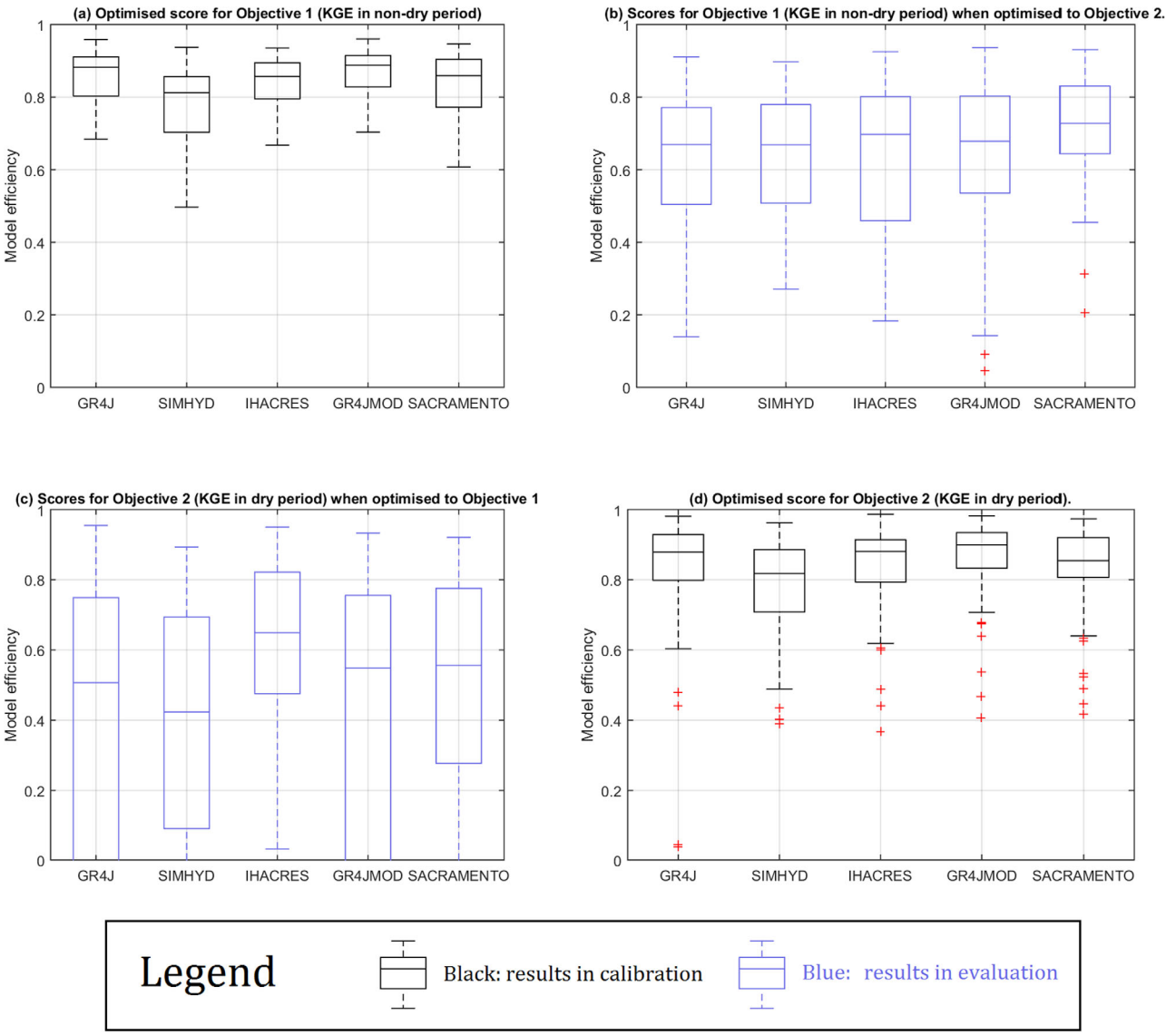

Figure 6. Values of Kling Gutpa Efficiency (KGE) for calibration and evaluation when optimized to Period 1, the nondry period (top) and Period 2, the dry period (bottom). Note that negative values exist but are not shown. The whiskers extend a maximum of 1.5 times the interquantile range. Values beyond the whiskers are marked as outliers and are denoted as +.

of catchments. IHACRES was comparatively better, with a median score of 0.67. Nonetheless, in general, the performance was markedly reduced when moving from wetter to drier climatic periods. Furthermore, some of the lowest values of KGE in evaluation corresponded to relatively high KGE values in calibration (Figure 7). These findings are consistent with the literature review [e.g., *Vaze et al.*, 2010; *Coron et al.*, 2012; *Thirel et al.*, 2015b].

If the dry period (Period 2) was used as the calibration period instead of the evaluation period, results demonstrated that the rainfall runoff models are generally able to replicate the flows during dry conditions, provided they are directly calibrated to them in isolation. However, there were some exceptions, particularly for the GR4J model structure, as shown by the outliers in Figure 6d. The reduction in performance between the calibration period (dry period, Figure 6d) and the evaluation period (nondry, Figure 6b) is less pronounced than in the previous case (Figures 6a and 6c) but is still evident. As above, some of the lowest values of KGE in evaluation corresponded to relatively high KGE values in calibration (Figure 7).

In summary, the model structures tested were generally able to replicate flows over a given set of climatic conditions, whether dry or wet, provided that they were directly calibrated to those conditions [*Li et al.*,

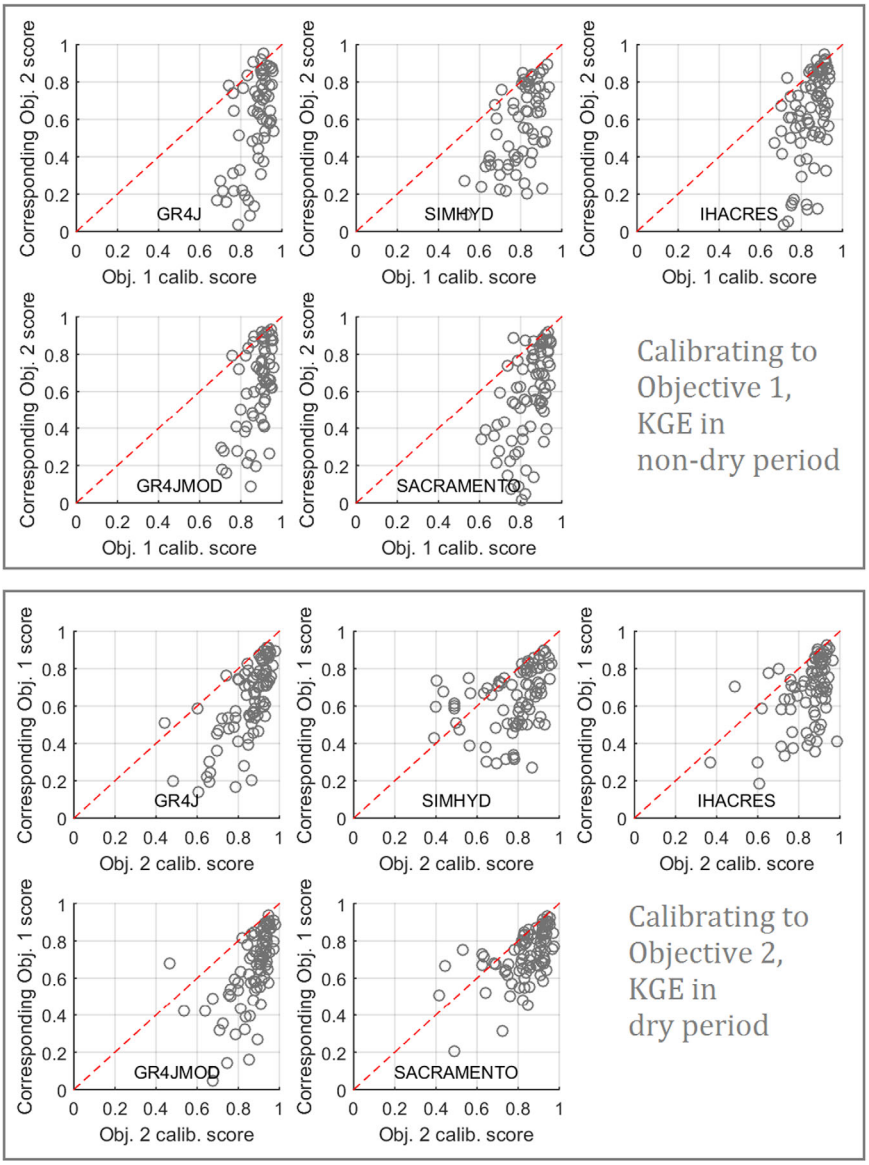

Figure 7. Scatter plots of calibration versus evaluation KGE values when calibrating to Period 1, the nondry period (top) and Period 2, the dry period (bottom). Each circle represents a catchment, with values of calibration KGE scores on the x axis plotted against evaluation KGE scores for the same parameter set on the y axis. Note that negative values exist but are not shown.

2012]. The key problem was that the parameter sets identified by optimization to one set of climatic conditions performed poorly in different conditions; that is, the mathematically optimum parameter sets identified were not robust to changes in climate. In subsequent discussion, the results presented in this section will be referred to as the results of a "single-objective DSST," since the models were calibrated to only one objective at a time; i.e., KGE in one set of climatic conditions, with subsequent evaluation in different climatic conditions. These single-objective DSST results are used in this paper as a baseline method representing common practice.

