Response to the comments by Reviewers

We appreciate the efforts and comments the reviewers have made in the reviewing process of our paper. Thanks for the opportunity to revise our paper. As the reviewer suggested, we have removed R^2 from the objective function. Also we have corrected Equation (20) for the water balance error (WBE). We hope these changes have adequately addressed the reviewer's concern. Supplements include revised manuscript and response to the comments by reviewers.

Comments

The changes you have made to your manuscript based on reviewer comments have been well received by the reviewer. Only some remarks are being made: first is to remove R^2 as an objective function: please consider whether you wish to do so or not. If you would like to maintain its inclusion, then please add some concerns with respect to this objective function.

Response:

We agree with the reviewer. We have removed R^2 from the objective function. The revision has been made. We also update our paper such as descriptions of results and tables according to this change.

--Line 879-881.

Table 3 Summary results of the model calibration under different climatic conditions (*i.e.* dry and wet periods).

Indicator	SIMHYD calibrated on dry period	SIMHYD calibrated on wet period	DWBM calibrated on dry period	DWBM calibrated on wet period
25th NSE	0.84	0.85	0.71	0.77
Median NSE	0.70	0.77	0.58	0.66
75th NSE	0.61	0.68	0.43	0.54
Average NSE	0.70	0.76	0.57	0.65
25th d_1	0.77	0.79	0.71	0.75
Median d_1	0.72	0.76	0.67	0.71
75th d_1	0.70	0.74	0.61	0.68
Average d_1	0.73	0.76	0.65	0.71
25th WBE	22	16	25	24
Median WBE	13	8	15	12
75th WBE	6	4	9	5
Average WBE	14	11	22	17

--Line 894-896.

Table 4 Summary results of the model validation when calibrated under different climatic conditions.

Model	Indicator	dry/dry	dry/wet	wet/dry	wet/wet
SIMHYD	25th NSE	0.72	0.74	0.68	0.77
	Median NSE	0.55	0.64	0.51	0.69
	75th NSE	0.42	0.44	0.41	0.55
	Average NSE	0.57	0.61	0.54	0.66
	25th d_1	0.74	0.78	0.74	0.78
	Median d_1	0.71	0.74	0.70	0.75
	75th d_1	0.66	0.70	0.63	0.72
	Average d_1	0.69	0.73	0.68	0.74
	25th WBE	34	30	39	23
	Median WBE	20	19	28	13
	75th WBE	14	8	16	7
	Average WBE	24	21	29	17
DWBM	25th NSE	0.56	0.65	0.51	0.72
	Median NSE	0.46	0.48	0.45	0.61
	75th NSE	0.34	0.35	0.30	0.42
	Average NSE	0.48	0.52	0.45	0.59
	25th d_1	0.69	0.73	0.68	0.74
	Median d_1	0.65	0.69	0.63	0.70
	75th d_1	0.58	0.64	0.56	0.66
	Average d_1	0.62	0.68	0.61	0.69
	25th WBE	35	29	53	25
	Median WBE	22	20	33	18
	75th WBE	15	12	18	11
	Average WBE	27	23	36	19

Comments

Second, please check upon equation 20, as now, indeed, the formula does not refer to the absolute water balance error (MAE) but to the relative mean absolute error: or equation 20 is adjusted appropriately, or, the MAE is used throughout the remainder of the text.

Response:

We have corrected Equation (20) for the water balance error (WBE). The revision has been made. We also update our paper such as descriptions of results and figures according to this change.

--Line 352.

$$WBE = \frac{\sum_{i=1}^{N} \left| Q_{sim,i} - Q_{obs,i} \right|}{\sum_{i=1}^{N} Q_{obs,i}} \times 100\% \tag{19}$$

--Line 943-946.

