Seth Westra, Mark Thyer, Michael Leonard, Dmitri Kavetski, and Martin Lambert
A strategy for diagnosing and interpreting hydrological model nonstationarity
Water Resources Research, 2014; 50(6):5090-5113

DOI: 10.1002/2013WR014719

AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2013WR014719

A strategy for diagnosing and interpreting hydrological model nonstationarity

Seth Westra[1], Mark Thyer[1], Michael Leonard[1], Dmitri Kavetski[1], and Martin Lambert[1]

[1]School of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide, South Australia, Australia

Key Points:
- A strategy to diagnose and interpret hydrological nonstationarity is presented
- Time-varying parameters are used to represent model nonstationarity
- The strategy reduces predictive biases over an independent confirmatory period

Correspondence to:
S. Westra,
seth.westra@adelaide.edu.au

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

Received 9 SEP 2013
Accepted 27 MAY 2014
Accepted article online 30 MAY 2014
Published online 24 JUN 2014

Abstract This paper presents a strategy for diagnosing and interpreting hydrological nonstationarity, aiming to improve hydrological models and their predictive ability under changing hydroclimatic conditions. The strategy consists of four elements: (i) detecting potential systematic errors in the calibration data; (ii) hypothesizing a set of "nonstationary" parameterizations of existing hydrological model structures, where one or more parameters vary in time as functions of selected covariates; (iii) trialing alternative stationary model structures to assess whether parameter nonstationarity can be reduced by modifying the model structure; and (iv) selecting one or more models for prediction. The Scott Creek catchment in South Australia and the lumped hydrological model GR4J are used to illustrate the strategy. Streamflow predictions improve significantly when the GR4J parameter describing the maximum capacity of the production store is allowed to vary in time as a combined function of: (i) an annual sinusoid; (ii) the previous 365 day rainfall and potential evapotranspiration; and (iii) a linear trend. This improvement provides strong evidence of model nonstationarity. Based on a range of hydrologically oriented diagnostics such as flow-duration curves, the GR4J model structure was modified by introducing an additional calibration parameter that controls recession behavior and by making actual evapotranspiration dependent only on catchment storage. Model comparison using an information-theoretic measure (the Akaike Information Criterion) and several hydrologically oriented diagnostics shows that the GR4J modifications clearly improve predictive performance in Scott Creek catchment. Based on a comparison of 22 versions of GR4J with different representations of nonstationarity and other modifications, the model selection approach applied in the exploratory period (used for parameter estimation) correctly identifies models that perform well in a much drier independent confirmatory period.

1. Introduction

The development of hydrological models that produce credible predictions under a changing climate is one of the most challenging aspects of hydrological modeling [*Klemes*, 1986]. This challenge is particularly pertinent when models are extrapolated outside the range of observed data used for parameter estimation, which is often necessary when looking at long lead times or high warming scenarios [*Milly et al*., 2008]. Under such conditions, model evaluation and selection require methods that make the best use of available historical data to assess the model's extrapolative ability [*Anderson and Woessner*, 1992; *Oreskes et al*., 1994].

One of the most stringent tests of hydrological model credibility is "differential split-sample testing" [*Klemes*, 1986]. In these tests, the performance of a calibrated model is evaluated on one or more periods that are climatologically different from the period used for parameter estimation; for example, a model calibrated under "wet" conditions can be tested on a "dry" period, and vice versa. For a model capable of such extrapolation, parameter estimates and predictive performance should remain similar across the two periods. However, numerous studies concluded that parameter estimates depended on the calibration period [*Gan and Burges*, 1990; *Wagener et al*., 2003; *Choi and Beven*, 2007; *Le Lay et al*., 2007; *Marshall et al*., 2007; *Wu and Johnston*, 2007; *Vaze et al*., 2010; *Merz et al*., 2011; *Zhang et al*., 2011; *Coron et al*., 2012; *Seiller et al*., 2012]. Furthermore, seasonal variations of hydrological parameters have been reported by *Ye et al*. [1997] and *Paik et al*. [2005].

We define the term "hydrological model nonstationarity" as the situation where hydrological model parameters vary in time, and thus depend on the period of record used for their estimation. Such nonstationarity can lead to poor predictions, especially when the model is applied to a climatologically different period [*Gharari et al*., 2013]. For example, *Coron et al*. [2012] found that models calibrated to a period with a wetter

climate overestimated the mean annual runoff when applied to a drier period, and vice versa. The severity of the nonstationarity problem and its implications on model prediction depend on multiple factors, including: (i) the length and variability of the historical record; (ii) the magnitude of future climate change; and (iii) the hydrological model [e.g., *Brigode et al*., 2012].

There are many possible reasons for hydrological model nonstationary, including systematic data errors, weaknesses in calibration procedures, numerical artefacts, model structural deficiencies, and others [*Beven and Binley*, 1992; *Wagener et al*., 2003; *Clark et al*., 2011; *Kavetski et al*., 2011]. For example, streamflow records can become biased due to siltation of weirs and changes in the channel flow geometry [*Guerrero et al*., 2012]; rainfall records can be affected by changes in the location and quality of rain gauges [*Molini et al*., 2005], and so forth. Similarly, poor choice of objective function can cause nonstationarity in the calibrated model parameters. For example, *Thyer et al*. [2009] showed that calibration to different time periods using a standard least squares objective function produced distinctly different estimates of hydrological parameters; these discrepancies were substantially reduced when a weighed least squares objective function was used.

A fundamental concern with hydrological nonstationarity is the possible implication that one or more important physical processes are not adequately represented [*Lin and Beck*, 2007; *de Vos et al*., 2010], or that changes in the catchment (e.g., land use changes) are occurring but are not explicitly represented by the model. We therefore argue that, provided that robust data, numerical methods and calibration procedures are used, hydrological model nonstationarity must be caused by the approximate nature of the hydrological models [*Anderson and Woessner*, 1992]. From this perspective, models with time-invariant parameters are more likely to be reliably representing the key physical processes. This is particularly important when predicting catchment response to future climatic forcings, as accurate process representation is critical when extrapolating a model outside of its calibrated range. Stationarity of model parameters can therefore be viewed as a necessary condition for the hydrological model to provide credible projections under extrapolation, and tests for stationarity can be useful as part of model selection for climate impact studies [*Seiller et al*., 2012].

A pragmatic approach to detect and mitigate nonstationarity is to calibrate the model to one or more historical periods that are analogous to the expected future hydroclimatic conditions [e.g., *Vaze et al*., 2010]. Provided such historical analogues are available, this approach reduces the extent of model extrapolation, and thus may be adequate for short future time horizons and small levels of climate change. An obvious limitation is that there may not be any historical periods that are sufficiently representative of the projected future conditions. This limitation can be particularly significant when it is recognized that hydroclimatic changes are expressed not only in terms of changes in annual average precipitation and potential evapotranspiration, but, just as importantly, in terms of the seasonality, intermittency, and intensity of future precipitation events [*Bates et al*., 2008; *Westra et al*., 2013]. Furthermore, by maximizing the "similarity" of the historical climate sequences to the projected future climate, it becomes necessary to use only relatively short portions of the historical record for model calibration, so that potentially valuable information on catchment behavior is ignored during parameter estimation. This is a type of bias-variance trade-off: to maximize the similarity between the calibration period and expected future climate (and hence reduce parameter bias), we need to use shorter periods of the historical record as the basis for calibration (which will usually increase parameter variance) [*Brigode et al*., 2012]. Finally, this approach does not characterize and/or resolve the cause of suspected model nonstationarity.

This paper develops a strategy to diagnose nonstationarity in hydrological model parameters and identify possible causes that require further investigation. The major distinct elements of the strategy are the characterization of parameter nonstationarity by representing hydrological model parameter(s) as a function of a set of time-varying covariates, the trialing of alternative model structures, and the assessment of empirical support for each proposed description of nonstationarity and/or alternative model structures using multiple model selection criteria. Compared to the existing approach of separately calibrating the hydrological model to different historical periods, the proposed approach has the following advantages:

1. A larger portion of the historical record is used for parameter estimation. This avoids the potential loss of information when discarding large portions of observed data.

2. By representing selected hydrological model parameters as continuous functions of selected covariates, it becomes possible to at least tentatively extrapolate these parameters to different hydroclimatic regimes (note that the difficulties of model evaluation under extrapolation described by *Klemes* [1986] still apply). Such extrapolation is not possible when the parameters are kept constant at values calibrated to a subset of the historical record.

3. The use of model selection techniques such as split-sample testing and/or information-theoretic approaches allows an assessment of whether the additional model complexity associated with the description of parameter nonstationarity produces a significant improvement in the model's predictive ability. In contrast, it is not clear how to evaluate the trade-off between model fit, complexity and length of record when calibrating parameters to different historical periods.

4. Additional insights are provided on the nature of possible deficiencies in the model structure [*de Vos et al*., 2010]. As parameter nonstationarity can be symptomatic of poor representation of important hydrological processes, it can serve as a valuable diagnostic of the suitability of the existing model for extrapolation. The nature of the suggested nonstationarity can help guide model improvement, especially when the nonstationarity can be attributed to a specific cause, such as a particular poorly represented process in the model or a major change in catchment conditions.

The paper is structured as follows. The key elements of the proposed strategy for diagnosing and interpreting hydrological nonstationarity are presented in section 2, followed by a description of the case study catchment in section 3. Section 4 provides a detailed investigation of data quality, including the analysis of possible systematic changes in the quality of rainfall, evapotranspiration, or streamflow data. Section 5 describes a set of 22 candidate hydrological models with different combinations of nonstationarity parameters to be evaluated, and section 6 describes the approach to parameter estimation. Section 7 details an AIC-based approach for model selection and diagnosis of hydrological nonstationarity. Results are presented in section 8, followed by discussion in section 9 and conclusions in section 10.

2. Overview of the Strategy for Diagnosing and Interpreting Hydrological Nonstationarity

Our strategy for developing hydrological models for predicting catchment runoff under changing hydroclimatic conditions follows the philosophical approach of "multiple working hypotheses," described originally by *Chamberlain* [1890] and more recently in the hydrological context by *Clark et al*. [2011]. In this approach, a set of candidate models (hypotheses) is constructed and evaluated, with each model providing an alternative representation of catchment behavior. The models are calibrated to observed data in an exploratory period, and an information-theoretic measure (the AIC) is used to evaluate the level of support from the data for each model. A selected subset of models is then tested on an independent confirmatory period that is climatologically different from the period used for parameter estimation, thus representing a differential split-sample test [*Klemes*, 1986]. The four elements of the strategy are outlined next.

2.1. Detecting Systematic Errors in the Calibration Data

Biases and systematic changes in the measurement of hydrological data can significantly affect model calibration and can lead to nonstationarity in the estimated model parameters. In situations where biases and/or changes in data quality cannot be excluded a priori, they must be retained among the working hypotheses to be evaluated a posteriori as part of model calibration and analysis. In this study, we use a set of standard diagnostics to assess the quality of the rainfall, potential evapotranspiration, and runoff data (section 4).

2.2. Modeling One or More Parameters as Functions of Time-Varying Covariates

As we define hydrological model nonstationarity as the case where hydrological model parameters change in time, a practical strategy for detecting nonstationarity is to allow the model parameters to vary in time as functions of selected covariates and examine the resulting impact on model performance. In this study, the covariates are selected to represent the major time scales of hydrological variability. For example, we use a sinusoidal function to represent seasonal changes in the catchment storage capacity. The covariates are discussed further in section 5.1, and resemble some of the time scales of variability used in the "unobserved components" of the data-based mechanistic modelling (DBM) approach [*Young and Beven*, 1994; *Young*, 1998]. In DBM, however, time-varying covariates are used as part of model identification and development using transfer functions, whereas in our case the purpose is largely as a diagnostic for structural errors in conceptual hydrological models.

2.3. Construction of Alternative Model Structures

One of the possible causes of nonstationarity in hydrological model parameters is poor process representation within the model. This step therefore aims to identify missing or poorly represented processes, and can

be used both to improve the hydrological model and to better characterize its predictive uncertainty. In this paper, we use multiple hydrological diagnostics, such as flow-duration curves stratified by season and by the phase of the hydrograph (rising and falling limbs), to isolate possible weaknesses in the conceptual model GR4J when simulating runoff in Scott Creek catchment. Based on this assessment, we make two modifications to the standard GR4J model; these are discussed in section 5.2.

Alternative approaches for model development and comparison include flexible model frameworks such as FUSE [*Clark et al*., 2008] and SUPERFLEX [*Fenicia et al*., 2011]. These frameworks can be used to analyze larger and more diverse sets of model structures. However, incorporating flexible model structures into the second step of the nonstationarity analysis strategy (section 2.2) requires further work to support nonnested structures with distinctly different conceptualizations and parameterizations. For example, in nonnested models, it may not be possible to apply nonstationary covariates to the same parameter, making it difficult to consistently compare the extent of parameter nonstationarity across all models under consideration.

2.4. Model Selection and Evaluation

The final step is to evaluate the empirical support for the model structures hypothesized and calibrated in Steps 2 and 3. Many model selection approaches can be used, including:

1. Cross-validation based methods [e.g., *Schoups et al*., 2008; *Hastie et al*., 2009], including split-sample testing, in which one or more models are fitted using a portion of the historical record (usually referred to as the "calibration" period) and tested on the remainder of the record (usually referred to as the "validation" or "verification" period). This has been the preferred approach in the hydrological literature for estimating "out of sample" model error [e.g., *Hastie et al*., 2009];

2. Information theory [e.g., *Burnham and Anderson*, 2010], which is receiving increased interest in the hydrological literature [e.g., *Gupta et al*., 2008; *Weijs et al*., 2010]. The information-theoretic framework aims to estimate the "in-sample" prediction error from the likelihood (objective) function calculated during model calibration, while also attempting to account for the expected model "optimism" arising from the assessment of model performance over the calibration period itself [*Hastie et al*., 2009]. The Akaike Information Criterion (AIC) [*Akaike*, 1974] and its small sample approximation (AICc) [*Sugiura*, 1978] are widely used model selection criteria derived using information theory.