3.2. Pareto Curve Results

For each of the five model structures, AMALGAM was applied to derive a Pareto Front between the two objectives (i.e., between KGE in Period 1 (nondry) and Period 2 (dry)) in each of the 86 study catchments. To explain and interpret these results, we first use the example of the Rocky River upstream of Gorge Falls (Station A5130501), a 190 km^2 catchment on Kangaroo Island, South Australia (mean annual rainfall = 730 mm/

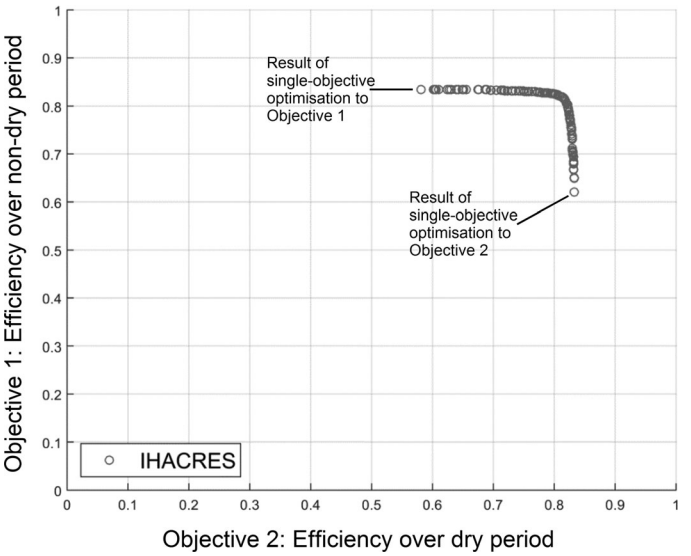

Figure 8. Pareto Front identified by AMALGAM between the two objectives, for the Rocky River upstream of Gorge Falls (A5130501), using the IHACRES rainfall runoff model structure.

yr; rainfall-runoff ratio 0.1). The dry period in this catchment was found to be 2001–2007 inclusive, and average streamflow over this period was only 40% of the long-term average. For this station and the IHACRES model structure, the single-objective DSST metrics were:

1. $KGE_{nondry}(\text{calibration}) = 0.835$, $KGE_{dry}(\text{evaluation}) = 0.581$;
2. $KGE_{dry}(\text{calibration}) = 0.833$, $KGE_{nondry}(\text{evaluation}) = 0.621$.

These figures indicate that the single objective approach identified parameter sets that performed well in one set of climatic conditions or the other, but not both. Let us now consider whether the Pareto approach can identify robust parameter sets that perform well in both periods. Figure 8 shows the Pareto Front identified by AMALGAM, displayed in two-dimensional objective function performance space. The results quoted above form the endpoints of the Pareto curve in this space (i.e., the endpoints are [0.581, 0.835] and [0.833, 0.621]). Since AMALGAM is an evolutionary method that uses a finite population, the front is displayed not as a continuous line but as a set of discrete points, one for each parameter set (in this case N = 100). A number of those parameter sets are in the region of the objective space where KGE_{dry} and KGE_{nondry} both exceed 0.8. Thus, given that the values of both objectives are favourable *for the same parameter set*, we may cautiously conclude that the IHACRES model structure is capable of providing robust simulations over changing climatic conditions (assuming that the KGE can be considered a suitable indicator of simulation performance). Henceforth in this paper we will use the terminology "false negative" to refer to cases such as this where suitable parameter set(s) exist within a model structure, but the DSST fails to find them.

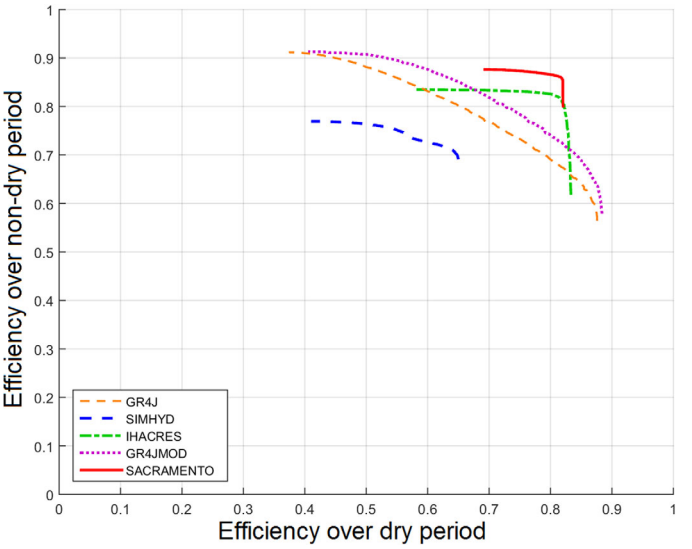

Figure 9. Pareto curves for each model structure for the Rocky River upstream of Gorge Falls (A5130501).

Next, let us consider the results for other model structures applied to the same catchment. While one model structure (SACRAMENTO) performed better, the remainder did not, as shown in Figure 9. The GR4J and GR4JMOD structures were capable of high KGE scores in either the dry period or the nondry period, as indicated by the end-points of the Pareto curves (cf. Figure 6). However, the curves joining these points do not approach the region of favourable trade-off mentioned above; that is, there were no parameter sets robust to changes in climate for these structures in this catchment. Note that in Figure 9, the individual markers have been replaced by lines for ease of viewing.

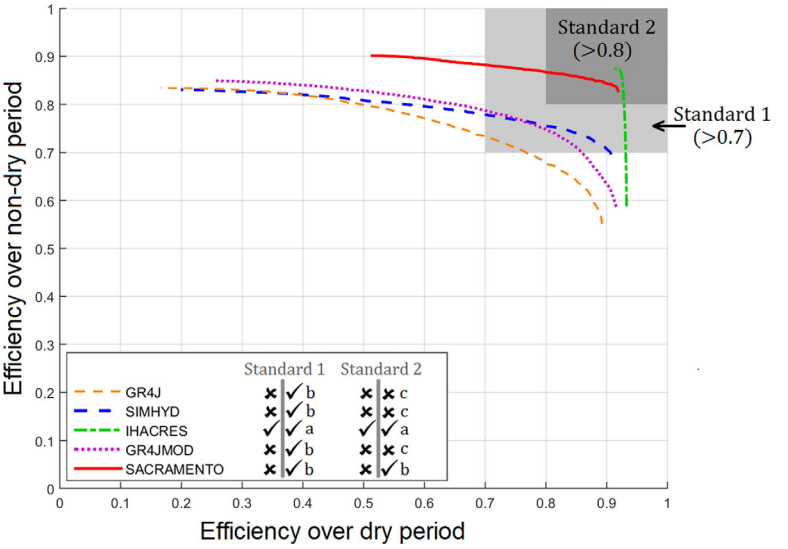

Figure 10. Pareto fronts for Catchment 405274, with annotations regarding the meeting of modeling standards 1 (light grey) and 2 (dark grey). The ticks and crosses refer to results for the single-objective DSST to Objective 1 (left of the line) and AMALGAM (right of the line).

The results for GR4J and GR4JMOD in Figure 9 provide an instructive case study in model assessment. The endpoints of the curves are similarly placed for each of these two models. Thus, use of a single-objective DSST (as presented in the previous section) would lead to the erroneous conclusion that the alterations to GR4JMOD by *Hughes et al.* [2013] made negligible difference to the model's capabilities. In contrast, Figure 9 shows that this is not the case by the divergence of the purple GR4JMOD curve from the orange GR4J curve. Although the difference in this case is relatively modest, it is not an isolated case––four further case studies are shown in the supporting information (supporting information Figure S5). Thus, use of the single-objective Differential Split Sample Test may result in situations where highly successful model improvements are discarded as ineffective.

3.3. Identifying Model Structures That Meet Modeling Standards

One difficulty in moving from a single catchment example to the full set of 86 catchments is the challenge of displaying the results meaningfully across such a large sample. For the interested reader, the Pareto curves for every combination of model structure and catchment are provided in full in the supporting information Figures S1 and S2. Although we experimented (not shown) with measures to characterize the shape of the Pareto curve, here we focus instead on whether or not a given rainfall runoff model structure is capable of robust simulations under changing climatic conditions, as indicated by high KGE values. Graphically, such model structures have Pareto Curves that contain parameter sets that are relatively close to the "perfect" point, [1, 1].

Although it is difficult to say exactly how close is "sufficient" for a given case study, for the present study it is useful to define some subjective performance standards. By defining what "success" is (albeit in a subjective fashion), these standards allow us to more easily summarize the skill of the Differential Split Sample Test in identifying "successful" model structures. Two attempts at defining such a standard are depicted in Figure 10. In Standard 1, a "successful" model is one in which the model efficiency (KGE) at some point on the Pareto Curve exceeds 0.7 in *both* the dry and nondry periods. Standard 2 is similar except that the KGE benchmark is now 0.8; a higher standard of performance. Many other different standards could be formulated, and it is expected that the most suitable standard may depend upon the particular objectives of the study at hand. For the purposes of interpreting results in this paper, we will proceed with these two standards, and denote any parameter set that meets a given standard to be "suitable" (note that concepts of model adequacy are discussed in section 4.2).

For each case study (i.e., combination of model structure and catchment) we now ask two questions:

1. Would a suitable parameter set be found by a single-objective DSST calibrating over the nondry period (y-axis) and evaluating over the dry period (x axis)? (Note that this corresponds to the left hand extreme of the Pareto Curve, such that the y-axis ordinate is maximized).

2. Would a suitable parameter set be found by AMALGAM? i.e., is any portion of the Pareto Curve within the boxes of Figure 10?

We note that the DSST in point (1) above could equally be defined the other way around, with calibration over the dry period and evaluation over the nondry. However, climate projections for southern Australia generally agree that long-term average rainfall is likely to reduce under climate change [e.g., *Chiew et al.*, 2009]. Thus, it is more relevant within this study area to evaluate models in conditions that are drier than the calibration period.