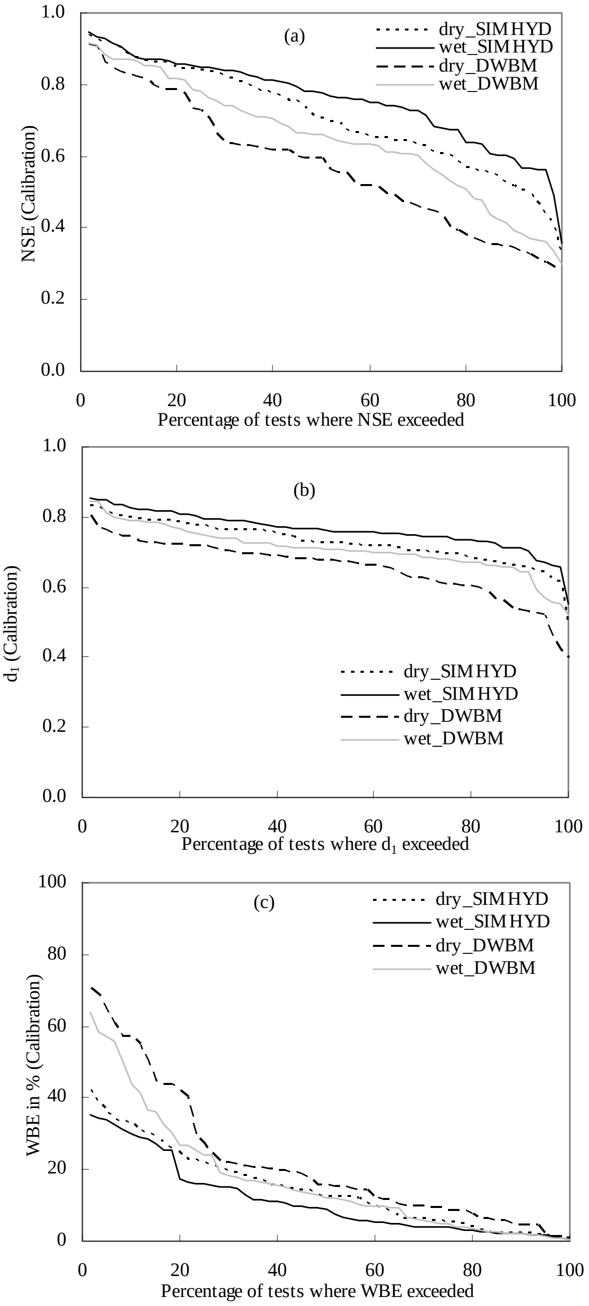
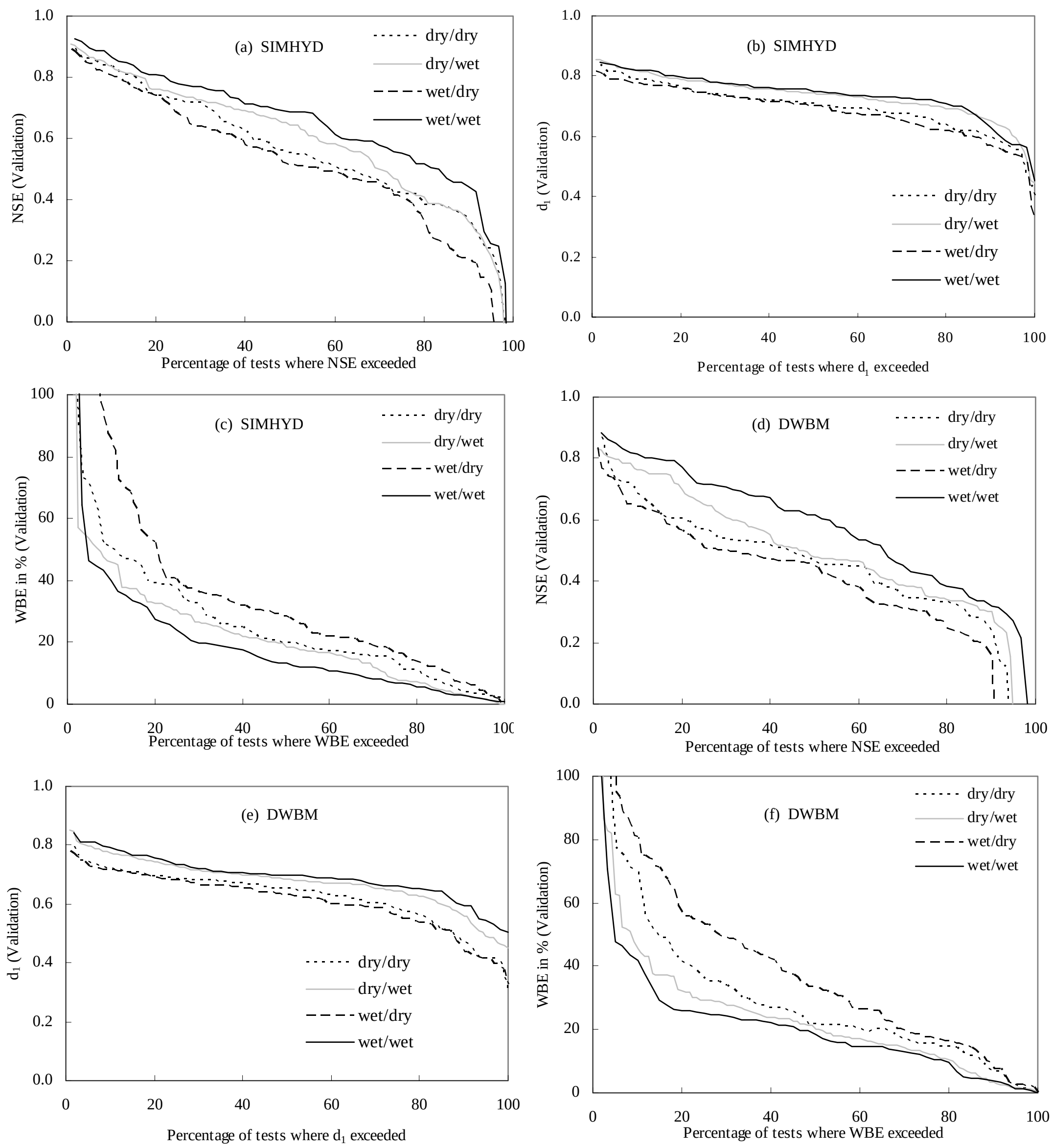