3. Bayesian approaches, such as the Bayesian Information Criterion (BIC) [*Schwarz*, 1978; *Marshall et al*., 2005; *Martinez and Gupta*, 2011], Kashyap's Information Criteria (KIC) [*Kashyap*, 1982; *Martinez and Gupta*, 2011], and Bayesian model averaging [*Hoeting et al*., 1999; *Claeskens and Hjort*, 2008].

There are ongoing debates in the hydrological and broader communities on the advantages, limitations, and interpretations of different model selection criteria [e.g., *Gupta et al*., 2008; *Ye et al*., 2008; *Burnham and Anderson*, 2010]. An increasing number of studies compare multiple model selection approaches, often with contradictory results that appear to depend on specific features of the data and models being investigated [*Schoups et al*., 2008; *Ye et al*., 2008; *Burnham and Anderson*, 2010; *Dai et al*., 2012; *Engelhardt et al*., 2013]. In this study, we adopt the AIC because it is a simple yet widely used model selection criterion that seeks to maintain parsimony while selecting the model with the greatest predictive ability [*McQuarrie and Tsai*, 2007; *Burnham and Anderson*, 2010]. The key properties of this criterion are given in section 7.

Note that this paper uses the term "exploratory period" to refer to the period used for parameter estimation (calibration), model comparison, and selection. Furthermore, the term "confirmatory period" refers to the period used for independent model evaluation. The confirmatory period is commonly referred to as the validation or verification period in the hydrological literature, however the term "confirmatory" is intended to emphasize that future model performance cannot be "validated" or "verified" from past performance alone [*Oreskes et al*., 1994].

3. Case Study Catchment

The four steps of the strategy for analysing nonstationarity of hydrological model parameters are illustrated using the Scott Creek catchment in South Australia. This catchment has an area of 29 km^2 and forms a part of the larger Onkaparinga catchment—Adelaide's primary surface water source (Figure 1). The median

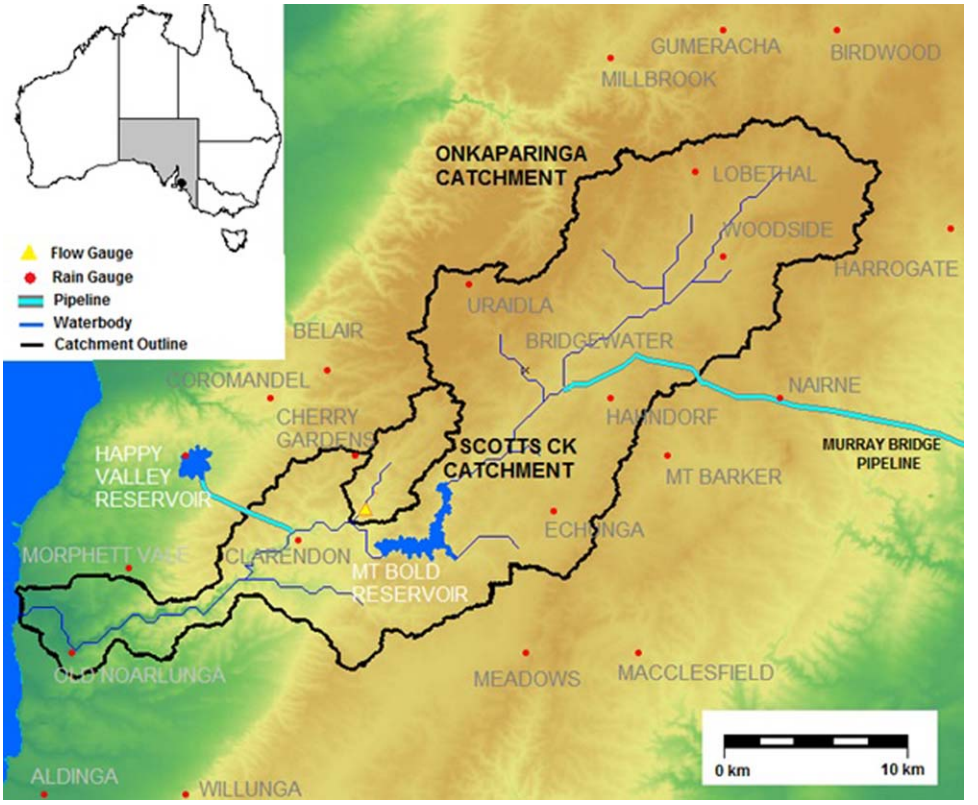

Figure 1. Map of the Onkaparinga catchment.

annual rainfall (P) in Scott Creek is 905 mm, and the median annual potential evapotranspiration (PET) is 1600 mm. The long-term average runoff is 123 mm, giving a runoff coefficient of 0.14.

The Scott Creek catchment is classified as semiarid and has a winter-dominated rainfall regime. February is the driest month (monthly average of 20 mm), while July is the wettest (monthly average of 130 mm). In contrast, monthly PET varies from 50 mm in July to 250 mm in January. Therefore, in summer the catchment is water-limited ($P << PET$), whereas in winter it is energy-limited ($P >> PET$). The combined effect of seasonality in P and PET is that, in an average year, the runoff is highly seasonal, with over 75% occurring in the 3 month period from July to September. The seasonality of the catchment suggests that different physical mechanisms may be governing the rainfall-runoff relationships in summer and winter.

In addition to seasonal variations, the runoff characteristics of the Scott Creek catchment also vary interannually. At the aggregated annual scale, the relationship between catchment-average rainfall and runoff is approximately linear (with a Pearson correlation R^2 of 0.80), and a 1% change in annual rainfall yields an approximately 3% change in runoff. This catchment sensitivity is within the typical range for semiarid catchments in southeast Australia [*Chiew*, 2006]. The runoff coefficient, when calculated for each calendar year, varies from 0.06 in the driest year (2006) to 0.22 in the wettest year (1986).

The streamflow varies over four orders of magnitude, with approximately 21% of days over the exploratory and confirmatory periods having flows below 0.01 mm/d, and with only 22 days having flows above 10 mm/d. Approximately 30% of the total flow volume occurs in the top 1% of flow days, and 68% of the total flow volume occurs in the top 10% of flow days.

The 1985–1999 period is used for the exploratory analysis (parameter estimation and model selection), and the 2000–2009 period is used for the confirmatory analysis (model evaluation). Prior to both periods, a 4 year spin-up period is used to reduce the impact of unknown initial conditions. The confirmatory period is much drier than the exploratory period, with 19% less runoff on average, and therefore provides a stringent differential split-sample test.

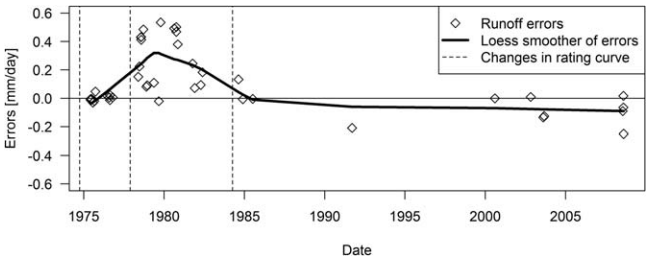

Figure 2. Runoff error time series at Scott Creek over the exploratory (1985–1999) and confirmatory (2000–2009) periods. Runoff errors are defined as the differences between the streamflows predicted by the rating curve and the actual streamflow gauging. Error analysis using the loess smoother [*Hastie et al*., 2009] shows a clear overprediction of streamflow prior to the last rating curve change in 1984.

4. Identifying Systematic Errors in Calibration Data

The first element of the strategy is to identify systematic errors in the observed data. In this study, we examine the quality of observed streamflow, potential evapotranspiration, and rainfall.

Streamflow estimates for Scott Creek were obtained from a rectangular stepped weir located near the catchment outlet and operated continuously since 1969. Analysis of the differences between streamflow gaugings and streamflow estimates from the rating curve suggests a significant increase in rating curve errors during 1980–1984, with some evidence of systematic bias (Figure 2). Furthermore, the gauging station metadata indicates that a major rating curve change occurred in 1984. Hence, to avoid the impact of potentially biased streamflow data on the inference of nonstationarity, our analysis is based exclusively on post-1984 data. The drawback of selecting this time period is that it has a smaller number of rating curve measurements, so that all flows greater than 10 mm (1-in-6-month flow) are extrapolated.

Catchment-average PET was estimated using Morton's areal potential evapotranspiration (APET) method [*Morton*, 1983; *McMahon et al*., 2013], which is based on temperature, vapor pressure, and incoming solar radiation data from the Australian SILO 0.05° latitude/longitude gridded data set [*Jeffrey et al*., 2001]. The time series of annual APET have a slight upward trend from 1985 to 2009. A similar trend is present in Morton's APET estimated at the high-quality Kent Town weather station (the nearest high-quality weather recording station), indicating that this trend is unlikely to be caused by measurement errors.

Three rainfall gauges are located within or very close to Scott Creek catchment. Continuous rainfall data for these gauges were obtained from the SILO patched point database, and these data are occasionally infilled using interpolated data when observed data are missing or suspect [*Jeffrey et al*., 2001]. Therefore, to detect potential systematic errors, a homogeneity analysis [*Allen et al*., 1998] was performed by comparing the rainfall time series at each gauge in Scott Creek catchment to time series from the rain gauge at Happy Valley, which is part of Australia's high-quality gauge network [*Lavery et al*., 1992]. No statistically significant evidence of inhomogeneity was found. The catchment-average rainfall for Scott Creek was obtained by kriging the three gauges, and is dominated by a single gauge at Cherry Gardens (see Figure 1), which has a weight of 0.9.

Based on the analysis of streamflow, PET, and rainfall data in Scott Creek catchment, we conclude that these data are of relatively high quality from 1985 onward, and we therefore use only post-1984 data for model development and evaluation. A negative consequence of using stringent criteria for data selection is that potentially long portions of the historical record might be discarded from the analysis. For the present case study, the record retained is sufficient for the intended analysis, and reduces the potential contribution of poor data quality to parameter nonstationarity.

An alternative way of addressing data quality is to develop more comprehensive data error models. For example, rainfall error models could be based on detailed geostatistical analysis [*Renard et al*., 2011]. However, this requires considerable additional information and was not pursued in this study.

5. Candidate Hydrological Models

All hydrological models considered in this work are derived from the lumped conceptual rainfall-runoff model GR4J [*Perrin et al*., 2003]. The published version of GR4J has four calibration parameters, namely the production store capacity (θ_1, units of mm), the groundwater exchange coefficient (θ_2, units of mm), the

1 day-ahead maximum capacity of the routing store (θ_3, units of mm), and the time base of the unit hydrograph (θ_4, units of days).

GR4J was developed to provide, on average, good performance across a wide range of catchment conditions [*Perrin et al*., 2003]. This makes GR4J particularly suitable as a starting point for model modifications and refinements, including the versions constructed in this work as part of detecting and quantifying hydrological nonstationarity. The GR4J modifications are described next.

5.1. Simulating Hydrological Model Nonstationarity

Parameter θ_1 is allowed to vary in time to represent several potential time scales of model nonstationarity. We focus on θ_1 because it represents the primary storage of water in the catchment. Previous studies [*Kuczera et al*., 2006; *Renard et al*., 2011] have indeed suggested that θ_1 is the most sensitive GR4J parameter, with *Renard et al*. [2011] showing through a sensitivity analysis that stochastic variations of θ_1 have the largest impact on model predictions. By treating θ_1 as a function of multiple covariates representing seasonal, annual, and longer-term variability, we attempt to characterize the major potential time scales of nonstationarity, as follows:

1. Seasonal-scale variability in catchment characteristics is represented by conditioning θ_1 on a sine function with a yearly period, parameterized by its amplitude and phase. In the Scott Creek catchment, a major source of seasonality might be the switch from water limitations in summer to energy limitations in winter (section 3).

2. Annual-scale variability due to hydrometeorological changes is represented by conditioning θ_1 on the 365 day antecedent daily rainfall and potential evapotranspiration. This conditioning aims to account for nonstationarity in the predictive errors, such as when a hydrological model systematically overestimates flows during dry years and underestimates flows during wet years [e.g., *Coron et al*., 2012; *Pathiraja et al*., 2012].

3. Long-term changes in catchment response are represented using a linear trend in θ_1.

The full nonstationary model for θ_1 is:

$$\theta_1(t|\boldsymbol{\lambda}) = \lambda_1 + \underbrace{\lambda_2 t}_{\text{linear trend}} + \underbrace{\lambda_3 \sin\left(2\pi \frac{t+\lambda_4}{365}\right)}_{\text{seasonal variability}} + \underbrace{\lambda_5 P_{365} + \lambda_6 PET_{365}}_{\text{annual variability}} \quad (1)$$

where t is the number of days since the start of simulation and $\lambda_1, \ldots, \lambda_6$ are six "nonstationarity" parameters. Parameter λ_1 is a constant term, λ_2 represents the linear trend, $\{\lambda_3, \lambda_4\}$ represent the amplitude and phase of the sine term, and $\{\lambda_5, \lambda_6\}$ represent the influence of previous 365 day rainfall (P_{365}) and potential evapotranspiration (PET_{365}). Note that parameters λ_1 and λ_4 depend on the starting date of the simulation (here selected as 1 January in both the exploratory and confirmation periods).