There are three possible combinations of answers to the above questions:

a. Suitable parameter set(s), i.e., parameter set(s) that meet the performance standard, were found by *both* the single-objective DSST *and* AMALGAM;
b. Suitable parameter set(s) were *not* found by the single-objective DSST but *were* found by AMALGAM; and
c. Suitable parameter set(s) were *not* found by either method.

To explain these categories graphically, consider the curves in Figure 10, which show Pareto results for Home Creek at Yaark (Station 405274, 181.6 km^2, mean annual rainfall = 744 mm/yr; rainfall-runoff ratio 0.18). As an example, we consider the results for Standard 2 (dark grey). Only two of the model structures have a portion of the Pareto Curve within the box for Standard 2––SACRAMENTO (red) and IHACRES (green). This means that GR4J, GR4JMOD and SIMHYD are all in category (c) with respect to Standard 2. With respect to SACRAMENTO, although it can fulfil Standard 2, the parameter set that would be chosen by the single-objective DSST to Objective 1 (i.e., the endpoint [0.51, 0.90]) does not fulfil Standard 2. Thus, SACRAMENTO is in category (b) with respect to Standard 2. For IHACRES, the parameter set that would be chosen by the single-objective DSST to Objective 1 (i.e., the endpoint [0.92, 0.88]) does fulfil Standard 2. Thus, IHACRES is in category (a) with respect to Standard 2.

Hypothetically, if the results across all catchments and model structures indicated a dominance of case (a), then we would conclude that there are in fact few problems with current rainfall runoff model structures simulating changing climates (in the "engineering" sense; section 4.2), although there might still be some scope to improve them. However, the results presented above (e.g., Figure 6) have already demonstrated that this is not the case. Thus we are left with (b) or (c). Dominance of (b) would indicate that common single-objective calibration methods (as commonly used in the Differential Split Sample Test) generate an abundance of false negatives, and thus the problem is with the calibration methods, not with the model structures themselves. Dominance of (c) would support the argument that the model structures themselves need to be improved in order to provide an empirical match with streamflow data.

The results (Figure 11) depend on the modeling standard used, and on the model structure tested. Looking first at the lower of the two standards (Standard 1), for some model structures (e.g., GR4J, GR4JMOD, SACRAMENTO) the cases are relatively evenly split between cases (a), (b) and (c). This means that it was just as common for failure in a DSST to be the result of the calibration method as it was the result of the model structure. Thus, with regards to the hypothesis, both the models and the calibration methods need improvement in order to successfully model changing climatic conditions.

The IHACRES model structure once again provides an interesting case study. IHACRES was able to attain Standard 1 in a very high proportion of catchments: 74 out of 86 (i.e., 37 catchments in category (a) plus 37 catchments in category (b)). This would suggest that the model structure itself is relatively well suited to simulating changes in climate and does not require change to provide an empirical match with data. However, of the 74 that were successfully modeled by IHACRES, the single-objective DSST was only able to find a suitable parameter set in 37 cases (category a). The remainder (category b) were catchments where AMALGAM found a suitable parameter set but the single-objective DSST did not. This is not a particularly favourable success rate for the DSST, and suggests the need to review the use of single objective optimization methods in model calibration.

However, the interpretation shifts if Standard 2 is adopted instead of Standard 1. In this case the number of catchments where the modeling standard is not met is around 50% in the case of GR4JMOD and SACRAMENTO, and greater for GR4J and SIMHYD. The IHACRES model is able to meet this modeling standard in 59% of cases (24 + 27=51 out of 86 catchments) compared to 87% (74) for Standard 1. Thus, if this higher

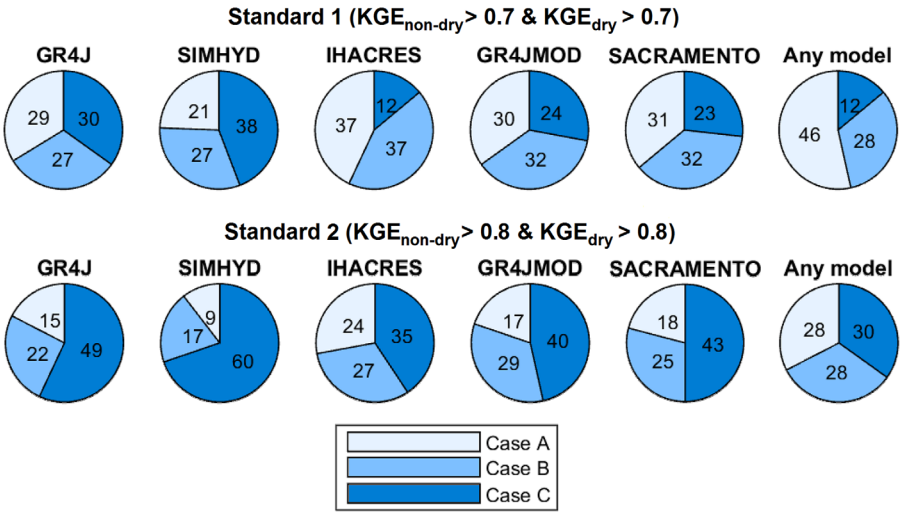

Figure 11. Effectiveness of the single-objective Differential Split Sample Test (DSST) and AMALGAM in finding parameter sets that meet the two performance standards. Case A means suitable parameter sets were found by *both* the DSST and AMALGAM; Case B means suitable parameter sets were *not* found by the DSST but *were* found by AMALGAM; and Case C means neither method found suitable parameter sets. "Any model" means that the modeller able to apply all five of the model structures and has the freedom to adopt the best model whatever it may be.

standard is adopted, one possible conclusion is that the current generation of rainfall runoff model structures, including IHACRES, require improvements to simulate changes in climatic conditions in order to produce an empirical match with data. An alternative explanation is that the failure to attain the modeling standard is due to data errors (section 4.3).

The pie charts to the far right of Figure 11 present the results in the case where a modeller is able to apply all five of the model structures to every catchment and has the freedom to adopt the best model whatever it may be. In this case, suitable parameter sets are found during the single-objective Differential Split Sample Test in 53% of catchments (46 out of 86) for Standard 1, and 33% of catchments (28 out of 86) for Standard 2. There still remains a significant portion of catchments that are not modeled satisfactorily by any of the 5 model structures: 12 catchments out of 86 (14%) in the case of Standard 1, and 30 catchments out of 86 (35%) in the case of Standard 2.

3.4. Examination of Catchments Where Models Failed

We examined those catchments where the model structures failed to meet Standard 1 and/or Standard 2. Two main avenues were explored: first, we analyzed the Pareto Curves and considered what the form of these curves may indicate about the type of model failure; and second we examined the physical and climatic properties of these catchments. For brevity, some elements of this discussion are summaries only, with a reference to the supporting information for more detail.

Having failed to meet the standard, every instance considered was one where no single parameter set could simulate flows satisfactorily in both wet and dry periods. However, we categorized failure instances further according to whether or not the model structure was able to meet the standard in a given objective when optimized to it in isolation. The results varied by model type: for example, GR4J and GR4JMOD were exceptionally good at meeting the modeling standards in a given objective provided that they were calibrated to it in isolation; i.e., the maximum possible value in each objective was high, but there was also a high degree of tradeoff in between these endpoints. We categorized this type of failure with the phrase "Model structure can simulate both dry and wet periods, but not with the same parameter set." In contrast, this type of failure was not common with IHACRES and SIMHYD, particularly for Standard 2, in which the category "Appears to be deficient in this catchment, regardless of climate" claimed the highest proportion of failures. The full results of this analysis are shown in the supporting information Figure S7.

Next, we focussed on the physical properties, including location, of the catchments where the model structures failed. For this analysis, we examined only those catchments where none of the model structures were

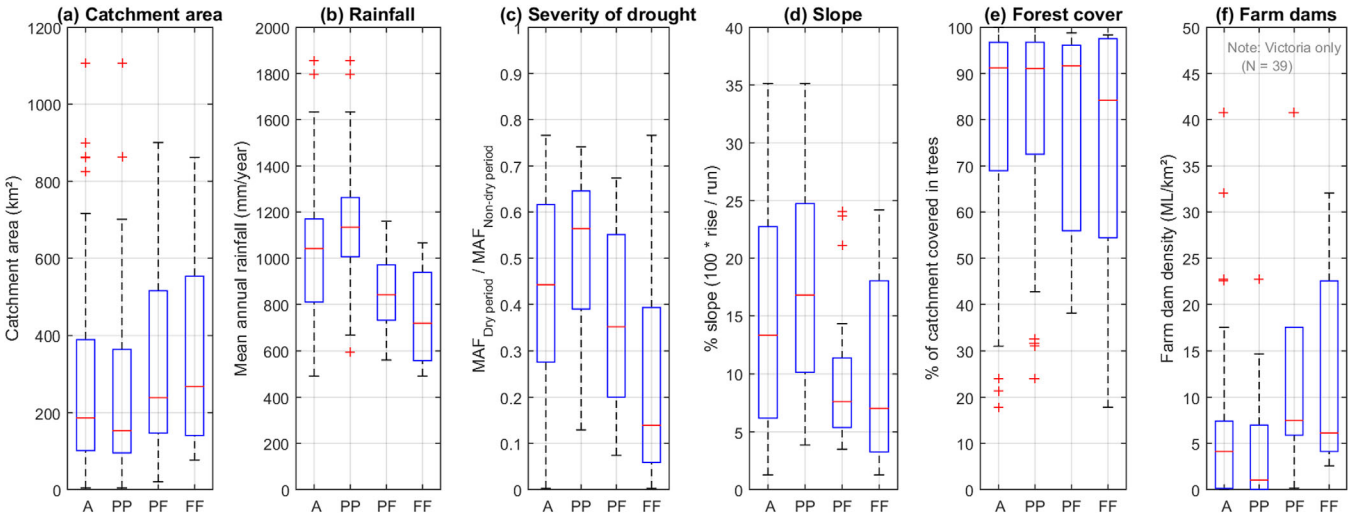

Figure 12. Physical characteristics of catchments where model structures failed one or both standards. Four boxplots are shown for each characteristic: all catchments (marked A; N = 86); cases where both standards were passed (marked PP; N = 56); cases where Standard 1 was passed but Standard 2 was failed (marked PF; N = 18); and cases where neither standard was met (marked FF; N = 12). Farm dam data were only available for Victoria, so that the N values are different in plot f ($N_A = 39$, $N_{PP} = 27$, $N_{PF} = 6$ and $N_{FF} = 6$). The whiskers extend a maximum of 1.5 times the interquantile range. Values beyond the whiskers are marked as outliers and are denoted as +.

able to meet the required standard. As per Figure 11, there were 12 such catchments in the case of Standard 1 (labeled "FF" since they failed both standards) and a further 18 in the case of Standard 2 (labeled "PF" since they passed one standard and failed the other).