Figure 5 (a) Percentage of model calibration tests with a NSE value greater than or equal to a given NSE value. Similarly, **Figure 5 (b-c)** are corresponding plots of the modified index of agreement (d_1) and the water balance error (*WBE*), respectively.

--Line 947-953.



Figures 6 (a) and (d) Percentage of model validation tests with a NSE value greater than or equal to a given NSE value. Similarly, **Figures 6 (b) and (e)**, **Figures 6 (c) and (f)** are corresponding plots of the modified index of agreement (*d_1*), the water balance error (*WBE*), respectively.

The transferability of hydrological models under nonstationary climatic conditions

Chuanzhe Li[1, 2, 3], Lu Zhang[2, *], Hao Wang[1], Yongqiang Zhang[2], Fuliang Yu[1] and Denghua Yan[1]

[1] State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin, China Institute of Water Resources and Hydropower Research, Beijing 100038, P.R. China

[2] CSIRO Land and Water, Canberra ACT 2601, Australia

[3] State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing 210098, P.R. China

* Corresponding author: Lu Zhang, CSIRO Land and Water, GPO Box 1666, Canberra ACT 2601, Australia

Tel: +61(2)6246-5802

Fax: +61(2)6246-5800

Email: lu.zhang@csiro.au

Submission to: Hydrology and Earth System Sciences

Submission date: September, 2011

Abstract: This paper investigates issues involved in calibrating hydrological models against observed data when the aim of the modelling is to predict future runoff under different climatic conditions. To achieve this objective, we tested two hydrological models, DWBM and SIMHYD, using data from 30 unimpaired catchments in Australia which had at least 60 years of daily precipitation, potential evapotranspiration (PET), and streamflow data. Nash–Sutcliffe efficiency (NSE), modified index of agreement (d_1) and water balance error (WBE) were used as performance criteria. We used a differential split-sample test to split up the data into 120 sub-periods and 4 different climatic sub-periods in order to assess how well the calibrated model could be transferred different periods. For each catchment, the models were calibrated for one sub-period and validated on the other three. Monte Carlo simulation was used to explore parameter stability compared to historic climatic variability. The chi-square test was used to measure the relationship between the distribution of the parameters and hydroclimatic variability. The results showed that the performance of the two hydrological models differed and depended on the model calibration. We found that if a hydrological model is set up to simulate runoff for a wet climate scenario then it should be calibrated on a wet segment of the historic record, and similarly a dry segment should be used for a dry climate scenario. The Monte Carlo simulation provides an effective and pragmatic approach to explore uncertainty and equifinality in hydrological model parameters. Some parameters of the hydrological models are shown to be significantly more sensitive to the choice of calibration periods. Our findings support the idea that when using conceptual hydrological models to assess future climate change impacts, a differential split-sample test and Monte Carlo simulation should be used to quantify uncertainties due to parameter instability and non-uniqueness.