As discussed in section 9.3, there may be physically interpretable reasons for temporal changes in catchment storage capacity. For example, an increase in on-farm dams [*Teoh*, 2002] in Scott Creek catchment may lead to an increase in the total available storage volume, and thus to a larger value of the storage parameter θ_1. Other forms of nonstationarity might be less physically interpretable. For example, in Scott Creek, the total volume of available storage in the soil matrix is unlikely to change regularly each season, so that the presence of a sinusoidal pattern in θ_1 does not immediately indicate a seasonal change in actual catchment storage capacity. Therefore, we view the primary purpose of the covariates described in this section as diagnostic: by representing the main time scales of likely variation in model parameters, it becomes possible to identify deficiencies in the model structure, which in turn can be used to identify areas for model improvement.

5.2. Modifying the Structure of GR4J

Nonstationarity in hydrological model parameters can indicate that a hydrological process is either absent or incorrectly represented in the hydrological model. We test this proposition by making several modifications to GR4J, based on the results of model diagnostics (discussed further in section 7).

5.2.1. Representation of Recession Dynamics

Inspection of hydrographs predicted using the standard GR4J model indicated systematic deficiencies in the representation of the falling limb (section 8). To improve the representation of recession behavior, an

additional parameter θ_5 is introduced to provide greater flexibility in the GR4J equation that controls the partitioning of net rainfall between the production and routing stores:

$$P_s = \frac{\theta_1\left(1-\left(\frac{S}{\theta_1}\right)^{\theta_5}\right)\tanh\left(\frac{P_n}{\theta_1}\right)}{1+\frac{S}{\theta_1}\tanh\left(\frac{P_n}{\theta_1}\right)} \quad (2)$$

where P_s is the portion of net rainfall P_n that enters the production store and S is the water content in the production store [compare with *Perrin et al*., 2003, equation 3].

5.2.2. Representation of Evapotranspiration Dynamics

The low runoff coefficient in the Scott Creek catchment and the general aridity of its regional environment indicate a large contribution of evapotranspiration to the overall water balance. Analysis of the GR4J simulations found that almost 30% of the rainfall is converted to actual evapotranspiration on rainy days, before the rainfall enters the production store.

In the original version of GR4J, actual evapotranspiration (AET) is determined from two different model processes. The first process occurs on all rainy days when P > PET; here the net rainfall is calculated $P_n = P - PET$, and AET occurs at the potential rate. The second process occurs on days when P < PET, and the AET is calculated as a function of the water level in the production store.

In the modified GR4J, an alternative formulation is considered, in which $P_n = P$ (i.e., removing the first process), and AET is only a function of the volume of water in the production store. This representation is common in hydrological models, including HBV [*Bergstrom*, 1995], TOPMODEL [*Beven et al*., 1995], and others.

5.3. Model Structure Groupings

To assist in the systematic comparison of predictive performance across a large number of candidate models, we define the following three model structure groupings, as described in Table 1:

1. A set of eight model structures, labeled $g_{1.1},\ldots,g_{1.8}$, are used to cover all possible combinations of the three nonstationarity components developed in section 5.1. Note that the individual terms within each distinct nonstationarity component in equation (1) are always considered jointly (e.g., we do not split the annual variability representation into individual P and PET terms).

2. A set of four model structures, labeled $g_{2.1},\ldots,g_{2.4}$, are used to examine the impact of GR4J structural modifications presented in section 5.2 for improving the representation of recession and evapotranspiration dynamics. Note that the original GR4J model ($g_{1.1}$) is included in this grouping as model $g_{2.1}$, and is used as a reference against which this set of model modifications are compared.

3. A set of 12 model structures, labeled $g_{3.1},\ldots,g_{3.12}$, are given by different combinations of nonstationarity models and GR4J structural modifications. In this grouping, the reference model ($g_{3.1}$) is selected to be model $g_{2.2}$, as this model was found to be the best model in grouping $g_{2.x}$ (section 8.2). Note that the model grouping $g_{3.x}$ does not include all possible combinations of covariates for nonstationary θ_1 and other model modifications, as this would have led to an excessively large number of candidate models. Rather, important groups of parameters were identified based on the analysis of the first two model groupings ($g_{1.x}$ and $g_{2.x}$); this is discussed further in section 8.

6. Parameter Estimation

This section describes the method of maximum likelihood used in this study to estimate the model parameters. This method requires the construction of a likelihood function, followed by parameter optimization through likelihood maximization.

6.1. Specification of the Likelihood Function

The likelihood function $\mathcal{L}()$ is defined as the joint probability of the observed streamflow given the observed forcings and the parameters $\boldsymbol{\theta}$ of a predictive model, i.e., $\mathcal{L}(\boldsymbol{\theta}) = p(\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_n | \tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n, \boldsymbol{\theta}) = p(\tilde{\mathbf{y}} | \tilde{\mathbf{x}}, \boldsymbol{\theta})$.

Table 1. Modified GR4J Models Used in This Paper[a]

					Antecedent						
New Process:		Trend	Seasonality		Rain	PET					Net Precip
Model	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	θ_2	θ_3	θ_4	θ_5	P_n
$g_{1.1}$	✓						✓	✓	✓		$P-E$
$g_{1.2}$	✓	✓					✓	✓	✓		$P-E$
$g_{1.3}$	✓		✓	✓			✓	✓	✓		$P-E$
$g_{1.4}$	✓				✓	✓	✓	✓	✓		$P-E$
$g_{1.5}$	✓	✓	✓	✓			✓	✓	✓		$P-E$
$g_{1.6}$	✓		✓	✓	✓	✓	✓	✓	✓		$P-E$
$g_{1.7}$	✓	✓			✓	✓	✓	✓	✓		$P-E$
$g_{1.8}$	✓	✓	✓	✓	✓	✓	✓	✓	✓		$P-E$
$g_{2.1}=g_{1.1}$	✓						✓	✓	✓		$P-E$
$g_{2.2}$	✓						✓	✓	✓	✓	$P-E$
$g_{2.3}$	✓						✓	✓	✓		P
$g_{2.4}$	✓						✓	✓	✓	✓	P
$g_{3.1}=g_{2.2}$	✓						✓	✓	✓	✓	$P-E$
$g_{3.2}$	✓	✓					✓	✓	✓	✓	$P-E$
$g_{3.3}$	✓		✓	✓			✓	✓	✓	✓	$P-E$
$g_{3.4}$	✓	✓	✓	✓			✓	✓	✓	✓	$P-E$
$g_{3.5}$	✓	✓					✓	✓	✓		P
$g_{3.6}$	✓		✓	✓			✓	✓	✓		P
$g_{3.7}$	✓	✓	✓	✓			✓	✓	✓		P
$g_{3.8}$	✓	✓					✓	✓	✓	✓	P
$g_{3.9}$	✓		✓	✓			✓	✓	✓	✓	P
$g_{3.10}$	✓	✓	✓	✓			✓	✓	✓	✓	P
$g_{3.11}$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	$P-E$
$g_{3.12}$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	P

[a]The parameters for the nonstationary model of θ_1 are represented by $\lambda_1,\ldots,\lambda_6$ as described in equation (1). Parameter θ_5 is described in equation (2). The last column describes the approach used to calculate net rainfall (P_n).

The predictive model is constructed by combining a hydrological model with a description of predictive uncertainty, as detailed next.

Consider a deterministic hydrological model $h()$, such as GR4J. At time step t, the model predictions of streamflow y_t are:

$$y_t = h(\tilde{\mathbf{x}}_{1:t}; \boldsymbol{\theta}_h) \tag{3}$$

where $\tilde{\mathbf{x}}_{1:t}$ is the time series ($t=1,\ldots,n$) of observed hydrological inputs (here, daily rainfall and PET) and $\boldsymbol{\theta}_h$ is the vector of hydrological model parameters (here, $\boldsymbol{\theta}_h=\{\theta_1,\ldots,\theta_5\}$).

Next, consider an additive residual error model, defined as

$$\varepsilon_t = \tilde{y}_t - y_t \tag{4}$$

where $\tilde{y}_t$ is the observed streamflow at time step t.

The residual error model in equation (4) provides an aggregate representation of all data and structural errors responsible for the differences between observed and predicted streamflows [*Kennedy and O'Hagan*, 2001].

We assume that the residuals ε are independent in time and follow a Gaussian distribution with zero mean and standard deviation σ_ε, i.e., $\varepsilon \sim N(0,\ \sigma_\varepsilon)$. As hydrological model residuals are typically heteroscedastic [*Sorooshian*, 1981; *Schoups and Vrugt*, 2010], we allow σ_ε to vary in time as a linear function of predicted streamflow, i.e.,

$$\sigma_{\varepsilon(t)} = a_\varepsilon + b_\varepsilon y_t \tag{5}$$

The error model parameters $\boldsymbol{\theta}_\varepsilon=\{a_\varepsilon,\ b_\varepsilon\}$ are unknown and are therefore estimated as part of the inference.

Under the residual error assumptions listed above, the following log likelihood is obtained:

$$\log \mathcal{L}(\boldsymbol{\theta}) = \log \mathcal{L}(\boldsymbol{\theta}_h, \boldsymbol{\theta}_\varepsilon) = \sum_{t=1}^{n} \log f\left[\varepsilon_t(\boldsymbol{\theta}_h); 0, \sigma_{\varepsilon(t)}(\boldsymbol{\theta}_h, \boldsymbol{\theta}_\varepsilon)\right] \quad (6)$$

where $f(z; \mu, \sigma)$ is the Gaussian probability density function with mean μ and standard deviation σ, evaluated at point z. Note that the residuals depend solely on hydrological parameters, whereas the standard deviations of the residuals depend on both hydrological and error model parameters.

The impact of the assumptions underlying the residual error model (equations (4) and (5)) on the model selection technique is discussed further in section 7.

6.2. Extension to Models With Nonstationary Parameters

As detailed in section 2.2, hydrological nonstationarity can be investigated by allowing one or more hydrological model parameters θ_h to vary in time as functions of selected covariates. For example, θ_1 is modeled as a function of covariates as described in equation (1). This can be accommodated within the likelihood function by no longer calibrating θ_1 and instead calibrating $\lambda_1, \ldots, \lambda_6$.

The remainder of the paper uses the short-hand notation g to represent the combined hydrological and error models,

$$\begin{aligned} g &= g(\tilde{\mathbf{x}}_{1:t}; \boldsymbol{\theta}) \\ &= h(\tilde{\mathbf{x}}_{1:t}; \boldsymbol{\theta}_h, \boldsymbol{\lambda}) + \varepsilon \end{aligned} \quad (7)$$

As discussed in section 5.3, we compare the performance of 22 alternative models listed in Table 1. The individual models are identified by an index on g. Note that the models have different numbers of calibrated parameters, e.g., model $g_{1.2}$ can be written as $g_{1.2} = h(\tilde{\mathbf{x}}_{1:t}; \theta_2, \theta_3, \theta_4, \lambda_1, \lambda_2) + \varepsilon$.

6.3. Mitigating Deficiencies in the Assumed Likelihood Function

The assumption of independent residual errors in equation (6) is poor in most hydrological applications [*Sorooshian and Dracup*, 1980; *Evin et al.*, 2014]. Moreover, near-zero flows exert a strong influence on the inference when using a likelihood that represents error heteroscedasticity using the linear relationship in equation (5). Therefore, two changes are made to the likelihood function, as detailed below.

6.3.1. Handling Low (Close to Zero) Flows in the Likelihood Function

The Scott Creek catchment is highly seasonal, typically with very little runoff during summer. The handling of low flows in the likelihood function is the subject of ongoing research [e.g., *Smith et al.*, 2010]. To avoid this issue negatively impacting on the analysis, observed daily flows below a threshold of 0.09 mm are censored from the likelihood function. The resulting streamflow data set is referred to as $\tilde{\mathbf{y}}^{(>0.09)}$. This censoring threshold corresponds to the streamflow value for which, based on the rating curve analysis, there is a 95% probability that the streamflow predicted by the rating curve is greater than zero. Over the exploratory period, 55% of days have flows below 0.09 mm, yet these censored days contribute less than 5% of the total catchment flow volume.

Residual error diagnostics were checked in all cases, and are presented here for the simplest model ($g_{1.1}$) and for one of the most complex models ($g_{3.11}$). Density plots of the standardized residuals in Figure 3 provide empirical support for the Gaussian assumption used in the error model. The reliability of the total predictive uncertainty was assessed using a predictive quantile-quantile plot [*Thyer et al.*, 2009] (not shown). The observed p-values are very close to the 1:1 line, suggesting that the error model provides a reasonably reliable approximation of the probability distribution of the residuals.

6.3.2. Handling Autocorrelation in the Residuals

Autocorrelation of residual errors can significantly influence model inference and selection, yet is omitted in equation (6). In this case study, statistically significant error autocorrelation was found for all models. For the most complex model ($g_{3.11}$), the lag-1 autocorrelation coefficient for the residuals (after the low flow threshold is applied) is 0.32, which, although relatively low in the context of rainfall-runoff applications, is statistically significant at the 5% level. To reduce the impact of ignoring autocorrelation in the likelihood function, all hydrological models were recalibrated to a "thinned" streamflow set comprising every kth day of record. We trialed several values of k, and identified the minimal value of k for which the lag-1 autocorrelation coefficient was no longer significant at the 5% level. For almost all the models, this led to a 6 day sampling interval ($k = 6$).