In terms of geographic location, the instances of model failure are relatively well dispersed. In terms of failure to meet Standard 1, there appeared to be two regions where model structures were more likely to fail: the central part of the state of Victoria, and the northern-most catchments tested in the state of Queensland. There were also a number of catchments failing Standard 2 in the eastern highlands of New South Wales. A map is provided in the supporting information (supporting information Figure S6).

Figure 12 shows the physical characteristics of catchments where model structures failed one or both standards. We selected five characteristics for testing, based on their perceived importance to hydrology: catchment area; rainfall; slope; forest cover; and degree of development of private farm dams. Soil type and geology are also perceived to be important, but there are few high-quality national soil type/geology data sets that are numerical (i.e., noncategorical). In addition to the five characteristics above, the observed severity of drought was also included, measured as the ratio of mean annual flows during the dry period to mean annual flows during the nondry period. From Figure 12a, catchment area appears to have little bearing on the failure of the model structures. However, Figures 12b and 12c show that cases of model structure failure tended to be in drier catchments, and where flow reductions during Period 2 were greatest.

To further investigate these results, we applied the nonparametric one-sided Rank-Sum Test, otherwise known as the Wilcoxon-Mann-Whitney test or the Mann-Whitney U test (as described by e.g., *Wilks* [2011], see also *Wilcoxon* [1945] and *Mann and Whitney* [1947]). This evaluates the probability of the null hypothesis that two groups of data (in this case, characteristics of catchments where a modeling standard was/was not met, respectively) came from the same underlying distribution. By concentrating only on relative ranks rather than actual values, this test resists being influenced by one or two extreme values, which is important because some catchment characteristics have quite skewed distributions. The results (Table 2) confirmed that the catchments where the modeling standards were not met tended to be those with lower rainfall, lower slope and a greater relative reduction in flow during the 7 driest consecutive years. Catchment area was less strongly related to modeling performance than these three, and forest cover less so again.

Since the group that failed Modeling Standard 1 (Group FF) is such a small sample size, we provide a catchment-by-catchment list of characteristics for each member of group FF in supporting information Table S1. Inspection of the individual characteristics of group FF reveals that although there appears to be differences

Table 2. Results From the Nonparametric Rank-Sum Test to Test Whether Catchment Characteristics Differed Between Catchments Where a Given Modeling Standard Was Not Met (by Any Model Structure) and Those Where It Was[a]

	Relating to Modeling Standard 1		Relating to Modeling Standard 2	
	p Value	Significant at 95% Level?	p Value	Significant at 95% level?
Catchment area	0.0953	no	0.0304	yes
Mean annual rainfall	0.0002	yes	<0.0001	yes
Dry period flow ratio	0.0003	yes	<0.0001	yes
Catchment average slope	0.0165	yes	0.0001	yes
Forest Cover	0.4161	no	0.2343	no
Farm Dam Development	0.0553	no	0.0041	yes

[a]Columns two and four indicate the probability that the observed differences in characteristics between the two groups of catchments arose purely by chance.

between the boxplots for catchment average slope and forest cover, the reality is more complex, with group FF being spread across a relatively wide range in both cases.

In terms of farm dams, estimates of farm dam volume were only available for catchments in the State of Victoria (N = 39). Two of the three catchments where farm dam density exceeds 20 ML/km^2 were catchments where modeling standard 1 was not met, and it is possible that harvesting of water by farm dams in these catchments is causing difficulties in modeling. However, the other catchments had much lower levels of development of farm dams so it is unlikely that farm dams are degrading model performance in these catchments. Further research is required to investigate whether rainfall-runoff modeling in the two catchments with farm dam density exceeding 20 ML/km^2 might be aided by quantification of farm dam interception.

The results of this study are partially consistent with recent findings of *Saft et al* [2015] who analyzed changes to the relationship between rainfall and runoff on an annual time step, in the same study area. They found that changes to the relationship were more likely in drier catchments (upheld here) with low slope (upheld here) and low forest cover (not upheld here, although the catchments used in this study generally had greater forest cover than those in *Saft et al.* [2015]). Note that although bushfires are relatively common throughout Australia, we could not find any evidence linking bushfire history with the failure of models to attain the modeling standards (supporting information Text S2).

4. Discussion

4.1. Results Summary

Although the results above are specific to the catchments, data, models and objective functions used, they are potentially relevant to any study that has rejected a model structure based on a poor match with streamflows in an independent evaluation period (e.g., a DSST). The results show that a significant proportion of such rejections may be spurious because parameter sets may exist that fulfil a given set of performance criteria but remain undetected during calibration. Thus, poor performance in evaluation in a split sample test is a poor basis on which to reject a model hypothesis, although it is adequate for rejecting the model/calibration method combination.

As noted in the method section, the Pareto framework used here was intended only to critically assess existing methods of model calibration and evaluation; in this paper we are not suggesting that the method should be adopted for use in rainfall-runoff model calibration. The reasons for this are explained below (sections 4.3 and 4.7).

4.2. Getting the Right Answers for the Wrong Reasons?

We now consider whether or not a parameter set or model structure that is found to fulfil the adopted KGE performance criteria (i.e., get the "right answer") can be considered "adequate" or "valid." First, it is widely acknowledged that one performance criteria (e.g., KGE) is insufficient to ensure a holistic match with observed flows [*Oudin et al.*, 2006; *Gupta et al.*, 2009; *Berthet et al.*, 2010; *Andreassian et al.*, 2012], even if jointly considered over two contrasting periods as demonstrated here. As an example, consider the progression of modeling bias with time for the parameter sets shown in Figure 13. Even though long-term bias is a

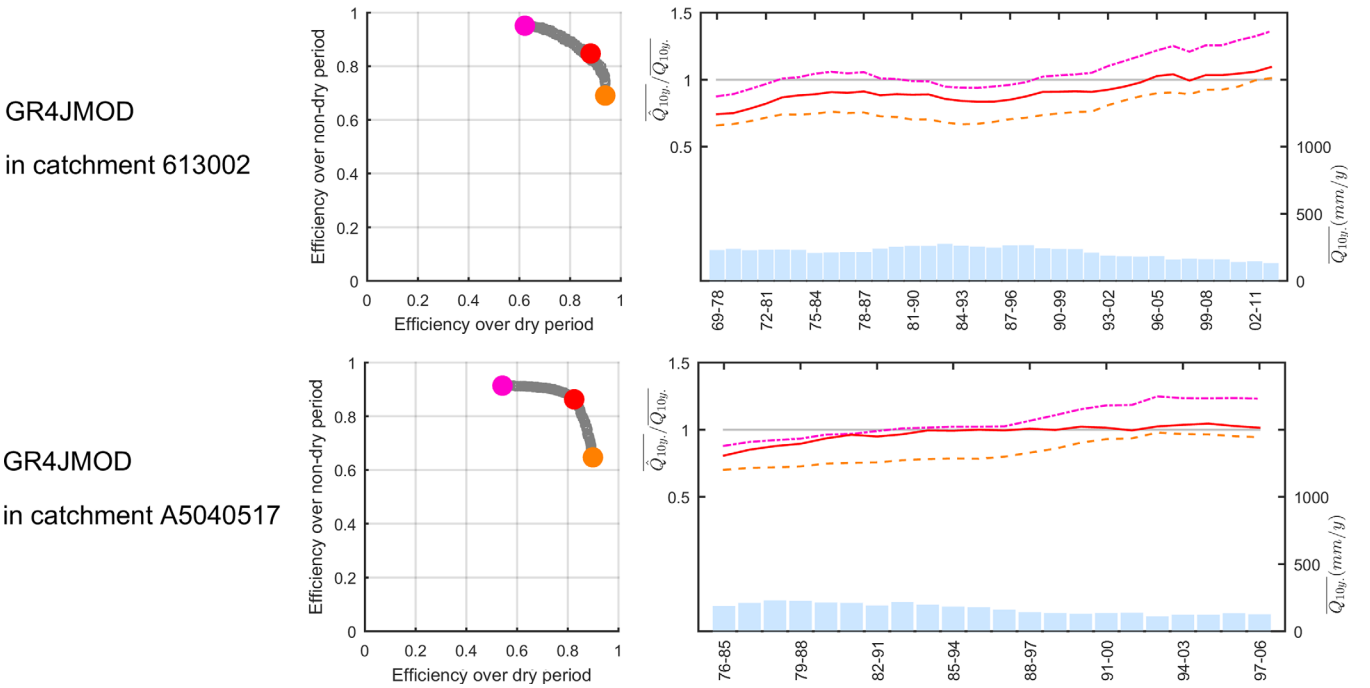

Figure 13. Long-term simulation bias (right) for three selected parameter sets (left), after *Coron et al.* [2014, Figure 5), for two selected case studies. Simulation bias is plotted as a 10 year moving average, and the 10 year moving average streamflows are also plotted for reference, in blue. The case study catchments are 613002 (Harvey River at Dingo Road, Western Australia; 147.5 km^2; mean annual rainfall = 992 mm/yr; rainfall-runoff ratio 0.21) and A5040517 (First Creek at Waterfall Gully, South Australia; 5.1 km^2; mean annual rainfall = 992 mm/yr; rainfall-runoff ratio 0.19). Despite similar positions in objective space for the red parameter set, changes in long term average streamflow are more faithfully tracked in A5040517 than 613002.