KEY WORDS: Hydrological models; nonstationarity; calibration; validation; climate change

1 Introduction

Climate change caused by increasing atmospheric concentration of greenhouse gases may have significant effects on the hydrological cycle and water availability, hence affecting agriculture, forestry, and other industries (*Rind et al*., 1992; *IPCC*, 2007). Changes in the hydrological cycle may mean more floods and droughts, and increased pressure on water supply and irrigation systems. It is important for us to be able to estimate the potential impact of climate change on water resources and develop sustainable management strategies. One of the challenges in predicting hydrological response to climate change is the issue of hydrological nonstationarity (*Milly et al*., 2008). There are numerous factors that can affect hydrological stationarity and these include vegetation responses to elevated CO_2, changes in land use and rainfall characteristics. It is crucial to improve our understanding of the effect of nonstationarity on hydrological assessments of climate change.

Hydrological models are important tools for predicting the impact of climate change on future water resources and associated socioeconomic impacts. A number of models have been used to evaluate hydrological effects of climate change (*Rind et al*., 1992). Predicting the hydrological impacts of climate change involves two key steps: downscaling the outputs from global climate models (GCMs) and then running hydrological models. At present, outputs from different GCMs have been used to drive hydrological models for predicting streamflow under a changed climate (Chiew *et al*., 2009). There are many factors that can affect the accuracy of a rainfall-runoff model in predicting the hydrological responses to climate change, including the particular hydrological model chosen, the GCM used, the optimisation technique employed, and the calibration period of the model. Most researchers usually use an ensemble of these techniques to minimise the uncertainty in predicting climate change impacts. For instance, *Chiew et al*. (1995) used results from 5 separate GCM experiments and reported that, in certain parts of Australia, the GCMs did not even agree on the direction of change in rainfall (i.e. increasing or decreasing rainfall). *Boorman and Sefton*. (1997) evaluated effects of climate change on mean runoff, flood magnitude, and low flow for 3 catchments in UK using 2 conceptual rainfall–runoff models. In their study, they considered 2 climate scenarios and 8 climate sensitivity tests. *Minville et al*. (2008) produced an uncertainty envelope of future hydrological variables by considering 10 equally weighted climate projections from a combination of 5 GCMs and 2 greenhouse gas emission scenarios. *Monomoy and O'Connor* (2007) used 6 automatic optimisation techniques to calibrate a conceptual rainfall–runoff model, and there have been a number of more recent studies for estimating the impact of climate change on hydrological processes (*Chiew et al*., 2009, *Vaze et al.,* 2010, *Boyer et al*., 2010). An implicit assumption in all these studies is that rainfall–runoff models calibrated over the historical period are valid for predicting the future hydrological regime under a changed climate and this relates directly to the assumption of hydrological stationarity. However, little has been carried out to test the validity of this assumption.

Calibration of hydrological models generally involves optimizing model parameters to match measured streamflow using observed rainfall as input. Performance of the model is usually tested using a simple spilt-sample test, i.e. the model is calibrated for one period of the record and tested for another period. The simple split-sample test may be sufficient for applications where hydroclimatic conditions between the calibration period and validation period are similar. However, when the model needs to be applied to simulate streamflow from periods with different conditions from those in the calibration periods, a more powerful test is required (*Klemes*, 1986, *Xu*, 1999, *Seibert*, 2003). In a recent paper, *Andreassian et al* (2009) used crash test to advocate for more comprehensive model testing in hydrology. For predicting the impact of climate change on streamflow, the input rainfall series are varied according to an assumed future climate scenario and this often means different climatic conditions. But is it appropriate to use these models for future climatic conditions when rainfall–runoff relations could be very different to those experienced historically?