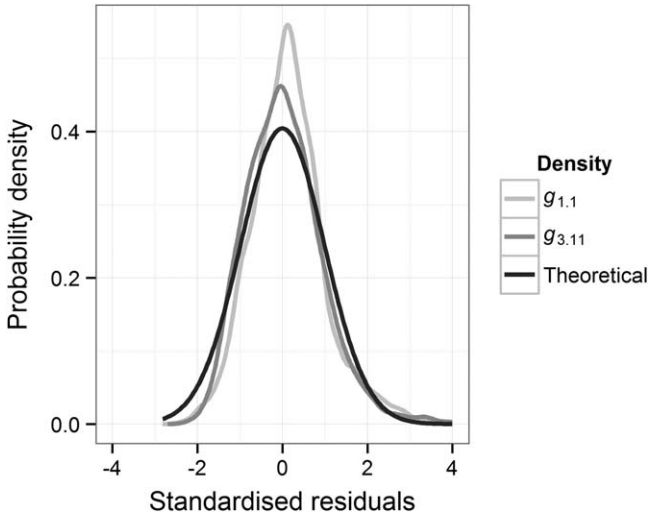

Figure 3. Density plots of standardized residuals in the exploratory period for models $g_{1.1}$ and $g_{3.11}$. The standard Gaussian distribution is shown for reference.

The thinning is incorporated into the likelihood function in equation (6) by only including the model residuals from every sixth day of record, while still censoring days with observed flows below the 0.09 mm threshold, i.e., $\varepsilon_t : t \in \{1, 7, 13, ...\} \cap \tilde{y}_t > 0.09$. The corresponding streamflow set is referred to as $\tilde{\mathbf{y}}^{(>0.09)}_{t:=1+6j}$.

The sensitivity of the results to the particular choice of thinned period is investigated by calibrating (separately) to six nonoverlapping sets of thinned residuals, defined as $\varepsilon_t : t \in \{2, 8, 14, ...\} \cap \tilde{y}_t > 0.09$, $\varepsilon_t : t \in \{3, 9, 15, ...\} \cap \tilde{y}_t > 0.09$, and so on. The corresponding streamflow sets are referred to as $\tilde{\mathbf{y}}^{(>0.09)}_{t:=2+6j}$, $\tilde{\mathbf{y}}^{(>0.09)}_{t:=3+6j}$, and so on.

More complex residual error models, such as those including specialized treatment of low flows [*Smith et al.*, 2010] and direct treatment of error autocorrelation [*Evin et al.*, 2014], are clearly of interest to improve the specification of the likelihood function. However, practical difficulties have been encountered when jointly inferring error autocorrelation and heteroscedasticity, including strong interactions of the error autocorrelation parameter with the GR4J mass balance parameter θ_2 [*Evin et al.*, 2014]. Moreover, combined treatment of error autocorrelation and low flows requires separate theoretical development. Hence, censoring of low flows and calibrating to thinned streamflow sets was used in this work as a pragmatic approach to reduce the violations of the likelihood assumptions.

6.4. Parameter Optimization

The parameter values that maximize the likelihood function in equation (6) were estimated using a quasi-Newton optimization method. Optimization was repeated with 100 random starting points to reduce the probability of being trapped in local optima.

7. Model Evaluation and Selection

We use an information-theoretic approach combined with multiple hydrologically oriented diagnostics to evaluate the performance of the hydrological models described in section 5. This section describes the specific metrics used.

7.1. The Akaike Information Criterion

Information-theoretic techniques use the Kullback-Leibler information to compare an approximate probability model $p(\tilde{\mathbf{y}}|\boldsymbol{\theta})$ against the (unknown) "true" probability density function $p_{true}(\tilde{\mathbf{y}})$ describing the system of interest [*Burnham and Anderson*, 2010]:

$$I_{KL}(p_{true}(\tilde{\mathbf{y}}) \parallel p(\tilde{\mathbf{y}}|\boldsymbol{\theta})) = \int \log \frac{p_{true}(\tilde{\mathbf{y}})}{p(\tilde{\mathbf{y}}|\boldsymbol{\theta})} p_{true}(\tilde{\mathbf{y}})\ \mathrm{d}\tilde{\mathbf{y}}$$
$$= \int \log\left(p_{true}(\tilde{\mathbf{y}})\right) p_{true}(\tilde{\mathbf{y}}) \mathrm{d}\tilde{\mathbf{y}} - \int \log\left(p(\tilde{\mathbf{y}}|\boldsymbol{\theta})\right) p_{true}(\tilde{\mathbf{y}})\ \mathrm{d}\tilde{\mathbf{y}} \quad (8)$$

The Kullback-Leibler information $I_{KL}(p_{true}(\tilde{\mathbf{y}}) \parallel p(\tilde{\mathbf{y}}|\boldsymbol{\theta}))$, often referred to as the "Kullback-Leibler divergence of $p(\tilde{\mathbf{y}}|\boldsymbol{\theta})$ from $p_{true}(\tilde{\mathbf{y}})$," can be interpreted as the information lost when an approximate likelihood $p(\tilde{\mathbf{y}}|\boldsymbol{\theta})$ is used to represent the "true" likelihood $p_{true}(\tilde{\mathbf{y}})$. Since $p_{true}(\tilde{\mathbf{y}})$ represents the "truth," it does not vary as a

function of the parameters, whereas $p(\tilde{\mathbf{y}}|\boldsymbol{\theta})$ varies over the parameter space $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Also note that the conditioning of $p(\tilde{\mathbf{y}}|\boldsymbol{\theta})$ on $\tilde{\mathbf{x}}$ indicated previously has been suppressed for notational convenience.

We stress that $p(\tilde{\mathbf{y}}|\boldsymbol{\theta})$ refers to the complete probability model of the data, which here is constructed by combining a deterministic component (i.e., the hydrological model) and a stochastic component (i.e., the error model).

In real environmental systems $p_{true}(\tilde{\mathbf{y}})$ is unknown and therefore the Kullback-Leibler information cannot be calculated. However, since the term $\int \log\left(p_{true}(\tilde{\mathbf{y}})\right) p_{true}(\tilde{\mathbf{y}}) \mathrm{d}\tilde{\mathbf{y}}$ in equation (8) is a constant that depends only on the (unknown) "truth," it is possible to calculate the difference in Kullback-Leibler information between any two models $p_1(\tilde{\mathbf{y}}|\boldsymbol{\theta})$ and $p_2(\tilde{\mathbf{y}}|\boldsymbol{\theta})$. This difference can be treated as a measure of relative empirical support in favor of one of the models.

The Akaike Information Criterion (AIC) is derived such that, under a set of assumptions discussed below, choosing the model that maximizes the AIC yields the smallest Kullback-Leibler divergence from the true model p_{true} [*Akaike*, 1974]. The derivation of the AIC, $\mathcal{A}$, from the Kullback-Leibler information is described in *Burnham and Anderson* [2010], and uses the maximum-likelihood estimate of the parameter vector, $\hat{\boldsymbol{\theta}}$:

$$\mathcal{A} = \log \mathcal{L}(\hat{\boldsymbol{\theta}}) - K \tag{9}$$

The term K denotes the number of calibrated parameters in the model, and is often described as a "complexity penalty" that accounts for the fact that the model parameters $\hat{\boldsymbol{\theta}}$ are being calibrated to the (finite) observed data.

The AIC differences, denoted by $\Delta\mathcal{A}$, can be interpreted as the loss of information when model i is used instead of the AIC-best model in a set of models under comparison:

$$\Delta\mathcal{A}_i = \mathcal{A}_i - \mathcal{A}_{\min} \tag{10}$$

This metric can be evaluated for each model $i = 1,\ldots,M$ in the set of M models being compared, with $\mathcal{A}_{\min}$ being the lowest (best) AIC value produced by the models in the set.

Akaike "weights," $w^{(\mathcal{A})}$, defined for model i from the set of M models as:

$$w_{i|M}^{(\mathcal{A})} = \frac{\exp\left(-\frac{1}{2}\Delta\mathcal{A}_i\right)}{\sum_{j=1}^{M} \exp\left(-\frac{1}{2}\Delta\mathcal{A}_j\right)} \tag{11}$$

can then be interpreted as the "weight of evidence in favor of model i," i.e., the probability that, given the set of M models, model i will obtain the highest likelihood value when predicting new data arising from the same system.

The Akaike weights facilitate a probabilistic interpretation of AIC differences. Values of $\Delta\mathcal{A}_i$ less than 2 are usually interpreted as indicating "substantial" support for model i, whereas values greater than 10 indicate that there is "virtually no support" for that model [*Burnham and Anderson*, 2010].

Two major assumptions underlying the AIC should be considered. First, the term K in equation (9) is derived under the assumption that the sample size is "large." A "large" sample is usually defined when $n/K > 40$ [*Burnham and Anderson*, 2010], and in this study the criterion is met in all cases. Second, the AIC is derived under the assumption that the likelihood function provides a "good" approximation to the actual system. This assumption is questionable in this study, in particular because the error model used to derive the likelihood in equation (6) assumes the residuals are independent (section 6.3.2). Since neglecting the serial dependence of the errors results in an overestimation of the information content of the data and may affect the AIC assumptions, this paper does not use the full interpretation of the AIC weights described in the preceding paragraph. However, despite these limitations, we proceed on the assumption that AIC-based rankings and the relative magnitudes of the AIC differences and the AIC weights can still help guide model selection (see section 8).

We also note that the likelihood function used in this work to calibrate the hydrological models and to compute the AIC criterion is based on an aggregate residual error model that does not distinguish between

structural and data errors (refer to equation (4)). Although such aggregate error models may still provide a suitable basis for predictive applications [e.g., *Evin et al*., 2014], the associated likelihood functions may not provide all the information required to detect and quantify model structural errors, unless an additional error decomposition is carried out [e.g., *Renard et al*., 2011]. In other words, the diagnostic power of such likelihood functions for model selection is limited [*Gupta et al*., 2008]. This important limitation provides further motivation for incorporating multiple hydrologically oriented model diagnostics into hydrological model selection and evaluation studies.

7.2. Hydrologically Oriented Model Diagnostics

Given the limitations of model comparison based on single criteria such as the AIC, additional metrics with hydrological interpretation are used for a more thorough model comparison:

1. The Nash-Sutcliffe coefficient of efficiency (NSE), which is widely used in the hydrological literature and therefore enables direct comparison with other studies;

2. The differences between modeled and observed annual total flow volume, which is a measure of the catchment water balance;

3. Daily-scale flow-duration curves, which allow comparing the probability distributions of observed and modeled flows and can provide a visual indication of potential biases (e.g., compensating behavior with overestimation of low flows and underestimation of high flows). We consider stratified flow-duration curves for: (i) all flows throughout the year; (ii) flows in individual seasons; and (iii) flows in the rising and falling limb of the hydrographs.

This list is not intended as a comprehensive set of diagnostics for hydrological model evaluation. In addition to general metrics, the diagnostics should reflect any specific modeling goals. We refer the reader to *Martinez and Gupta* [2011] for further details.

8. Results

This section examines the performance of the 22 hydrological models (Table 1) over the exploratory and confirmatory periods (section 3). For convenience, the comparison makes use of the model structure groupings described in section 5.3. The impact of thinning the streamflow set used in the calibration (section 6.3.2) is also investigated.

Figure 4 shows the AIC differences, the residual error parameters (a_ε and b_ε), the NSE and the magnitude of the groundwater flux calculated over the exploratory period using streamflow set $\tilde{\mathbf{y}}^{(>0.09)}$. The AIC differences when estimating parameters using streamflow set $\tilde{\mathbf{y}}^{(>0.09)}_{t:=1+6j}$ and the AIC values for the AIC-best model in each model structure grouping are also shown.

8.1. Model Grouping $g_{1.x}$: Nonstationary GR4J

The results for the model grouping $g_{1.x}$ are presented as red bars in Figure 4. When calibrating to streamflow set $\tilde{\mathbf{y}}^{(>0.09)}$, the best AIC value is achieved by model $g_{1.8}$, which is the most complex model in the comparison and includes all forms of nonstationarity. In contrast, when calibrating to streamflow set $\tilde{\mathbf{y}}^{(>0.09)}_{t:=1+6j}$, model $g_{1.6}$ is the AIC-best model, with model $g_{1.8}$ very closely behind. The only difference between these two models is that $g_{1.8}$ has the linear trend in parameter θ_1.