component of the KGE, the 10 year rolling average bias still deviates considerably from zero for the three parameter sets shown in each case, and shows some similarity with the results of *Coron et al.* [2014], particularly in catchment 613002. Choosing a parameter set that performs well in both periods (red) does not guarantee unbiased simulations over the modeling period, although GR4JMOD performs better in this aspect in the second case (A5040517) than the first (613002). This analysis of bias, based on the format of *Coron et al.* [2014], is shown for other selected case studies in the supporting information (supporting information Figures S8–S11). From these examples it is clear that a high KGE score may mask underlying discrepancies in matching the observed data. Furthermore, even a near-perfect match with observed streamflows would not necessarily imply that a rainfall runoff model is "adequate" or "valid," depending on the philosophical viewpoint. As discussed in the Introduction, a near-perfect match with observed streamflows corresponds to adequacy in an operational or "engineering" sense [*Gupta et al.*, 2012] but a "physical science" approach would ask whether the model is getting the right answers for the right reasons [*Kirchner*, 2006; *Gupta et al.*, 2012]. Under this viewpoint, models are adequate only if consistent with dominant physical processes. As noted in the introduction, this is difficult to test in practice for a large sample of catchments, and thus we do not assess the adequacy of models in this physical science sense. Given that some processes that are thought to be important are not represented by some conceptual models used in this study (e.g., interception in the case of IHACRES––*Jakeman and Hornberger* [1993]; *Savenije* [2004]) it is unlikely that such models could be considered adequate in the physical science sense, regardless of their goodness of fit.

4.3. The Role of Data Errors

Data errors are ubiquitous in hydrology and can confound the results of hydrologic studies. For example, for the data used in this study, the streamflow data are subject to uncertainty in the stage-discharge relationship [*McMillan et al.*, 2010], while the gridded rainfall data are subject to measurement error in the underlying point rainfall data [e.g., *Nešpor and Sevruk*, 1999] plus interpolation error associated with creating a spatial grid of values based on point measurements [*Jones et al.*, 2009; *Tozer et al.*, 2012].

Although optimization to a single performance measure (e.g., KGE or NSE) remains common in practice, during optimization the mathematical compensation for input and output errors can lead to spurious

results [*Thyer et al.*, 2009]. The mathematically optimum parameter set is actually a function of the input and output errors, and a different set of errors may result in an entirely different "optimum" set. In this paper, since the input and output errors were not explicitly accounted for, the Pareto Fronts generated are similarly a function of the errors in the input and output data. The complex interactions of model structural error with input and output error further complicate the situation [*Renard et al.*, 2010].

The uncertainty in model inputs and flow data propagate through to uncertainty in parameters and projections, and this can be quantified in various ways [e.g., *Beven and Binley*, 1992; *Freer et al.*, 1996; *Kavetski et al*, 2006a,b; *Renard et al.*, 2010, 2011]. Common methods identify not a single parameter set (as in optimization) but an ensemble of parameter sets, which together are consistent with knowledge of input and output uncertainty, and allow quantification of uncertainty through consideration of multiple possible model simulations.

We affirm that the quantification of uncertainty is an important aspect of any study aiming to provide model projections or forecasting to inform decision making. In contrast, the aim of this study was to revisit the conclusion that rainfall runoff models suffer from poor performance if applied in climatic conditions different to those against which they were calibrated. Given that previous studies have used single objective optimization and the DSST to make conclusions about model validity [*Vaze et al.*, 2010] and parameter stationary [*Merz et al.*, 2011], we tailored our method to specifically investigate how reliable the outcomes of such tests may be. The Pareto approach proved useful in this context, but we reiterate that the method used here is not recommended as a general calibration method, in part due to its inability to estimate predictive uncertainty.

4.4. Relevance to Future Model Improvements

The results of this study are instructive toward future efforts to improve rainfall runoff models. The key lesson for model improvements is this: where improvements are trialed, it is possible that their full benefit will not be seen if evaluated using the DSST in isolation, due to the chance of false negatives. This was shown very clearly (see Figure 9 and supporting information Figure S5) for the comparison between GR4J [*Perrin et al.*, 2003] and the modified version GR4JMOD by *Hughes et al.* [2013]. Numerous cases were found where the DSST led to a false conclusion of negligible benefit from the changes of *Hughes et al.* [2013].

Some studies, such as *Brigode et al.* [2013] demonstrated a DSST using a method (e.g., DREAM––*Vrugt et al.* [2008]) that generated an ensemble of parameter sets. Because such ensemble methods inherently provide information about a wider range of parameter sets, they may be more likely to identify sets that demonstrate the true capabilities resulting from a model improvement. However, this depends strongly on details of methodology, with a key choice being whether or not to explicitly represent the uncertainty of inputs and outputs [e.g., *Renard et al.*, 2010, 2011] or adopt objective functions that compensate for these errors without representing them explicitly (as adopted by *Brigode et al.*, 2013, cf. *Schoups and Vrugt*, 2010].

As discussed above, in this study we did not account for data errors, and so instances of apparent model failure may be related to cases of particularly poor data quality. However, we observed tendencies among catchments where failure was common––namely, they tended to be drier, flatter and have more severe droughts [see also *Saft et al.*, 2015]––and these systematic tendencies support the case for research to better simulate flow generation mechanisms in such catchments, as opposed to assuming that all remaining deficiencies are the result of data errors [*Brigode et al.*, 2013].

4.5. Minimizing False Negatives

This paper has demonstrated that DSST results may provide a false negative impression of the capabilities of a model. Geometrically, this is associated with Pareto Fronts that had an "inverted L" shape, intersecting regions of robust performance (e.g., shaded regions of Figure 10), but with endpoint(s) distant from these regions (e.g. IHACRES in Figure 10). Shapes less prone to false negatives included Pareto curves that formed quasi-linear diagonal lines in the objective space (e.g., GR4J in Figure 10) and the ideal case (in the sense of parameter stationarity) where the Pareto curve is so compact as to appear as a dot in the objective space (e.g., GR4JMOD in 410057, supporting information Figure S1).

It is difficult to generalize about the relation between model complexity (number of parameters) and the tendency to produce false negatives. In a separate analysis (not shown) we examined the shape of the

Pareto curves on a model-by-model basis, which demonstrated that the parsimonious model GR4J tended to produce Pareto Fronts of the "quasi linear diagonal" type, and thus have a lower tendency to generate false negative impressions of model capabilities. However, higher model complexity did not necessarily lead to more false negatives, as shown by a comparison of IHACRES (8 parameters, 37 false negatives for Standard 1) and SACRAMENTO (13 free parameters, 32 false negatives for Standard 1).

It is also possible that careful selection of objective functions may minimize false negatives. In the ideal case listed above (Pareto curve collapsed to a dot in the Objective Space), the parameter set identified as optimal in one set of climatic conditions is optimal or near-optimal in other climatic conditions––a desirable attribute for an objective function and/or model structure. In the present context, the tendency to produce this ideal case could be evaluated for a given objective function either by (a) assessing only the endpoints of the Pareto curves (one-at-a-time single objective optimization [cf. *Coron et al.*, 2012, 2014]); or (b) via full Pareto analysis as shown in this paper. Future research could conduct this analysis individually for a number of objective functions from the literature in turn, and then compare the results. It is likely that a more nuanced objective function such as a meta-function incorporating responses over multiple time scales [*Hartmann and Bárdossy*, 2005; *Shamir et al.*, 2005] may have more success than commonly used functions that consider only the daily time step. Such analysis would be relevant to the discussion of the value (or lack of value) of single objective optimization in hydrology [e.g., *Gupta et al.*, 2008].

4.6. Climate Change: Beyond the Scope of Historical Observations?

While climate change may be outside of the range of current observations in many regions of the world, in South East Australia the changes in streamflow projected in some climate change studies are of a similar order to the historic streamflow declines during the Millennium Drought. For example, *Chiew et al.* [2009] projected future runoff across South East Australia using the outputs of 15 Global Climate Models (GCMs). Although there was a high degree of uncertainty, in most locations and for all GCMs the percentage change in long-term average annual flows was generally less than 55% (ibid. Figure 9), which was the median observed reduction during the Millennium Drought for the catchments in this study (section 2.6, cf. Figure 5). However, *Chiew et al.* [2009] used GCM runs based on a 0.9°C increase in temperature, and scenarios with greater temperature increases would result in greater reductions in streamflow that may be beyond the range of observations. Nonetheless, we suggest that it is reasonable to assume that the observed behaviour of catchments during historic dry periods like the Millennium Drought can be used to inform our understanding of possible future behaviour of these catchments under climate change.