This paper investigates the transferability of hydrological models under nonstationary climatic conditions. We compare results obtained with different hydrological models calibrated under different climatic conditions. The paper first presents two hydrological models chosen for this study – the Dynamic Water Balance Model (DWBM) and the SIMHYD model – and then describes the data used to calibrate them. We describe different methods of applying the data, including a differential split-sample test, a Monte Carlo simulation, and a performance criterion. Finally, we analyse the performance of the models under different calibration conditions and discuss the optimal parameters for each.

2 Description of Hydrological Models and Data

Two lumped hydrological models with daily inputs were chosen for this study: the Dynamic Water Balance Model (DWBM) (*Zhang et al*., 2008) and the SIMHYD model (*Chiew et al*., 2002), and detailed description of the two models is presented below.

2.1 The Dynamic Water Balance Model (DWBM)

The DWBM model used in this study was developed by *Zhang et al*. (2008). It is a lumped conceptual water balance model with two stores: a near surface root-zone store and a deeper zone store (**Figure 1**). The model is based on Budyko's concept of water availability and atmospheric demand (Budyko, 1958) or the concept of "*limits and controls*" (*Calder*, 1998). Fundamental to this model is a functional form that represents a smooth transition between supply and demand limits (*Fu*, 1981):

$$\frac{E}{P} = 1 + \frac{E_0}{P} - \left[1 + \left(\frac{E_0}{P}\right)^w\right]^{1/w} \tag{1}$$

where w is a model parameter ranging between 1 and ∞. For the purpose of model calibration, we define α = 1-1/w so that α varies between 0 and 1. This definition also conveniently associates an increase in α with an increase in evapotranspiration efficiency. P is rainfall and E_0 is potential evapotranspiration at mean annual timescale. More details of this mean annual water balance model are given in *Zhang et al.* (2004) and *Zhang et al.* (2008).

It is assumed that rainfall $P(t)$ in time step t will be partitioned into direct runoff $Q_d(t)$ and catchment rainfall retention:

$$P(t) = Q_d(t) + X(t) \tag{2}$$

where $X(t)$ is called catchment rainfall retention and is the amount of rainfall retained by the catchment for evapotranspiration $ET(t)$, change in soil moisture storage $S(t)$-$S(t-1)$ and recharge $R(t)$.

The demand limit for $X(t)$ is the sum of available storage capacity (S_{max}–$S(t-1)$) and potential evapotranspiration ($E_0(t)$) and is denoted as $X_0(t)$, while the supply limit can be considered as rainfall $P(t)$. Following a similar argument to *Budyko* (1958), we can postulate that:

$$X(t)/P(t) \to 1 \quad as \quad X_0(t)/P(t) \to \infty \quad \text{(very dry conditions)} \tag{3}$$

$$X(t) \to X_0(t) \quad as \quad X_0(t)/P(t) \to 0 \quad \text{(very wet conditions)} \tag{4}$$

The catchment rainfall retention $X(t)$ can be calculated as:

$$X(t) = P(t)F\left(\frac{X_0(t)}{P(t)}, \alpha_1\right) \tag{5}$$

where $F(\,)$ is Fu's curve – equation (1), α_1 is rainfall retention efficiency, i.e., a larger α_1 value will result in more rainfall retention and less direct runoff.

From equations (2) and (5), direct runoff is calculated as:

$$Q_d(t) = P(t) - X(t) \tag{6}$$

At sub-annual time scales, water availability $W(t)$ can be defined as:

$$W(t) = X(t) + S(t-1) \tag{7}$$

Combining the definition of $X(t)$ with equation (7), one obtains:

$$W(t) = ET(t) + S(t) + R(t) \tag{8}$$

While equation (7) defines the source of the water availability, Equation (8) determines the partitioning. Next define evapotranspiration opportunity (*Sankarasubramanian and Vogel*, 2002) as $Y(t) = ET(t) + S(t)$, we obtain:

$$W(t) = Y(t) + R(t) \tag{9}$$

The demand limit for $Y(t)$ can be considered as the sum of potential evapotranspiration ($E_0(t)$) and soil water storage capacity (S_{max}) and is denoted as $Y_0(t)$, while the supply limit is the available water $W(t)$. Similar to *Budyko* (1958), we can postulate that:

$$Y(t)/W(t) \to 1 \quad as \quad Y_0(t)/W(t) \to \infty \quad \text{(very dry conditions)} \tag{10}$$

$$Y(t) \to Y_0(t) \quad as \quad Y_0(t)/W(t) \to 0 \quad \text{(very wet conditions)} \tag{11}$$