8.1.1. Interpretation of the AIC Weights

The AIC-best model estimated using the streamflow set $\tilde{\mathbf{y}}^{(>0.09)}$ has an AIC weight of 0.994, while the second-best model has a weight of 0.006. In contrast, the AIC-best and AIC-second best models estimated using streamflow set $\tilde{\mathbf{y}}^{(>0.09)}_{t:=1+6j}$ have near-equal weights of 50.1 and 49.9, respectively, with almost no weight for the remaining models. This would indicate that the remaining models have almost no probability of being selected as AIC-best under an independent confirmatory period, but it is unlikely that this interpretation holds in this case. This is because the assumption of independence in the model residuals is unlikely to hold exactly, even when sampling every sixth day. Other deficiencies in the likelihood function, including the GR4J hydrological model, the Gaussian distribution and linear heteroscedasticity of the residual errors, may also affect the "good model" assumption underlying the AIC and reduce its interpretability. Despite these limitations, the relative magnitude of AIC differences in Figure 4 is instructive, and suggests which

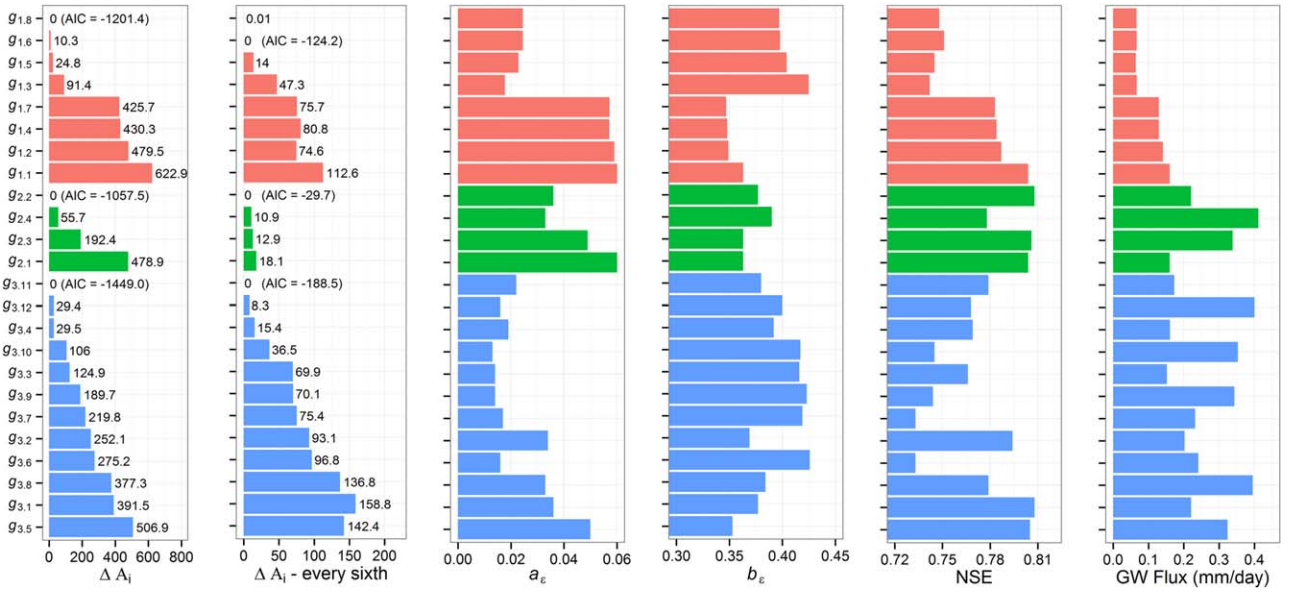

Figure 4. Model comparison in the exploratory period: Akaike differences ($\Delta\mathcal{A}_i$) when using every day and every sixth day in the likelihood function, residual error model parameters (a_ε and b_ε), Nash Sutcliffe coefficient of efficiency (NSE) and groundwater flux for all models. The red, green, and blue colors indicate the model structure groupings $g_{1.x}$, $g_{2.x}$, and $g_{3.x}$, respectively. Within each grouping, the models are ordered from best to worst performance, as given by the AIC differences.

model modifications are responsible for the greatest improvements in model performance. For example, comparison of models $g_{1.2}$, $g_{1.3}$, and $g_{1.4}$ shows that the sinusoid representation of the seasonal-scale non-stationarity in θ_1 delivers by far the greatest improvement in predictive ability.

8.1.2. Increasing Trend in Parameter θ_1

Figure 5 shows the time variation of parameter θ_1 (i.e., the catchment storage capacity) and the actual storage in the production store for the two AIC-best models, $g_{1.6}$ and $g_{1.8}$, over the exploratory period. The sinusoidal variation is prominent for both models. There is also an apparent trend, with higher values of the production store observed in the second half of the record. The magnitude of this trend is similar regardless of whether a linear trend is included ($g_{1.8}$) or not included ($g_{1.6}$) as one of the covariates. It is likely that covariation exists between parameters λ_2 (representing the linear trend) and λ_5 (representing the previous 365 day PET) as a trend was found in PET (section 3), and this could explain the similarity in performance between the two models. In both models, the increase in θ_1 over the exploratory period means that the responsiveness of the catchment to rainfall is decreasing through time (as a larger storage capacity provides a stronger damping of the effects of rainfall variability on the streamflow).

The actual water level in the production store is highly seasonal, with the store reaching a maximum value in late winter/early spring, and a minimum value in summer. This is not surprising given the seasonal nature of rainfall and PET in this catchment. More interesting is the timing of the sinusoid function for θ_1, with a maximum value occurring at the beginning of the year and a minimum value occurring in the middle of the year. As the production store affects the catchment responsiveness and the partitioning of rainfall between actual evapotranspiration and runoff/groundwater recharge, this result suggests that, in summer, the model without a sinusoidal term in θ_1 is overestimating the runoff responses and/or underestimating the actual evapotranspiration flux. The opposite effect is present in the winter predictions.

8.1.3. Other Measures of Model Performance

The residual error model parameters a_ε and b_ε can serve as additional measures of hydrological model performance and are shown in Figure 4. These parameters need to be interpreted jointly, as a_ε describes the standard deviation of the residuals at low flows, while b_ε describes the rate of increase in the standard deviation of the residuals with respect to the predicted flow. Figure 4 shows that, as we consider models with lower AIC, a_ε decreases faster than b_ε. In fact, b_ε for the AIC-best model (b_ε=0.40) is only slightly larger than b_ε for the AIC-worst model (b_ε=0.36). This indicates that the GR4J modifications provide the greatest improvements when predicting low flows.

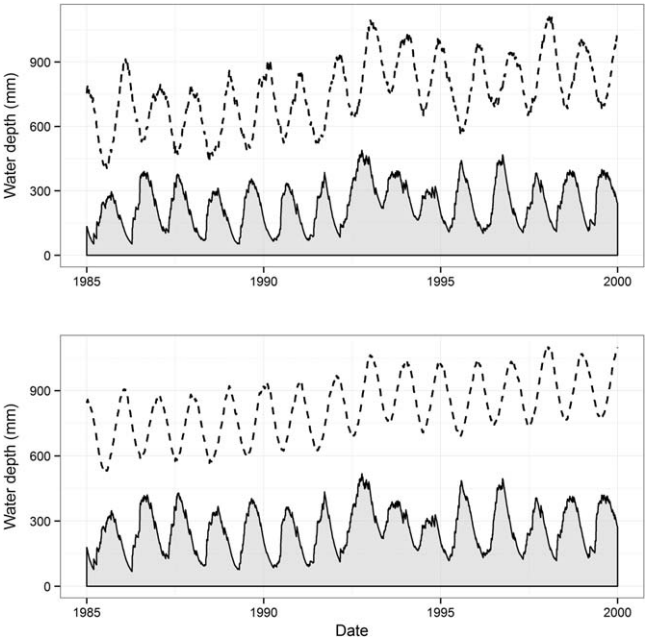

Figure 5. Time series of the production store capacity parameter θ_1 (dotted line) and the actual storage S in the production store (solid gray curves). (top) The results for the AIC-best model ($g_{1.8}$) obtained when calibrating to streamflow set $\tilde{\mathbf{y}}^{(>0.09)}$. (bottom) The results for the AIC-best model ($g_{1.6}$) obtained when calibrating to streamflow set $\tilde{\mathbf{y}}^{(>0.09)}_{t:=1+6j}$; see section 6.3 for a description of the streamflow sets.

The value of θ_1 for model $g_{1.1}$ (the original GR4J model, where θ_1 is constant in time) is approximately 400 mm, which is closer to the winter minimum value of θ_1 when the parameter is allowed to vary sinusoidally. This suggests that the nonstationarity model (equation (1)) provides the largest improvements during periods of low flow, particularly in summer and autumn. This is apparent when examining the autumn flow-duration curves for models $g_{1.1}$, $g_{1.2}$, $g_{1.3}$, and $g_{1.4}$ in Figure 6a, a significant improvement is provided by model $g_{1.3}$, in which θ_1 varies sinusoidally over the year, whereas the improvements are much more limited for the other models.

In contrast to the AIC-based rankings, the NSE yields a very different ranking of models, with model $g_{1.1}$ selected as the "best" model with respect to both streamflow sets $\tilde{\mathbf{y}}^{(>0.09)}$ and $\tilde{\mathbf{y}}^{(>0.09)}_{t:=1+6j}$. The NSE ranking is more consistent with the annually aggregated flow error metric: the models with the lowest error in total annual flow also have the highest NSE values. The relatively close correspondence between the annual flow metric and the NSE is probably due to the highly skewed nature of flows in the catchment, with the majority of flow volume occurring in a relatively small number of wet days, and hence with the NSE being dominated by the model performance in those few high-flow days. In contrast, the AIC is informed by the heteroscedastic likelihood model, which allows for a greater contribution from low flows whose total flow volume is small.

8.2. Model Grouping $g_{2.x}$: Improved the Process Representation

The models in grouping $g_{2.x}$ represent modifications to the recession and ET equations in the GR4J model (except for $g_{2.1}$ which represents the original GR4J). Figure 4 shows that these modifications yield substantial improvements to model performance. Regardless of whether the model was calibrated using streamflow set $\tilde{\mathbf{y}}^{(>0.09)}$ or $\tilde{\mathbf{y}}^{(>0.09)}_{t:=1+6j}$, model $g_{2.2}$ was selected as the AIC-best model, followed by model $g_{2.4}$. Both models contain the additional parameter θ_5, and model $g_{2.4}$ also includes the modified representation of actual evapotranspiration.

In contrast to the model selection results based on the AIC or the inferred parameters of the residual error model, model $g_{2.1}$ is the best in reproducing annual average flows. The most notable difference is the groundwater export volume, with model $g_{2.4}$ having either 2.5 or 1.9 times the groundwater flux compared to model $g_{2.1}$, depending on whether streamflow set $\tilde{\mathbf{y}}^{(>0.09)}$ or $\tilde{\mathbf{y}}^{(>0.09)}_{t:=1+6j}$ were used, respectively. For models $g_{2.3}$ and $g_{2.4}$ the total groundwater export is of a similar magnitude to the streamflow, thereby representing a major component of the catchment water balance.

Figure 6b shows the flow-duration curves for simulated runoff from models $g_{1.1}$ and $g_{2.2}$. The model predictions are adequate for flows greater than the 30% exceedance probability. However, model $g_{2.2}$ clearly outperforms $g_{1.1}$ for lower flows, supporting the earlier conclusion that the largest improvements occur for low flows. To further investigate the models' predictive performance, flow-duration curves were plotted separately for the rising and falling limbs of the hydrographs, as shown in Figures 6c and 6d. The most significant improvements occur in the falling limb. This is not surprising, as θ_5 was specifically

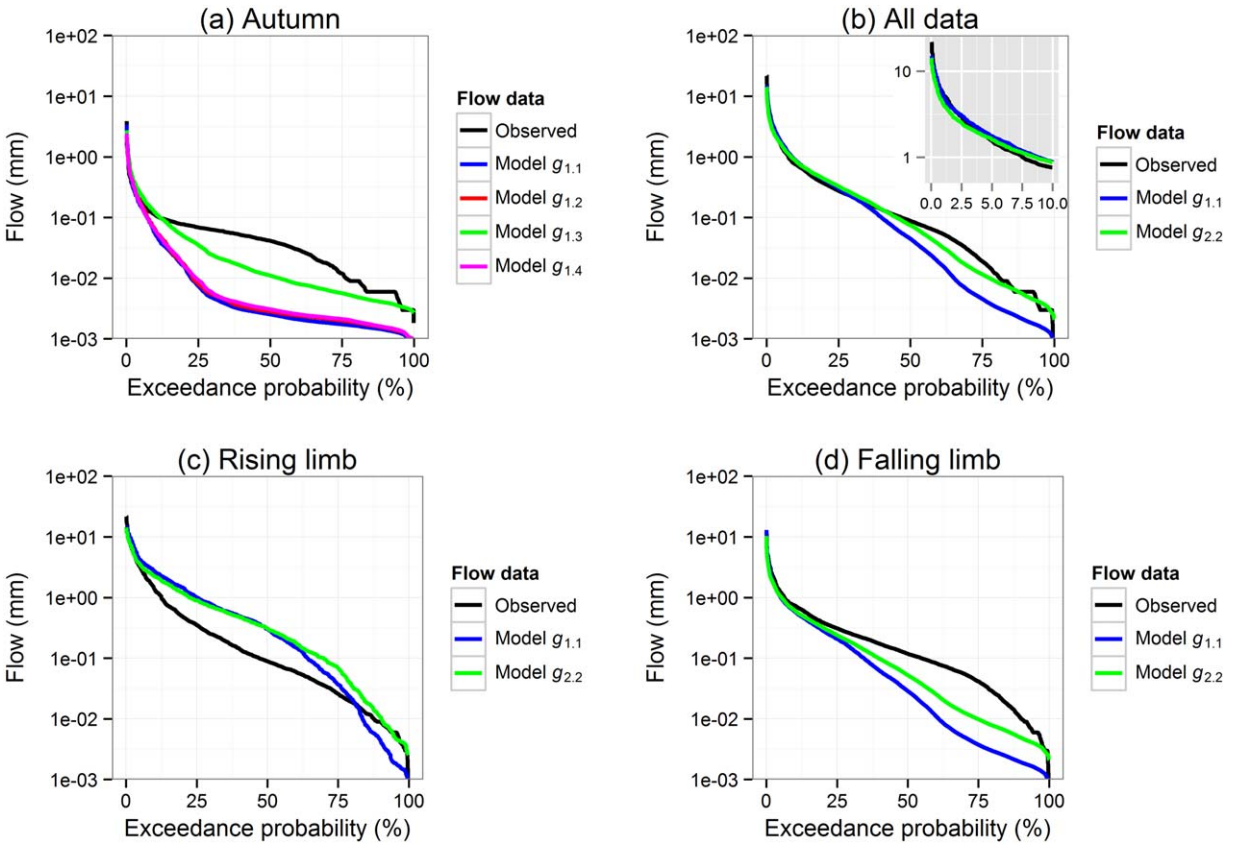

Figure 6. Comparison of observed and simulated flow-duration curves for the exploratory period (1985-1999). (a) Autumn data for models $g_{1.1}$, $g_{1.2}$, $g_{1.3}$ and $g_{1.4}$. (b) All data for models $g_{1.1}$ and $g_{2.2}$, with the inset zoom showing the highest 10% of flow days. (c) Rising limb of the hydrograph for models $g_{1.1}$ and $g_{2.2}$. (d) Falling limb of the hydrograph for models $g_{1.1}$ and $g_{2.2}$.

introduced to improve the shape of hydrograph recessions (by modifying the partitioning of net rainfall into the production and routing store).