4.7. Research Challenges

In this section, we summarize the research challenges to improve rainfall runoff modeling under a changing climate. These are not original ideas; rather, we aim to relate the present study to existing ideas and trends in the literature. We broadly group the challenges under two headings:

1. *Making better use of information content of measured data*: Figure 13 [see also e.g., *Oudin et al.*, 2006; *Gupta et al.*, 2009; *Berthet et al.*, 2010; *Andreassian et al.*, 2012] demonstrated that the use of global performance measures can mask significant deficiencies in simulations. Hydrologists should therefore favour measures that consider a breadth of characteristics about the historic data. The multitimescale objective functions mentioned above (section 4.5) are an example of this. We also note developments in using hydrologic signatures to inform calibration [*Wagener and Montanari*, 2011; *Vrugt and Sadegh*, 2013]. While signatures do not inherently take data errors into account, some signatures are less sensitive to data errors than others [*Westerberg and McMillan*, 2015], so that signature sets can be chosen with the intent of reducing the confounding effect of data errors [*Vrugt and Sadegh*, 2013] while maximizing the information content gained from observed data. This paper has demonstrated that some existing model structures were capable of simulations that provided robust performance before and after a change in climate. The challenge is to develop calibration methods that can identify these parameter sets using only "prechange" data. The Pareto method used here does not do this, and furthermore is not viable if the changed climate has not yet been observed.
2. *Improving process understanding in catchments under change*: Following the same logic as above, it may be that even with considerable advances in parameter estimation methods, it is still not possible to identify robust parameter sets using only "prechange" data. Further research is needed to investigate the physical reasons why runoff is more sensitive to changes in rainfall than current rainfall runoff models

would suggest. Such research would be consistent with the current research focus on change in hydrology and society (IAHS Panta Rhei decade 2013–2022––*Montanari et al.* [2013]). This knowledge could inform new rainfall runoff model(s) that, when calibrated to "prechange" data, would ideally provide more certainty about the trajectory of runoff after a change in rainfall, and be closer to "adequate" in both a physical science and engineering sense. However, it is noted that even if a model does have the correct structure to simulate flows under contrasting conditions, the relevant parameters may remain poorly identified during calibration [*Reichert and Omlin*, 1997], depending on the input data and method of calibration.

5. Conclusions and Recommendations

In this paper, five conceptual rainfall-runoff model structures were tested in 86 catchments, initially using a Differential Split Sample Test (DSST) that was intended to replicate common practice. When optimized to match the Kling-Gupta efficiency over the nondry period, the models generally had poor performance during the dry period, and vice versa. These results were consistent with existing literature [e.g., *Vaze et al*, 2010; *Coron et al*., 2012, 2014; *Thirel et al*., 2015b]. Therefore, the model structures largely failed the DSST, although this was catchment dependent. The model structures were then further tested using a Pareto approach via the AMALGAM algorithm. The AMALGAM results demonstrated that many of the cases of apparent failure under the DSST were false negatives. Of the 282 cases of apparent model failure under the DSST using the lower modeling standard (KGE $>$ 0.7), 155 were false negatives. The higher standard (KGE $>$ 0.8) resulted in 347 cases of apparent model failure, with 120 false negatives. Thus, regardless of the standard used, the DSST often missed potentially promising parameter sets within a given model structure.

These results can be used to answer the research question and hypothesis stated at the beginning of the paper. Responding to the recorded deficiencies of rainfall runoff model performance in the literature, the research question was: *Are current conceptual rainfall runoff model structures deficient in their ability to simulate streamflow responses to long term changes in climate?* The hypothesis to be tested was that *the observed poor performance is due to poor or insufficient model calibration and evaluation techniques rather than deficient model structures*. The results indicate that this hypothesis was true in around 55% of the cases (155 out of 282) or around 35% of the cases (123 out of 349) of poor performance in the DSST, depending on the modeling standard adopted. Thus, the answer to the research question is that *some* rainfall runoff model structures are deficient in *some* catchments, with the corollary that the deficiency is significantly less common than the Differential Split Sample Test might suggest. It was discussed that the definitions of "deficient" and "adequate" are themselves dependent on philosophical perspective [*Gupta et al*., 2012].

As noted throughout the paper, we are not proposing that the multiobjective approach trialed here is a viable alternative approach to the DSST. The logic expounded by *Klemeš* [1986] is valid and we affirm the need to withhold a portion of historic data for independent testing and evaluation. The multiobjective approach here does not do this, so the findings of this paper are based solely on calibration results, with no independent evaluation period. The Pareto approach trialed here is only useful insofar as it has demonstrated that commonly used model calibration and evaluation methods can give a false negative impression of the ability of a model structure to match observed streamflow.

Based on our results and discussion, we recommend:

1. *Caution when interpreting split sample results*. Split sample testing remains an essential test of models that will be used operationally (in the sense of *Klemeš* [1986]) and a useful "first test" of a model structure's capabilities. However, this paper has demonstrated that split sample test results can give a false negative impression of the ability of a model to match observed streamflow, and are thus a poor basis to reject a model hypothesis.
2. *Further work toward identifying parameter sets that are robust to changes in climate*. This paper has demonstrated that commonly used calibration and evaluation methods often fail to identify parameter sets that can simulate flows robustly when climatic conditions change, even when such parameter sets do exist within a model structure. New methods are needed that can more reliably identify such parameter sets.

3. *Further research aimed at understanding the physical processes occurring in catchments when climatic conditions change*, in line with the IAHS Panta Rhei Decade's focus on change in hydrology and society [*Montanari et al.*, 2013].

Acknowledgments
The authors gratefully acknowledge the support of the Australian Government in carrying out this work. Specifically, Keirnan Fowler's work was supported by an Australian Postgraduate Award and Murray Peel is the recipient of an Australian Research Council Future Fellowship (FT120100130). 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.*, 2011], https://www.longpaddock.qld.gov.au/silo/. The authors appreciate the work of the three anonymous reviewers and one Associate Editor whose feedback greatly improved the quality of the article.

References

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

Andréassian, V., C. Perrin, L. Berthet, N. Le Moine, J. Lerat, C. Loumagne, L. Oudin, T. Mathevet, M.-H. Ramons, and A. Valéry (2009), Crash tests for a standardized evaluation of hydrological models, *Hydrol. Earth Syst. Sci.*, *13*, 1757–1764.

Andreassian, V., N. Le Moine, C. Perrin, M. H. Ramos, L. Oudin, T. Mathevet, J. Lerat, and L. Berthet (2012), All that glitters is not gold: The case of calibrating hydrological models, *Hydrol. Processes*, *26*, 2206–2210.

Andrews, F. (2013), R code repository for HYDROMAD. [Available at http://hydromad.catchment.org/, last accessed 30 Mar. 2015.]

Arsenault, R., A. Poulin, P. Côté, and F. Brissette (2013), Comparison of stochastic optimization algorithms in hydrological model calibration, *J. Hydrol. Eng.*, *19*(7), 1374–1384.

Bárdossy, A. (2007), Calibration of hydrological model parameters for ungauged catchments, *Hydrol. Earth Syst. Sci.*, *11*(2), 703–710.

Bárdossy, A., and S. K. Singh (2008), Robust estimation of hydrological model parameters, *Hydrol. Earth Syst. Sci.*, *12*, 1273–1283.

Bekele, E. G., and J. W. Nicklow (2007), Multi-objective automatic calibration of SWAT using NSGA-II, *J. Hydrol.*, *341*(3), 165–176.

Berthet, L., V. Andréassian, C. Perrin, and C. Loumagne (2010), How significant are quadratic criteria? Part 2. On the relative contribution of large flood events to the value of a quadratic criterion, *Hydrol. Sci. J.*, *55*(6), 1063–1073.

Beven, K., and A. Binley (1992), The future of distributed models: Model calibration and uncertainty prediction, *Hydrol. Processes*, *6*(3), 279–298.

Booij, M. J., and M. S. Krol (2010), Balance between calibration objectives in a conceptual hydrological model, *Hydrol. Sci. J.*, *55*(6), 1017–1032.

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?, *J. Hydrol.*, *476*, 410–425.

Budyko, M. (1971), *Climate and Life*, Academic, New York.

Burnash, R. J. C., R. L. Ferral, and R. A. McGuire (1973), *A Generalized Streamflow Simulation System––Conceptual Modelling for Digital Computers*, 204 pp. Joint Fed.-State River Forecast Cent., Sacramento.

Chiew, F. H. S., and T. A. McMahon (1994), Application of the daily rainfall-runoff model MODHYDROLOG to 28 Australian catchments, *J. Hydrol.*, *153*(1), 383–416.

Chiew, F. H. S., M. C. Peel, and A. W. Western (2002), Application and testing of the simple rainfall-runoff model SIMHYD, in *Mathematical Models of Small Watershed Hydrology and Applications*, edited by V. P. Singh and D. K. Frevert, pp. 335–367, Water Resour. Publ., Littleton, Colo.

Chiew, F. H. S., J. Teng, J. Vaze, D. A. Post, J. M. Perraud, D. G. C. Kirono, and N. R. Viney (2009), Estimating climate change impact on runoff across southeast Australia: Method, results, and implications of the modeling method, *Water Resour. Res.*, *45*, W10414, doi:10.1029/2008WR007338.

Chiew, F. H. S., N. J.Potter, J. Vaze, C. Petheram, L. Zhang, J. Teng, and D. A. Post (2014), Observed hydrologic non-stationarity in far south-eastern Australia: Implications for modelling and prediction, *Stochastic Environ. Res. Risk Assess.*, *28*(1), 3–15.

Choi, H. T., and K. Beven (2007), Multi-period and multi-criteria model conditioning to reduce prediction uncertainty in an application of TOPMODEL within the GLUE framework, *J. Hydrol.*, *332*(3), 316–336.

Cieniawski, S. E., J. W. Eheart, and S. Ranjithan (1995), Using genetic algorithms to solve a multiobjective groundwater monitoring problem, *Water Resour. Res.*, *31*(2), 399–409.

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

Coron, L., V. Andréassian, C. Perrin, M. Bourqui, and F. Hendrickx (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.

Covey, C., K. M. AchutaRao, U. Cubasch, P. Jones, S. J. Lambert, M. E. Mann, T. J. Phillips, and K. E. Taylor (2003), An overview of results from the Coupled Model Intercomparison Project, *Global Planet. Change*, *37*(1), 103–133.