The evapotranspiration opportunity $Y(t)$ can be estimated from the following relationship:

$$Y(t) = W(t)F\left(\frac{E_0(t)+S_{\max}}{W(t)}, \alpha_2\right) \tag{12}$$

Thus groundwater recharge $R(t)$ can be calculated from Equation (9). The next step is to calculate evapotranspiration $ET(t)$. The demand limit for $ET(t)$ can be considered as potential evapotranspiration $E_0(t)$ and the supply limit is the available water $W(t)$. Similar to *Budyko* (1958), evapotranspiration $ET(t)$ can be calculated as:

$$ET(t) = W(t)F\left(\frac{E_0(t)}{W(t)}, \alpha_2\right) \tag{13}$$

where α_2 is a model parameter, representing evapotranspiration efficiency.

Soil water storage can now be calculated as:

$$S(t) = Y(t) - ET(t) \tag{14}$$

Finally, groundwater storage is treated as linear reservoir, so that baseflow and groundwater balance can be modelled as:

$$Q_b(t) = dG(t-1) \tag{15}$$

$$G(t) = (1-d)G(t-1) + R(t) \tag{16}$$

where Q_b is baseflow, G is groundwater storage, and d is a recession constant.

The DWBM model has been applied to 265 catchments in Australia and showed encouraging results (*Zhang et al*., 2008). The model has four parameters: retention efficiency(α_1); evapotranspiration efficiency(α_2); soil water storage capacity (S_{max}), and baseflow linear recession constant (d). The range of the parameter values is shown in **Table 1**.

[Figure 1 and Table 1 here]

2.2 The SIMHYD Model

The SIMHYD model is a lumped conceptual daily rainfall–runoff model (*Chiew et al*., 2002), driven by daily rainfall and PET, which simulates daily streamflow. It has been tested and used extensively across Australia (*Chiew et al*., 2002; *Siriwardena et al*., 2006; *Viney et al*., 2008; *Zhang et al*., 2008; *Zhang et al*., 2009). **Figure 2** shows the structure of the SIMHYD model and the algorithms controlling how water enters the system from precipitation, flows into several stores, and then flows out through evapotranspiration and runoff. The SIMHYD model has 7 parameters, and the useful ranges of them are shown in **Table 2**.

[Figure 2 and Table 2 about here]

In the SIMHYD model, daily rainfall is first intercepted by an interception store, which is emptied each day by evaporation. Incident rainfall, which occurs if rainfall exceeds the maximum daily interception, is then subjected to an infiltration function. The incident rainfall that exceeds the infiltration capacity becomes infiltration excess runoff. A soil moisture function diverts the infiltrated water to the river (as saturation excess runoff/interflow), groundwater store (as recharge) and soil moisture store. The saturation excess runoff/interflow is first estimated as a linear function of the soil wetness (soil moisture level divided by soil moisture capacity). The equation used to simulate interflow therefore attempts to mimic both the interflow and saturation excess runoff processes (with soil wetness used to reflect those parts of the catchment that are saturated and from which saturation excess runoff can occur). Groundwater recharge is then estimated, also as a linear function of the soil wetness. The remaining moisture flows into the soil moisture store. Evapotranspiration from the soil moisture store is estimated as a linear function of the soil wetness, but cannot exceed the potential rate (PET minus intercepted water). The soil moisture store has a finite capacity and overflows into the groundwater store, baseflow from which is simulated as a linear recession from the groundwater store. The model has therefore three runoff components: infiltration excess runoff, saturation excess runoff/interflow, and baseflow.