8.3. Model Grouping $g_{3.x}$: Combining the Nonstationary GR4J With Improved Process Representation

The models in grouping $g_{3.x}$ combine the nonstationarity characterization of parameter θ_1 with the recession and ET modifications to the standard GR4J model. As noted in section 5.3, the AIC-best model from grouping $g_{2.x}$ (i.e., model $g_{2.2}$) is included in model grouping $g_{3.x}$ as model $g_{3.1}$.

The AIC-based model ranking is almost identical, regardless of whether streamflow set $\tilde{\mathbf{y}}^{(>0.09)}$ or $\tilde{\mathbf{y}}_{t:=1+6j}^{(>0.09)}$ is used. The AIC-best and second-best models ($g_{3.11}$ and $g_{3.12}$, respectively) have the same model structure except for the evapotranspiration term. The third-best model is $g_{3.4}$, which does not use the previous 365 day rainfall and PET as covariates.

Models with the sinusoidal term ($g_{3.3}$, $g_{3.4}$, $g_{3.6}$, $g_{3.7}$, $g_{3.9}$, $g_{3.10}$, $g_{3.11}$, and $g_{3.12}$) generally rank much higher than the models without this term: all six top-ranked models include this term. This suggests that the model modifications described in section 8.2 are not able to eliminate the sinusoidal variation in θ_1. Models with parameter θ_5 perform better than models without this parameter, which is consistent with the results in section 8.2. In contrast, Figure 4 shows that the inclusion of the linear trend in the nonstationarity model of θ_1 has a much smaller effect on the model rankings (i.e., compare models $g_{3.3}$ versus $g_{3.4}$, $g_{3.6}$ versus $g_{3.7}$, and $g_{3.9}$ versus $g_{3.10}$).

Similar to the case for model groupings $g_{1.x}$ and $g_{2.x}$, there is no close relationship between the AIC value of models within the groupings $g_{3.x}$ and their errors in total annual flow volume. This implies that the AIC and likelihood values are not good predictors of annual flow error over the exploratory period. This is not surprising, as the likelihood used in this study is based on a heteroscedastic error model that provides a

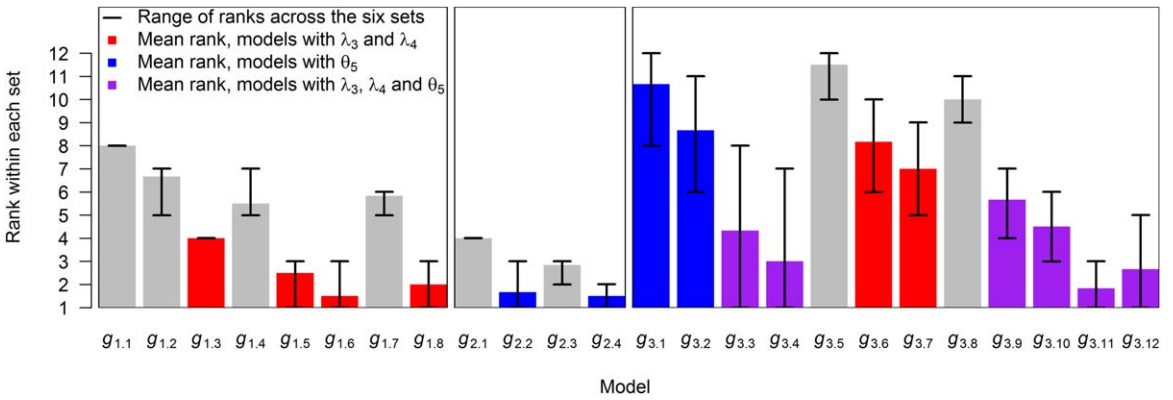

Figure 7. AIC ranking when models are calibrated separately to the six distinct thinned data sets (section 6.3.2). The model sets $g_{1.x}$, $g_{2.x}$, and $g_{3.x}$ are shown in separate figures. Within each figure, the mean rank and the range for each of the six sets is presented. The colors denote an alternative way of grouping the models, based on the presence of particular calibrated parameters (as indicated in the legend).

balanced fit to low and high flows. Consequently, the likelihood function is not overly sensitive to errors in the total flow volume.

8.4. Sensitivity of AIC-Based Model Rankings to the Choice of Thinned Data Set

This section reports the sensitivity of the AIC-based model selection to the choice of thinned data set (see section 6.3.2). Figure 7 shows the model ranks for each of the three sets of models (i.e., $g_{1.x}$, $g_{2.x}$, and $g_{3.x}$) using all six distinct thinned data sets. The colors are used to distinguish between three groupings of model structure that were found to perform similarly (sections 8.1–8.3): (i) models with a sinusoid parameterization of θ_1, (ii) models with the additional parameter θ_5, and (iii) models with both a sinusoid parameterization of θ_1 and the recession parameter θ_5.

The findings show reasonable consistency in the rankings obtained with different streamflow sets, particularly when accounting for the major model structural groupings. In set $g_{1.x}$, models with the sinusoid function consistently outperform those without this function. Model $g_{1.6}$ is the AIC-best for four of the thinned data sets, whereas $g_{1.5}$ and $g_{1.8}$ are each the AIC-best for one thinned data set. Similarly, in set $g_{2.x}$, models $g_{2.2}$ and $g_{2.4}$ consistently outperform the remaining models, except for the sixth thinned data set, in which model $g_{2.3}$ is ranked AIC-second best. Finally, in set $g_{3.x}$, the six models with both the sinusoid function and the recession parameter θ_5 are consistently ranked among the top four models, although there is some variation in their individual rankings. Thus, the conclusions in section 8.3 regarding the covariates with the most dominant influence on model performance appear to be reasonably robust with respect to the choice of thinned data set used in the exploratory period.

8.5. Model Evaluation Over an Independent Confirmatory Period

This section reports the performance of the models over the confirmatory period. The focus of this evaluation is to establish whether the AIC-best models identified over the exploratory period also perform well over the independent confirmatory period where, as discussed in section 3, the annual flows are on average 19% lower than in the exploratory period.

Figure 8 shows the observed and simulated hydrographs for representative half-year subperiods of the exploratory period (top) and confirmatory period (bottom). The prediction intervals are calculated using the estimated residual model parameters a_ε and b_ε. The cooler half-year (May to November) is shown as the majority of annual flow occurs during this period. The figure compares the predictions of the simplest model ($g_{1.1}$) and the AIC-best model ($g_{3.11}$). The models' ability to capture the observed hydrographs is difficult to determine by visual inspection alone, with both models underestimating some days and overestimating other days. Flow-duration curves (Figures 6–8) are arguably better diagnostics for assessing the predictive performance at individual streamflow quantiles. However, it can be seen from Figure 8 that, in the confirmatory period, the simplest model ($g_{1.1}$) significantly overpredicts the observed flows for the majority of flow events, while the AIC-best model ($g_{3.11}$) matches the observations much better. Another noticeable feature is the narrower prediction interval for model $g_{3.11}$, particularly for low flows. This

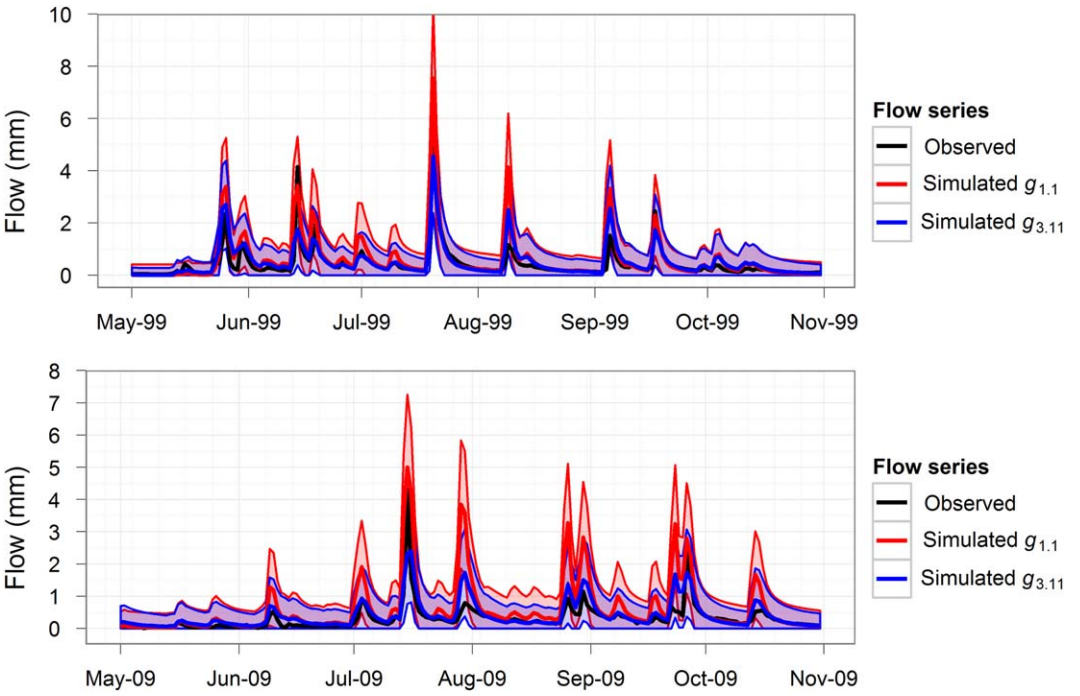

Figure 8. Observed and simulated flow series for a representative year from (top) the exploratory period and (bottom) an independent drier confirmatory period, for the standard GR4J model (model $g_{1.1}$, red lines) and the AIC-best model (model $g_{3.11}$, blue lines). The 90% prediction intervals are shown using shading.

highlights that the nonstationary model yields a significant improvement in predictive precision, while maintaining a good description of residual errors (Figure 3).

Figure 9 presents the performance metrics, namely the likelihood, NSE, and annual average flow volume error, for all models. Note that the AIC is not included in this comparison because the AIC is based on the maximum-likelihood parameter values estimated over the exploratory period, and it cannot be assumed that those parameters will also be the maximum-likelihood estimates over the confirmatory period. Based on likelihood values, the original GR4J model $g_{1.1}$ is the worst performing model in the confirmatory period. This model also incurs the largest errors in predicting the annual average flows, underestimating them by 18%. In contrast, the best model in the confirmatory period is model $g_{3.11}$, and this model underestimates the average flow rate by only 2.6%—a significant improvement. Figure 9 also shows that including a linear trend in θ_1 leads to an underestimation of flows in the confirmatory period (by 6.7% on average), while the absence of this term leads to an overestimation by a similar magnitude (7.7% on average). Potential reasons for this finding are discussed in section 9.3.

Figure 10 shows the AIC calculated over the exploratory period against the likelihood calculated over the confirmatory period, for both the full and thinned data sets. This plot examines the ability of the AIC to predict model performance in the confirmatory period. It can be seen that lower (better) AIC values over the exploratory period are associated with higher (better) log likelihood values in the confirmatory period. This association is statistically significant, with correlation coefficients of -0.66 and -0.44 for the full and thinned data sets, respectively. Therefore, even though the AIC tended to favor more complex models in the exploratory period, this complexity appears to be justified by the data: these more complex models also have the highest likelihood values in the confirmatory period.

9. Discussion: Model Selection for Future Climate Predictions

This section discusses three alternative perspectives for selecting one or more models to be used for prediction. The discussion is not intended as exhaustive (e.g., section 2 lists further approaches).