Deb, K., A. Pratap, S. Agarwal, and T. A. M. T. Meyarivan (2002), A fast and elitist multiobjective genetic algorithm: NSGA-II, *IEEE Trans. Evol. Comput.*, *6*(2), 182–197.

Department of Environment, Land, Water and Planning (2015a), Farm Dam Boundaries, Victorian Government, Melbourne, Australia. [Available at https://www.data.vic.gov.au/data/dataset/farm-dam-boundaries, last accessed 26 May 2015.]

Department of Environment, Land, Water and Planning (2015b), Farm Dam Points, Victorian Government, Melbourne, Australia. [Available at https://www.data.vic.gov.au/data/dataset/farm-dam-points, last accessed 26 May 2015.]

Donohue, R. J., T. R. McVicar, and M. L. Roderick (2010), Assessing the ability of potential evaporation formulations to capture the dynamics in evaporative demand within a changing climate, *J. Hydrol.*, *386*(1), 186–197.

Duan, Q., S. Sorooshian, and V. Gupta (1992), Effective and efficient global optimization for conceptual rainfall-runoff models, *Water Resour. Res.*, *28*(4), 1015–1031.

Duan, Q., S. Sorooshian, and V. Gupta (1994), Optimal use of the SCE-UA global optimization method for calibrating watershed models, *J. Hydrol.*, *158*(3), 265–284.

Efstratiadis, A., and Koutsoyiannis, D. (2010), One decade of multi-objective calibration approaches in hydrological modelling: A review, *Hydrol. Sci. J.*, *55*(1), 58–78.

Forster, P., et al. (2007), Changes in atmospheric constituents and in radiative forcing, in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited S. Solomon et al., Cambridge Univ. Press, Cambridge, U. K.

Freer, J., K. Beven, and B. Ambroise (1996), Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach, *Water Resour. Res.*, *32*(7), 2161–2173.

Gallant, J. C., T. I. Dowling, A. M. Read, N. Wilson, P. Tickle, and C. Inskeep (2011), 1 second SRTM derived digital elevation models user guide, Geoscience Australia report, Geoscience Australia, Canberra, Australia. [Available at www.ga.gov.au/topographic-mapping/digital-elevation-data.html.]

Gharari, S., M. Hrachowitz, F. Fenicia, and H. H. G. Savenije (2013), An approach to identify time consistent model parameters: Sub-period calibration, *Hydrol. Earth Syst. Sci.*, *17*(1), 149–161.
Gupta, H. V., S. Sorooshian, and P. O. Yapo (1998), Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, *Water Resour. Res.*, *34*(4), 751–763.
Gupta, H. V., T. Wagener, and Y. Liu (2008), Reconciling theory with observations: Elements of a diagnostic approach to model evaluation, *Hydrol. Process.*, *22*(18), 3802–3813.
Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez (2009), Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *J. Hydrol.*, *377*(1), 80–91.
Gupta, H. V., M. P. Clark, J. A. Vrugt, G. Abramowitz, and M. Ye (2012), Towards a comprehensive assessment of model structural adequacy, *Water Resour. Res.*, *48*, W08301, doi:10.1029/2011WR011044.
Gupta, H. V., C. Perrin, G. Bloschl, A. Montanari, R. Kumar, M. Clark, and V. Andréassian (2014), Large-sample hydrology: A need to balance depth with breadth, *Hydrol. Earth Syst. Sci.*, *18*(2) 463–477.
Haario, H., E. Saksman, and J. Tamminen (2001), An adaptive Metropolis algorithm, *Bernoulli*, *7*, 223–242.
Hansen, N., S. Müller, and P. Koumoutsakos (2003), Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES), *Evol. Comput.*, *11*(1), 1–18.
Hartmann, G., and A. Bárdossy (2005), Investigation of the transferability of hydrological models and a method to improve model calibration, *Adv. Geosci.*, *5*(5), 83–87.
Hughes, J., R. Silberstein, and A. Grigg (2013), Extending rainfall-runoff models for use in environments with long–term catchment storage and forest cover changes, in *MODSIM2013, 20th International Congress on Modelling and Simulation*, edited by J. Piantadosi, R. S. Anderssen, and J. Boland, pp. 2471–2477, Modelling and Simulation Society of Australia and New Zealand, Canberra, Australia.
Hughes, J. D., K. C. Petrone, and R. P. Silberstein (2012), Drought, groundwater storage and stream flow decline in southwestern Australia, *Geophys. Res. Lett.*, *39*, L03408, doi:10.1029/2011GL050797.
Hurst, H. E. (1951), Long-term storage capacity of reservoirs, *Trans. Am. Soc. Civ. Eng.*, *116*, 770–808.
Jakeman, A. J., and G. M. Hornberger (1993), How much complexity is warranted in a rainfall runoff model?, *Water Resour. Res.*, *29*(8), 2637–2649.
Jakeman, A. J., I. G. Littlewood, and P. G. Whitehead (1990), Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments, *J. Hydrol.*, *117*(1), 275–300.
Jeffrey, S. J., J. O. Carter, K. B. Moodie, and A. R. Beswick (2001), Using spatial interpolation to construct a comprehensive archive of Australian climate data, *Environ. Modell. Software*, *16*(4), 309–330.
Jones, D. A., W. Wang, and R. Fawcett (2009), High-quality spatial climate data-sets for Australia, *Aust. Meteorol. Oceanogr. J.*, *58*(4), 233–248.
Kavetski, D., G. Kuczera, and S. W. Franks (2006a), Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, *Water Resour. Res.*, *42*, W03407, doi:10.1029/2005WR004368.
Kavetski, D., G. Kuczera, and S. W. Franks (2006b), Bayesian analysis of input uncertainty in hydrological modeling: 2. Application, *Water Resour. Res.*, *42*, W03408, doi:10.1029/2005WR004376.
Kennedy, J., R. C. Eberhart, and Y. Shi (2001), *Swarm Intelligence*, 512 p., Morgan Kaufmann, San Francisco, Calif.
Kim, H. S., and S. Lee (2014), Assessment of a seasonal calibration technique using multiple objectives in rainfall–runoff analysis, *Hydrol. Processes*, *28*(4), 2159–2173.
Kirchner, J. W. (2006), Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, *Water Resour. Res.*, *42*, W03S04, doi:10.1029/2005WR004362.
Klemeš, V. (1986), Operational testing of hydrological simulation models, *Hydrol. Sci. J.*, *31*(1), 13–24.
Kollat, J. B., P. M. Reed, and T. Wagener (2012), When are multiobjective calibration trade-offs in hydrologic models meaningful?, *Water Resour. Res.*, *48*, W03520, doi:10.1029/2011WR011534.
Li, C. Z., L. Zhang, H. Wang, Y. Q. Zhang, F. L. Yu, and D. H. Yan (2012), The transferability of hydrological models under nonstationary climatic conditions, *Hydrol. Earth Syst. Sci.*, *16*(4), 1239–1254.
Lymburner, L., P. Tan, N. Mueller, R. Thackway, A. Lewis, M. Thankappan, L. Randall, A. Islam, and U. Senarath (2011), The National Dynamic Land Cover Dataset, Record 2011/31, Geosci. Aust. and the Bur. of Agric. and Resour. Econ. and Sci., Symonston, Canberra.
Madsen, H. (2003), Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives, *Adv. Water Resour.*, *26*(2), 205–216.
Mann, H. B., and D. R. Whitney (1947), On a test of whether one of two random variables is stochastically larger than the other, *Ann. Math. Stat.*, *18* (1), 50–60.
McMillan, H., J. Freer, F. Pappenberger, T. Krueger, and M. Clark (2010), Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions, *Hydrol. Processes*, *24*(10), 1270–1284.
McVicar, T. R., M. L. Roderick, R. J. Donohue, and T. G. Van Niel (2012), Less bluster ahead? Ecohydrological implications of global trends of terrestrial near-surface wind speeds, *Ecohydrology*, *5*(4), 381–388.
Meehl, G. A., et al. (2007), *Global climate projections, in Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon et al., pp. 747–845, Cambridge Univ. Press, Cambridge, U. K.
Merz, R., J. Parajka, and G. Bloschl (2011), Time stability of catchment model parameters: Implications for climate impact analyses, *Water Resour. Res.*, *47*, W02531, doi:10.1029/2010WR009505.
Mishra, A. K. and V. P. Singh (2010), A review of drought concepts, *J. Hydrol.*, *391*(1–2), 202–216.
Mishra, A. K., and V. P. Singh (2011), Drought modeling––A review, *J. Hydrol.*, *403*(1–2), 157–175.
Montanari, A., et al. (2013), "Panta Rhei—Everything Flows": Change in hydrology and society—The IAHS Scientific Decade 2013–2022, *Hydrol. Sci. J.*, *58:6*, 1256–1275.
Morton, F. I. (1983), Operational estimates of areal evapotranspiration and their significance to the science and practice of hydrology, *J. Hydrol.*, *66*(1), 1–76.
Muleta, M. K., and J. Nicklow (2005), Sensitivity and uncertainty analysis coupled with automatic calibration for a distributed watershed model, *J. Hydrol.*, *306*(1), 127–145.
Nash, J., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models part I—A discussion of principles, *J. Hydrol.*, *10*(3), 282–290.
Nathan, R., and L. Lowe (2012), The hydrologic impacts of farm dams, *Aust. J. Water Resour.*, *16*(1), 1–10.
Nešpor, V., and B. Sevruk (1999), Estimation of wind-induced error of rainfall gauge measurements using a numerical simulation, *J. Atmos. Oceanic Technol.*, *16*(4), 450–464.