2.3 Study Catchments and Data

In this study 30 catchments from Australia were selected with at least 60 years of unimpaired daily streamflow data (**Figure 3**). Unimpaired streamflow is defined as streamflow that is not subject to regulation or diversion. The catchment area ranges from 82 to 1891 km^2 with mean annual streamflow varied between 53 to 1363 mm. The mean annual precipitation (*P*) ranges from 628 to 2095 mm and annual potential evapotranspiration (*PET*) ranges from 817 to 2098 mm, representing diverse hydrological and climatic conditions. The runoff coefficient varies from 0.08 to 0.65.

Catchment averaged annual rainfall was estimated from gridded SILO daily rainfall (http://www.longpaddock.qld.gov.au/silo, *Jeffrey et al*., 2001). The SILO Data Drill provides surfaces of daily rainfall and other climate data interpolated from point measurements made by the Australian Bureau of Meteorology. The spatial resolution of the gridded daily rainfall data is 0.05 degrees based on interpolation of over 6000 rainfall stations across Australia. The interpolation uses monthly rainfall data, ordinary kriging with zero nugget, and a variable range. Monthly rainfall for each 5 × 5 km grid cell was converted to daily rainfall using daily rainfall distribution from the station closest to the grid cell (*Jeffrey et al*., 2001). The daily time series of maximum and minimum temperatures, incoming solar radiation, actual vapour pressure, and precipitation at 0.05 × 0.05 (~ 5 km × 5 km) grid cells from the SILO Data Drill (http://www.longpaddock.qld.gov.au/silo) were used.

Potential evaporation was calculated using the Priestley-Taylor equation (*Priestley and Taylor*, 1972) for each catchment with the Priestley-Taylor coefficient set to 1.26 following *Raupach* (2000). In the calculation, the available energy was taken as equal to the net radiation by neglecting ground heat flux. The net radiation was calculated from the incoming global shortwave and longwave radiation, surface albedo, surface emissivity, and surface temperature as described by *Raupach et al*. (2001).

Daily streamflow data were obtained from the Australian Land and Water Resources Audit project (*Peel et al*., 2000) and have been quality checked. Firstly, data quality codes were checked for any missing and poor-quality data as most gauging stations provide numerical codes indicating quality of streamflow data. Missing streamflow data were infilled by interpolating streamflow values at previous and following days.

Secondly, time series of daily rainfall and streamflow were plotted to identify any inconsistency and recording errors in the data (e.g. spikes, same streamflow value for a long period of time). The quality checks are to ensure good quality streamflow data are used in the study.

[Figure 3 here]

3 Methods

3.1 Differential Split-sample Test

In general, hydrological models rely on stationary conditions (*Xu*, 1999). Usually, model calibration requires a split-sample test, where the model is calibrated during one climatic period and validated on another independent period. The split-sample test is the classical test, being applicable to cases where there is sufficiently long time series of the climatic data for both calibration and validation and where the catchment conditions remain unchanged, i.e. stationary (*Refsgaard and Storm*, 1996). This test gives an indication how the model might perform for an independent period having similar conditions. Unfortunately, this test is unable to guarantee the applicability of hydrological models under nonstationary conditions (*Xu*, 1999; *Henriksen et al*., 2003).

In order to try to answer the question of whether the transfer of parameter values from the present-day climate to a future climate is justified, the 'differential split-sample test' proposed by *Klemes* (1986) was considered, in which the hydrological model is tested on calibration and validation periods under contrasting climatic conditions. In this case, different sub-periods are chosen with different historical rainfall conditions.

In this study, different periods with various climatic conditions were identified. First of all, we calculated annual and mean annual precipitation over the whole period of record for each catchment. Then sub-periods with consecutive annual precipitation greater than the mean were selected as the “wet” periods and sub-periods with consecutive annual precipitation less than the mean were selected as the “dry” periods. The precipitation in the “wet” periods is 10.2% to 47.1% above the long-term average annual precipitation, while the precipitation in the “dry” periods is 10.4% to 28.3% below the long-term average annual precipitation. In the selection, the minimum length of the sub-period was set to 5 years to ensure stable model calibration. If this process results in more than two “wet” or “dry” periods, then the two wettest periods or two driest periods were selected for model calibration and validation (**Figure 4**). The hydrological model was calibrated for each of the 4 sub-periods and validated on each of the remaining 3 sub-periods in turn, resulting in a total of 12 calibration and validation tests.