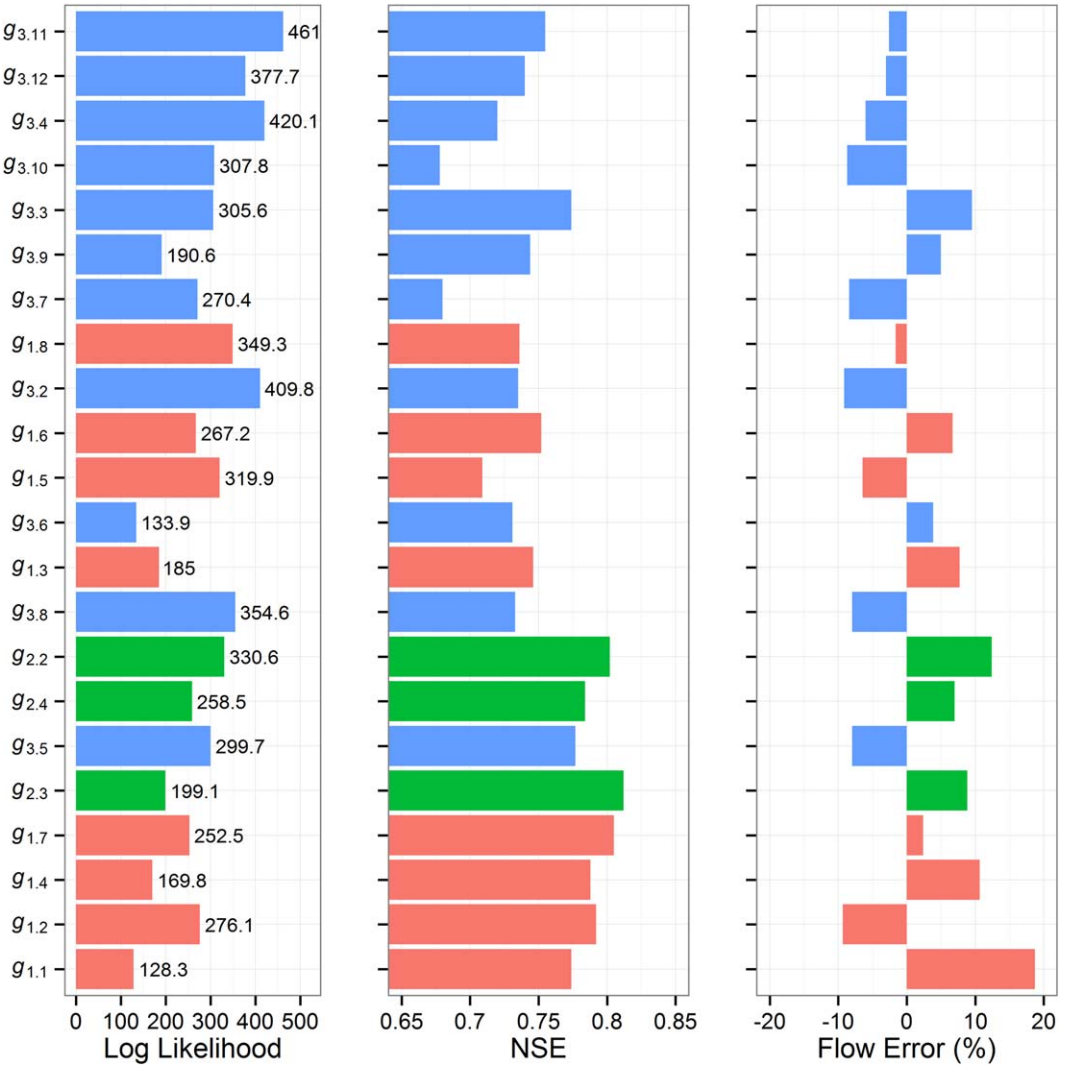

Figure 9. Model likelihood, NSE, and flow error (as a percentage of total annual flow) evaluated during the confirmatory period. Red, green, and blue colors indicate the model structure groupings $g_{1.x}$, $g_{2.x}$, and $g_{3.x}$, respectively. All models are ordered from best to worst performance, as given by the AIC differences over the exploratory period (see Figure 4).

9.1. An Information-Theoretic Approach to Model Selection

Section 6.1 discussed the use of the Kullback-Leibler information as a measure of the information lost when representing environmental processes using a (necessarily approximate) model. As noted by *Burnham and Anderson* [2010], a key theoretical appeal of the AIC is that, given a set of models, it identifies the model that approximately minimizes the Kullback-Leibler information. However, as emphasized in section 6.1, this appealing feature can be undermined if the assumptions in the likelihood function (and thus the AIC) are strongly violated. Note that this includes assumptions in both the deterministic and error models, i.e., AIC-based conclusions may be sensitive to deficiencies in either/both the physical process representation and the statistical description of uncertainty.

Of the 22 models, whether calibrating to data set $\tilde{\mathbf{y}}^{(>0.09)}$ or $\tilde{\mathbf{y}}^{(>0.09)}_{t:=1+6j'}$, model $g_{3.11}$ gives the lowest (best) AIC value, followed by model $g_{3.12}$. Models $g_{3.3}$ and $g_{3.4}$ also perform well in some of the thinned data sets, as shown in Figure 7. Models $g_{3.11}$ and $g_{3.12}$ are the most complex models considered in this work, incorporating all the covariates and differing only in the calculation of the actual ET. Similar results have been found in other studies [e.g., *Engelhardt et al*., 2013], where the AIC tended to favor very complex models when compared to Bayesian selection criteria such as the BIC or KIC. Nevertheless, the model with the lowest (best) AIC in the exploratory period was found to maximize the likelihood in the confirmatory period, although significant scatter is observed (Figure 10).

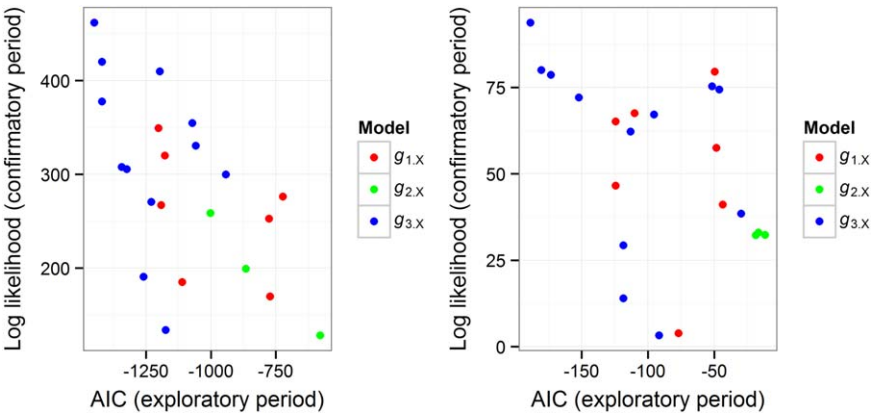

Figure 10. Likelihood function values computed over the confirmatory period (2000–2009) plotted against AIC values computed over the exploratory period (1985–1999). (left) The results for the full data set; (right) the results for the thinned data set. Correlation coefficients are −0.66 and −0.44, respectively, which are statistically significant at the 5% level.

The problems with applying the AIC weights in cases where the hydrological and error model assumptions are not met are demonstrated by comparing the weights of the candidate models. In the case where streamflow set $\tilde{\mathbf{y}}^{(>0.09)}$ is used for parameter estimation, the model with highest AIC rank has an Akaike weight of close to one, while all other models have weights close to zero. This may be due to the omission of error autocorrelation from the likelihood function, which results in an inference that overestimates the information content of the data. If streamflow set $\tilde{\mathbf{y}}^{(>0.09)}_{t:=1+6j}$ is used, the residual error autocorrelation is no longer statistically significant at the 5% significance level, yet the AIC weight of the preferred model decreases only slightly, to 0.98. Since the AIC is derived under the assumption that the entire predictive model (here, GR4J and the linear heteroscedastic error model) is a sufficiently "good" approximation of the real system, it may be that the AIC is affected more by deficiencies in the hydrological model (i.e., in GR4J and its variants) than by deficiencies in the error model. This reinforces the need to improve the specification of likelihood functions in hydrological modelling and understand the sensitivity of AIC weights to violations in the deterministic and stochastic components of the likelihood function. The use of "hydrologically meaningful" measures of model performance is hence of clear importance, as described next [see *Martinez and Gupta*, 2011].

9.2. A Multiple Diagnostics Approach to Model Selection

The limitations of single-metric approaches can be reduced by using multiple "hydrologically meaningful" diagnostics [e.g., see *Legates and McCabe*, 1999; *Martinez and Gupta*, 2011]. These diagnostics can be constructed to scrutinize the ability of the model to reproduce specific hydrological features of interest. For example, in this study, seasonal flow-duration curves were used to establish that model $g_{1.3}$ (for which parameter θ_1 is allowed to vary sinusoidally over the year) outperforms models based on other representations of nonstationarity (interannual, etc.). From a physical perspective, this can be attributed to the seasonality of the catchment, with summer being water-limited and winter being energy-limited. Flow-duration curves also helped to establish that, of all GR4J modifications considered in this study, the introduction of parameter θ_5 to control the portion of net rainfall directed to the production store yields the largest improvement in the simulation of hydrograph recessions.

The annual flow error (or bias) is another useful diagnostic, given its obvious relevance for studies such as reservoir yield analyses. However, in this study, the predictive power of this statistic appears limited, with no statistically significant correlation between the flow errors in the confirmatory versus exploratory periods. These results indicate that a "good" model in terms of overall mass balance over a calibration period may not be a "good" model when applied in prediction.

In contrast to the flow error, the NSE performed better as a diagnostic tool, with a statistically significant correlation between the NSE in the exploratory and confirmatory periods. In contrast to the AIC, the NSE generally favored simpler models, such as model $g_{2.2}$ (followed closely by $g_{2.3}$ and $g_{2.1}$) when calibrating to the streamflow set $\tilde{\mathbf{y}}^{(>0.09)}$, and $g_{1.1}$ when calibrating to the streamflow set $\tilde{\mathbf{y}}^{(>0.09)}_{t:=1+6j}$. As a result, in this study,

the models favored by the NSE are very different to the models favored by the AIC. This is likely to be due to the different weighting of low and high flows in the heteroscedastic likelihood (which attempts to balance the fitting of low and high flows) versus the NSE metric (which is generally insensitive to low flows).

Given that different models are favored by different metrics, it is unclear how to best use multiple diagnostics for model selection and for constructing a multimodel ensemble. Which models should be included in the ensemble, and how should they be weighted?

9.3. Use of Independent Information to Assist in Model Selection

In many cases, information on a particular catchment may be difficult to include directly into a hydrological modeling framework, but may nevertheless enable the physical realism of the model predictions to be assessed against the empirical evidence. This is referred to as the "principle of hydrological consistency" in *Martinez and Gupta* [2011].

In this study, the observed trend in parameter θ_1 might at least partially be explained by independent evidence suggesting an increase in farm dams in the catchment. In particular, the report by *Teoh et al*. [2002] shows that no farm dams were present in the catchment in 1987, increasing to 140 farm dams with a total storage volume of 118 mL in 1996, and to 161 farm dams with a total storage volume of 148 mL in 1999. The 1999 volume equates to a catchment-averaged depth of 5.1 mm, and represents 4% of the annual average catchment discharge over the exploratory period. Controls on the development of new farm dams have been instigated in the early 2000s [*Teoh*, 2002], and it is therefore likely that the total storage volume of farm dams would not have increased substantially since that time. Interestingly, during the confirmatory period all the models without a trend in θ_1 overestimated total annual flows, whereas all the models with a trend underestimated total annual flows (see section 8.5). This is consistent with the independent evidence on trends in farm dams, however other changes (e.g., groundwater extraction due to agricultural activities) may also have occurred over this time, and cannot be ruled out as alternative potential physical causes of the nonstationarity in θ_1.

Catchment groundwater flux is an alternative source of information that can be used to evaluate hydrological consistency. In most models calibrated in this study, groundwater represents an important component of the water balance, although the total groundwater flux estimates varies substantially between models, ranging from 0.064 to 0.411 mm/d (Figure 4). The best available estimate of groundwater export (calculated as net recharge minus base flow) was approximately 995 mL/yr (0.094 mm/d) when averaged over the 30 year period from 1975 to 2004, although the estimates are very approximate and uncertainty estimates are not available [*Adelaide and Mount Lofty Ranges Natural Resources Board*, 2013]. Therefore, available evidence on the groundwater flux is consistent with the modeling results presented here, in that all evidence points to a groundwater export. However, more detailed estimates of groundwater fluxes are needed before individual models can be more confidently excluded from the analysis.

10. Conclusions

This paper proposes and illustrates a strategy for diagnosing and interpreting hydrological nonstationarity. The major aim is to improve the ability of a hydrological model to provide extrapolative predictions under changing hydroclimatic conditions, since future hydroclimatic conditions may be outside of the domain of the data used for model selection and parameter estimation.

The strategy consists of four elements: (1) detecting, and where possible, eliminating, systematic errors in data; (2) allowing one or more hydrological model parameters to vary in time as functions of covariates intended to capture the relevant time scales of hydrological model nonstationarity (e.g., seasonal, annual, and interannual); (3) trialing alternative model structures, with the aim of reducing hydrological model nonstationarity; and (4) model selection and evaluation including the combined use of information-theoretic metrics (such as the AIC) and hydrologically oriented diagnostics (such as flow-duration curves).

The strategy is illustrated for a small catchment in South Australia, using the GR4J hydrological model as the initial hypothesis. A heteroscedastic error model likelihood is applied to a thresholded and thinned data set to reduce the impact of low flows and residual error autocorrelation, respectively. An exploratory period

is used for model calibration and selection, and a confirmatory period that is much drier than the exploratory period is used to test whether the models are robust under extrapolation.

The key conclusions of implementing the nonstationarity analysis strategy in the case study are:

1. Improved model predictions are obtained when the GR4J storage capacity parameter (θ_1) is made dependent on covariates describing seasonality, annual variability, and longer-term trends. No systematic errors were found in the calibration data itself, suggesting that the nonstationarity model of θ_1 is compensating for structural errors in how the model represents changes in the hydrological dynamics of the catchment.

2. The model selection analysis highlights the impact of the choice of model evaluation metrics and methodology. The AIC approach often reports a strong difference between models, compared to the NSE metric which has a much lower discriminatory power. Hydrological models with low AIC values in the exploratory period also perform well in terms of the AIC in the confirmatory period. In contrast, models selected using the NSE in the exploratory period performed poorly over the confirmatory period.

3. Hydrologically oriented model diagnostics, such as the flow-duration curves (stratified by season, rising and falling hydrograph limbs, etc.), are useful for detecting model weaknesses. For example, they can help detect systematic biases in predictions of low and high flows, motivate and guide changes in the model representation of recessions and actual evapotranspiration, and so on.

4. Overall, reasonable improvements in predictive performance are achieved: whereas the original GR4J model overestimates annual average flows in the confirmatory period by 18%, the best-performing modified models (incorporating parameter non-stationarity and other structural changes) underestimate the flows by only 3–7%.

When using the inferred nonstationarity models for developing streamflow projections for a future climate, scientific judgement is still required to estimate how the identified parameter trends might continue over time. For example, in this study, the identified trend of increasing model storage capacity could be tentatively explained by an increase in farm dams within the catchment, although other hypotheses such as changes in vegetation dynamics or groundwater extractions could not be excluded. Given this uncertainty, projections should be based on an ensemble of possible models, encompassing a range of possible future changes to catchment stores. This offers the best chance to adequately capture the uncertainty in future catchment behavior.