Oreskes, N., K. Shrader-Frechette, and K. Belitz (1994), Verification, validation, and confirmation of numerical models in the earth sciences, *Science*, *263*(5147), 641–646.
Oudin, L., V. Andreassian, T. Mathevet, C. Perrin, and C. Michel (2006), Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations, *Water Resour. Res.*, *42*, W07410, doi:10.1029/2005WR004636.
Pareto, V. (1927), *Manual of Political Economy*, translated from the French by A. S. Schwier, and A. N. Page, Augustus M. Kelley, N. Y.
Peel, M., B. Finlayson, and T. McMahon (2007), Updated world map of the Köppen-Geiger climate classification, *Hydrol. Earth Syst. Sci.*, *11*, 1633–1644.
Peel, M. C., T. A. McMahon, B. L. Finlayson, and F. G. Watson (2001), Identification and explanation of continental differences in the variability of annual runoff, *J. Hydrol.*, *250*(1), 224–240.
Perrin, C., C. Michel, and V. Andréassian (2001), Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments, *J. Hydrol.*, *242*(3), 275–301.
Perrin, C., C. Michel, and V. Andréassian (2003), Improvement of a parsimonious model for streamflow simulation, *J. Hydrol.*, *279*(1), 275–289.
Peterson, T. J., and A. W. Western (2014), Nonlinear time-series modeling of unconfined groundwater head, *Water Resour. Res.*, *50*, 8330–8355, doi:10.1002/2013WR014800.
Petrone, K. C., J. D. Hughes, T. G. Van Niel, and R. P. Silberstein (2010), Streamflow decline in southwestern Australia, 1950–2008, *Geophys. Res. Lett.*, *37*, L11401, doi:10.1029/2010GL043102.
Porter, J. W., and T. A. McMahon (1975), Application of a catchment model in southeastern Australia, *J. Hydrol.*, *24*(1), 121–134.
Potter, N. J., and F. H. S. Chiew (2011), An investigation into changes in climate characteristics causing the recent very low runoff in the southern Murray-Darling Basin using rainfall-runoff models, *Water Resour. Res.*, *47*, W00G10, doi:10.1029/2010WR010333.
Potter, N. J., F. H. S. Chiew, and A. J. Frost (2010), An assessment of the severity of recent reductions in rainfall and runoff in the Murray–Darling Basin, *J. Hydrol.*, *381*(1), 52–64.
Potter, N. J., L. Zhang, C. Petheram, and F. H. Chiew (2013), Hydrological non-stationarity in southeastern Australia, in *Proceedings of H01, 1062 IAHS-IAPSO-IASPEI Assembly*, edited by E. Boegh, pp. 358–363, IAHS Publ. 359, International Association of Hydrological Sciences, Gothenburg, Sweden.
Ramchurn, A. (2012), Improved modelling of low flows and drought impacts in Australian catchments using new rainfall-runoff model SpringSIM, in *Proceedings of the Australian Hydrology and Water Resources Symposium, 2012*, edited by S. Westra S. et al., pp. 429–440, Engineers Australia, Canberra, Australia.
Refsgaard, J. C., and J. Knudsen (1996), Operational validation and intercomparison of different types of hydrological models, *Water Resour. Res.*, *32*(7), 2189–2202.
Refsgaard, J. C., et al. (2014), A framework for testing the ability of models to project climate change and its impacts, *Clim. Change*, *122*, 271–282.
Reichert, P., and M. Omlin (1997), On the usefulness of overparameterized ecological models, *Ecol. Modell.*, *95*(2), 289–299.
Renard, B., D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks (2010), Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, *Water Resour. Res.*, *46*, W05521, doi:10.1029/2009WR008328.
Renard, B., D. Kavetski, E. Leblois, M. Thyer, G. Kuczera, and S. W. Franks (2011), Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation, *Water Resour. Res.*, *47*, W11516, doi:10.1029/2011WR010643.
Ritzel, B., J. Eheart, and S. Ranjithan (1994), Using genetic algorithms to solve a multiple objective groundwater pollution containment problem, *Water Resour. Res.*, *30*(5), 1589–1603.
Saft, M., A. W. Western, L. Zhang, M. C. Peel, and N. J. Potter (2015), The influence of multiyear drought on the annual rainfall-runoff relationship: An Australian perspective, *Water Resour. Res.*, *51*, 2444–2463, doi:10.1002/2014WR015348.
Savenije, H. H. (2004), The importance of interception and why we should delete the term evapotranspiration from our vocabulary, *Hydrol. Processes*, *18*(8), 1507–1511.
Schoups, G., and J. A. Vrugt (2010), A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, *Water Resour. Res.*, *46*, W10531, doi:10.1029/2009WR008933.
Seibert, J. (2000), Multi-criteria calibration of a conceptual runoff model using a genetic algorithm, *Hydrol. Earth Syst. Sci.*, *4*(2), 215–224.
Seibert, J., and J. J. McDonnell (2002), On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration, *Water Resour. Res.*, *38*(11), 1241, doi:10.1029/2001WR000978.
Shamir, E., B. Imam, H. V. Gupta, and S. Sorooshian (2005), Application of temporal streamflow descriptors in hydrologic model parameter estimation, *Water Resour. Res.*, *41*, W06021, doi:10.1029/2004WR003409.
Singh, R., T. Wagener, K. V. Werkhoven, K. V. Mann, and R. Crane (2011), A trading-space-for-time approach to probabilistic continuous streamflow predictions in a changing climate–accounting for changing watershed behavior, *Hydrol. Earth Syst. Sci.*, *15*(11), 3591–3603.
Storn, R., and K. Price (1997), Differential evolution––A simple and efficient heuristic for global optimization over continuous spaces, *J. Global Optim.*, *11*(4), 341–359.
Thirel, G., et al. (2015a), Hydrology under change: An evaluation protocol to investigate how hydrological models deal with changing catchments, *Hydrol. Sci. J.*, 60(7-8), 1184–1199.
Thirel, G.; V. Andréassian and C. Perrin (2015b), On the need to test hydrological models under changing conditions, *Hydrol. Sci. J.*, *60*(7-8) 1165–1173
Thyer, M., B. Renard, D. Kavetski, G. Kuczera, S. W. Franks, and S. Srikanthan (2009), Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis, *Water Resour. Res.*, *45*, W00B14, doi:10.1029/2008WR006825.
Tozer, C. R., A. S. Kiem, and D. C. Verdon-Kidd (2012), On the uncertainties associated with using gridded rainfall data as a proxy for observed, *Hydrol. Earth Syst. Sci.*, *16*(5), 1481–1499.
Turner, M. (2012), Hydrologic reference stations selection guidelines, version 1, report, Bureau of Meteorology, Melbourne, Australia.
Vaze, J., D. A. Post, F. H. S. Chiew, J. M. Perraud, N. R. Viney, and J. Teng (2010), Climate non-stationarity–validity of calibrated rainfall–runoff models for use in climate change studies, *J. Hydrol.*, *394*(3), 447–457.
Vogel, R. M., and A. Sankarasubramanian (2003), Validation of a watershed model without calibration, *Water Resour. Res.*, *39*(10), 1292, doi:10.1029/2002WR001940.
Vrugt, J. A., and B. A. Robinson (2007), Improved evolutionary optimization from genetically adaptive multimethod search, *Proc. Natl. Acad. Sci. U. S. A.*, *104*(3), 708–711.
Vrugt, J. A., and M. Sadegh (2013), Toward diagnostic model calibration and evaluation: Approximate Bayesian computation, *Water Resour. Res.*, *49*, 4335–4345, doi:10.1002/wrcr.20354.

Vrugt, J. A., H. V. Gupta, L. A. Bastidas, W. Bouten, and S. Sorooshian (2003), Effective and efficient algorithm for multiobjective optimization of hydrologic models, *Water Resour. Res.*, *39*(8), 1214, doi:10.1029/2002WR001746.
Vrugt, J. A., C. J. F. ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson (2008), Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, *Water Resour. Res.*, *44*, W00B09, doi:10.1029/2007WR006720.
Wagener, T., and A. Montanari (2011), Convergence of approaches toward reducing uncertainty in predictions in ungauged basins, *Water Resour. Res.*, *47*, W06301, doi:10.1029/2010WR009469.
Westerberg, I. K., and McMillan, H. K. (2015), Uncertainty in hydrological signatures, *Hydrol. Earth Syst. Sci.*, *19*, 3951–3968
Wilcoxon, F. (1945), Individual comparisons by ranking methods, *Biom. Bull.*, *1*(6), 80–83.
Wilks, D. (2011), Statistical Methods in the Atmospheric Sciences, *Int. Geophys. Ser.*, vol. 100, 3rd ed., Academic, Oxford, U. K.
Winsemius, H. C., B. Schaefli, A. Montanari, and H. H. G. Savenije (2009), On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information, *Water Resour. Res.*, *45*, W12422, doi:10.1029/2009WR007706.
Wöhling, T., L. Samaniego, and R. Kumar (2013), Evaluating multiple performance criteria to calibrate the distributed hydrological model of the upper Neckar catchment, *Environ. Earth Sci.*, *69*(2), 453–468.
Yadav, M., T. Wagener, and H. Gupta (2007), Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, *Adv. Water Resour.*, *30*(8), 1756–1774.
Yapo, P. O., H. V. Gupta, and S. Sorooshian (1998), Multi-objective global optimization for hydrologic models, *J. Hydrol.*, *204*(1), 83–97.
Ye, W., B. C. Bates, N. R. Viney, M. Sivapalan, and A. J. Jakeman (1997), Performance of conceptual rainfall-runoff models in low-yielding ephemeral catchments, *Water Resour. Res.*, *33*(1), 153–166.
Zhang, L., Dawes, W. R., and G. R. Walker (2001), Response of mean annual evapotranspiration to vegetation changes at catchment scale, *Water Resour. Res.*, *37*(3), 701–708.