Future research is recommended on: (1) extending the nonstationarity approach to multiple model parameters, to detect and quantify nonstationarity across nonnested models (e.g., models that do not share common parameters); (2) further exploring the AIC-based model selection methodology and comparing its results to other selection approaches such those identified in section 2; and (3) applying the nonstationary approaches and model selection strategy to flexible model structures such as FUSE [*Clark et al*., 2008] and SUPERFLEX [*Fenicia et al*., 2011; *Kavetski and Fenicia*, 2011], with the aim of finding model structures that minimize parameter nonstationarity.

Acknowledgments
This research was funded by the Goyder Institute for Water Research as part of the project: *C.1.1 Development of an agreed set of climate projections for South Australia*, and their support is gratefully acknowledged. We also wish to acknowledge Associate Editor Jasper Vrugt, Hoshin Gupta, and an anonymous reviewer, for constructive feedback that has helped improve the manuscript.

References

Adelaide and Mount Lofty Ranges Natural Resources Board (2013), *Water Allocation Plan: Western Mount Lofty Ranges*, Government of South Australia.

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Trans. Autom. Control*, *19*(6), 716–723.

Allen, R. G., L. S. Pereira, D. Raes, and M. Smith (1998), Statistical anlaysis of weather datasets, in *Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements*, Food and Agric. Organ. of the U. N.

Anderson, M. P., and W. W. Woessner (1992), The role of the postaudit in model validation, *Adv. Water Resour.*, *15*(3), 167–173.

Bates, B. C., Z. W. Kundzewicz, S. Wu, and J. P. Palutikof (2008), *Climate Change and Water*. Technical Paper of the Intergovernmental Panel on Climate Change, 210 pp., IPCC Secretariat, Geneva.

Bergstrom, S. (1995), The HBV model, in *Computer Models of Watershed Hydrology*, edited by V. P. Singh, pp. 443–476, Highlands Ranch, Water Resources Publications, Colo.

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

Beven, K., R. Lamb, P. F. Quinn, R. Romanowicz, and J. Freer (1995), TOPMODEL, in *Computer Models of Watershed Hydrology*, edited by V. P. Singh, Highlands Ranch, Colo.

Brigode, P., L. Oudin, and C. Perrin (2012), Hydrological model parameter instability: An additional uncertainty in estimating the hydrological impacts of climate change?, *J. Hydrol 476*, 410–425.

Burnham, K. P., and D. R. Anderson (2010), *Model Selection and Multimodel Inference*, Springer, N. Y.
Chamberlain, T. C. (1890), The method of multiple working hypotheses, *Science*, *15*(92).
Chiew, F. H. S. (2006), An overview of methods for estimating climate change impact on runoff, paper presented at 30th Hydrology and Water Resources Symposium, Engineers Australia, Launceston, Tasmania.
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*, 316–336.
Claeskens, G., and N. L. Hjort (2008), *Model Selection and Model Averaging*, Cambridge Univ. Press, Cambridge, U. K.
Clark, M. P., A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener, and L. E. Hay (2008), Framework for understanding structural errors (FUSE): A modular framework to diagnose differences between hydrological models, *Water Resour. Res.*, *44*, W00B02, doi:10.1029/2007WR006735.
Clark, M. P., D. Kavetski, and F. Fenicia (2011), Pursuing the method of multiple working hypotheses for hydrological modelling, *Water Resour. Res.*, *47*, W09301, doi:10.1029/2010WR009827.
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.
Dai, Z., A. Wolfsberg, P. Reimus, H. Deng, E. Kwicklis, M. Ding, D. Ware, and M. Ye (2012), Identification of sorption processes and parameters for radionuclide transport in fractured rock, *J. Hydrol.*, *414–415*, 516–526.
de Vos, N. J., T. H. M. Rientjes, and H. V. Gupta (2010), Diagnostic evaluation of conceptual rainfall-runoff models using temporal clustering, *Hydrol. Processes*, *24*, 2840–2850.
Engelhardt, I., J. G. De Aguinaga, H. Mikat, C. Schuth, and R. Liedl (2013), Complexity vs simplicity: Groundwater model ranking using information criteria, *Ground Water*, in press.
Evin, G., D. Thyer, D. Kavetski, D. Mclnerney, and G. Kuczera (2014), Comparison of joint versus postprocessor approaches for hydrological uncertainty estimatino accounting for error autocorrelation and heteroscedasticity, *Water Resour. Res.*, *50*, 2350–2375, doi:10.1002/2013WR014185.
Fenicia, F., D. Kavetski, and H. H. G. Savenije (2011), Elements of a flexible approach for conceptual hydrological modelling: 1. Motivation and theoretical development, *Water Resour. Res.*, *47*, W11510, doi:10.1029/2010WR010174.
Gan, T. Y., and S. J. Burges (1990), An assessment of a conceptual rainfall-runoff model's ability to represent the dynamics of small hypothetical catchments: 2. Hydrologic responses for normal and extreme rainfall, *Water Resour. Res.*, *26*(7), 1605–1619.
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*, 149–161.
Guerrero, J.-L., I. K. Westerberg, S. Halldin, C.-Y. Xu, and L.-C. Lundin (2012), Temporal variability in stage-discharge relationships, *J. Hydrol.*, *446*, 90–102.
Gupta, H. V., T. Wagener, and Y. Liu (2008), Reconciling theory with observations: Elements of a diagnostic approach to model evaluation, *Hydrol. Processes*, *22*, 3802–3813.
Hastie, T., R. Tibshirani, and J. Friedman (2009), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd edition, 745 p., Springer, USA.
Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999), Bayesian model averaging: A tutorial, *Stat. Sci.*, *14*(4), 382–417.
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.
Kashyap, R. L. (1982), Optimal choice of AR and MA part sin autoregressive moving average models, *IEEE Trans. Pattern Anal. Mach. Intell.*, *4*(2), 99–104.
Kavetski, D., and F. Fenicia (2011), Elements of a flexible approach for conceptual hydrological modelling: 2. Application and experimental insights, *Water Resour. Res.*, *47*, W11511, doi:10.1029/2011WR010748.
Kavetski, D., F. Fenicia, and M. P. Clark (2011), Impact of temporal data resolution on parameter inference and model identification in conceptual hydrological models: Insights from an experimental catchment, *Water Resour. Res.*, *47*, W05501, doi:10.1029/2010WR009525.
Kennedy, M. C., and A. O'Hagan (2001), Bayesian calibration of computer models, *J. R. Stat. Soc., Ser. B*, *63*(3), 425–464.
Klemes, V. (1986), Operational testing of hydrological simulation models, *Hydrol. Sci. J.*, *31*(1), 13–24.
Kuczera, G., D. Kavetski, S. W. Franks, and M. Thyer (2006), Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterizing model error using storm-dependent parameters, *J. Hydrol.*, *331*(1–2), 161–177.
Lavery, B., A. Kariko, and N. Nicholls (1992), A historical rainfall dataset for Australia, *Aust. Meteorol. Mag.*, *40*, 33–39.
Le Lay, M., S. Galle, G. M. Saulnier, and I. Braud (2007), Exploring the relationship between hydroclimatic stationarity and rainfall-runoff model parameter stability: A case study in West Africa, *Water Resour. Res.*, *43*, W07420, doi:10.1029/2006WR005257.
Legates, D. R., and G. J. McCabe (1999), Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation, *Water Resour. Res.*, *35*(1), 233–241.
Lin, Z., and M. B. Beck (2007), On the identification of model structure in hydrological and environmental systems, *Water Resour. Res.*, *43*, W02402, doi:10.1029/2005WR004796.
Marshall, L., D. Nott, and A. Sharma (2005), Hydrological model selection: A Bayesian alternative, *Water Resour. Res.*, *41*, W10422, doi:10.1029/2004WR003719.
Marshall, L., D. Nott, and A. Sharma (2007), Towards dynamic catchment modelling: A Bayesian hierarchical mixtures of experts framework, *Hydrological Processes*, *21*(7), 847–861.
Martinez, G. F., and H. V. Gupta (2011), Hydrologic consistency as a basis for assessing complexity of monthly water balance models for the continental United States, *Water Resour. Res.*, *47*, W12540, doi:10.1029/2011WR011229.
McMahon, T. A., M. C. Peel, L. Lowe, R. Srikanthan, and T. R. McVicar (2013), Estimating actual, potential, reference crop and pan evaporation using standard meteorological data: A pragmatic synthesis, *Hydrol. Earth Syst. Sci.*, *17*, 1331–1363.
McQuarrie, A. D. R., and C.-L. Tsai (2007), *Regression and Time Series Model Selection*, World Scientific Pub. Co. Inc., Singapore.
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.
Milly, P. C. D., J. Betancourt, M. Falkenmark, R. M. Hirsch, W. Zbigniew, Z. W. Kundzewicz, D. P. Lettenmaier, and R. J. Stouffer (2008), Stationarity is dead: Whither water management?, *Science*, *319*, 573–574.
Molini, A., L. G. Lanza, and P. La Barbera (2005), The impact of tipping bucket rain gauge measurement errors on design rainfall for urban-scale applications, *Hydrol. Processes*, *19*(5), 1073–1088.
Morton, F. I. (1983), Operational estimates of areal evapotranspiration and their significance to the science and practice of hydrology, *J. Hydrol.*, *66*, 1–76.

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.
Paik, K., J. H. Kim, H. S. Kim, and D. R. Lee (2005), A conceptual rainfall-runoff model considering seasonal variation, *Hydrol. Processes*, *19*, 3837–3850.
Pathiraja, S., S. Westra, and A. Sharma (2012), Why continuous simulation? The role of antecedent moisture in design flood estimation, *Water Resour. Res.*, *48*, W06534, doi:10.1029/2011WR010997.
Perrin, C., C. Michel, and V. Andreassian (2003), Improvement of a parsimonious model for streamflow simulation, *J. Hydrol.*, *279*, 275–289.
Renard, B., D. Kavetski, E. Leblois, M. Thyer, G. Kuczera, and S. W. Franks (2011), Toward a reliable decomposition of predictive uncertainty in hydrological modelling: Characterizing rainfall errors using conditional simulation, *Water Resour. Res.*, *47*, W11516, doi:10.1029/2011WR010643.
Schoups, G., and J. A. Vrugt (2010), A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroskedastic and non-Gaussian errors, *Water Resour. Res.*, *46*, W10531, doi:10.1029/2009WR008933.
Schoups, G., N. van de Giesen, and H. H. G. Savenije (2008), Model complexity control for hydrological prediction, *Water Resour. Res.*, *44*, W00B03, doi:10.1029/2008WR006836.
Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Stat.*, *6*(2), 461–464.
Seiller, G., F. Anctil, and C. Perrin (2012), Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions, *Hydrol. Earth Syst. Sci.*, *16*, 1171–1189.
Smith, T. J., A. Sharma, L. Marshall, R. Mehrotra, and S. A. Sisson (2010), Development of a formal likelihood function for improved Bayesian inference of ephemeral catchments, *Water Resour. Res.*, *46*, W12551, doi:10.1029/2010WR009514.
Sorooshian, S. (1981), Parameter estimation of rainfall-runoff models with heteroskedastic straemflow errors—The noninformative data case, *J. Hydrol.*, *52*(1–2), 127–138.
Sorooshian, S., and J. A. Dracup (1980), Stochastic parameter estimation procedures for hydrological rainfall-runoff models: Correlated and heteroscedastic error cases, *Water Resour. Res.*, *16*(2), 430–442.
Sugiura, N. (1978), Further analysis of the data by Akaike's information criterion and the finite corrections, *Commun. Stat. Theory Methods*, *A7*, 13–26.
Teoh, K. S. (2002), Estimating the impact of current farm dams development on the surface water resources of the Onkaparinga River Catchment, report, Dep. of Water, Land and Biodiversity Conserv, Government report by the Government of South Australia.
Thyer, M., B. Renard, D. Kavetski, G. Kuczera, S. W. Franks, and R. Srikanthan (2009), Critical evaluation of parameter consistency and predictive uncertainty in hydrological modelling: A case study using Bayesian total error analysis, *Water Resour. Res.*, *45*, W00B14, doi:10.1029/2008WR006825.
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*, 447–457.
Wagener, T., N. R. McIntyre, M. J. Lees, H. S. Wheater, and H. V. Gupta (2003), Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis, *Hydrol. Processes*, *17*, 455–476.
Weijs, S., G. Schoups, and N. de Giesen (2010), Why hydrological predictions should be evaluated using information theory, *Hydrol. Earth Syst. Sci.*, *14*(12), 2545–2558.
Westra, S., J. P. Evans, R. Mehrotra, and A. Sharma (2013), A conditional disaggregation algorithm for generating fine time-scale rainfall data in a warmer climate, *J. Hydrol.*, *479*, 86–99.
Wu, K., and C. A. Johnston (2007), Hydrologic response to climate variability in a Great Lakes Watershed: A case study with the SWAT model, *J. Hydrol.*, *337*, 187–199.
Ye, M., P. D. Meyer, and S. P. Neuman (2008), On model selection criteria in multimodel analysis, *Water Resour. Res.*, *44*, W03428, doi:10.1029/2008WR006803.
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.
Young, P. (1998), Data-based mechanistic modelling of environmental, ecological, economic and engineering systems, *Environ. Modell. Software*, *13*, 105–122.
Young, P., and K. Beven (1994), Data-based mechanistic modelling and the rainfall-flow non-linearity, *Environmetrics*, *5*, 335–363.
Zhang, H., G. H. Huang, D. Wang, and X. Zhang (2011), Multi-period calibration of a semi-distributed hydrological model based on hydroclimatic clustering, *Adv. Water Resour.*, *34*, 1292–1303.