To examine model performance under different calibration and validation conditions, results from the above tests are grouped as “**dry**/dry”, “**dry**/wet”, “**wet**/wet”, and “**wet**/dry” to represent climatic conditions in the calibration and validation periods respectively.

[Figure 4 about here]

3.2 Monte Carlo Simulation

It has been widely recognized that hydrological models can perform equally well against measured runoff estimates even with different parameter sets and this so-called parameter equifinality may result in large prediction uncertainty (*Beven*, 1993; *Boorman et al*., 1997; *Niel et al*., 2003; *Wilby et al*., 2005; *Minville et al*., 2008). The parameter equifinality is related to overparamterzation of hydrological models and poor parameter identifiability. For some practical applications, the parameter equifinality problem may not be an issue and any of the parameter sets may be appropriate. However, these equally good parameter sets may give different predictions when the model is used to estimate the effects of climate change and land use change on streamflow (*Uhlenbrook et al*., 1999). The need for improved model calibration and testing has been emphasized in recent years. Monte Carlo simulation is an effective way of calculating confidence limits of predicted time series and exploring parameter stability and identifiability in the context of historic climate variability (*Uhlenbrook et al*., 1999; *Wilby*, 2005; *Widen-Nilsson et al*., 2009).

For each catchment and each calibration period, a Monte Carlo simulation was undertaken with 1,000,000 runs, each with randomly generated parameter values within the given ranges listed in **Tables 1** and **2** for the two models respectively. We then selected assemblies of the 100 best parameter sets for each catchment and each calibration period according to a goodness-of-fit measure which is defined in section 3.3. Finally, the models were run during the validation periods with all the 100 best parameter sets. Calibration with the 100 best parameter sets gave very similar results and the means were used in subsequent analysis.

3.3 Model Performance Criteria

The Nash–Sutcliffe efficiency (NSE) (*Nash and Sutcliffe*, 1970) was used as the statistic criterion of the model performance. The objective function used in the model calibration is the Nash and Sutcliffe efficiency of daily runoff, which is defined as:

$$NSE = 1 - \frac{\sum_{i=1}^{N}(Q_{obs,i} - Q_{sim,i})^2}{\sum_{i=1}^{N}(Q_{obs,i} - \overline{Q_{obs,i}})^2} \tag{17}$$

where $Q_{sim,i}$ and $Q_{obs,i}$ are the simulated and observed daily runoff, respectively, $\overline{Q_{obs,i}}$ is the mean observed runoff, i is the ith day, and N is the number of days sampled and it varies with individual catchment.

Following recommendations by Legates and McCabe (1999) and Hogue et al., 2006, two statistics are used to indicate the accuracy of the SIMHYD and DWBM models: the modified index of agreement (d_1) and the water balance error (WBE):

$$d_1 = 1.0 - \frac{\sum_{i=1}^{N}\left|O_{obs,i} - O_{sim,i}\right|}{\sum_{i=1}^{N}\left(\left|O_{sim,i} - \overline{O_{obs,i}}\right| + \left|O_{obs,i} - \overline{O_{obs,i}}\right|\right)} \tag{18}$$

$$WBE = \frac{\sum_{i=1}^{N}\left|Q_{sim,i} - Q_{obs,i}\right|}{\sum_{i=1}^{N}Q_{obs,i}} \times 100\% \tag{19}$$

with the symbols defined above.

3.4 Analysis of Parameter Probability Distributions under Different Calibration Periods

For each of the models, we ended up with 100 best parameter sets for each catchment and for each calibration period. From these parameters sets we calculated a probability distribution of each parameter. For a given significance level α, the chi-square test (χ^2 test) was used to test the null hypothesis that the parameter distributions obtained for a dry period and a wet period were significantly different. A *p* value greater than 0.01 indicates a rejection of the null hypothesis, which means that the parameter probability distributions for the two different calibration periods are similar.

4 Results

4.1 Comparisons of Model Calibration under Different Climatic Conditions

Results of model calibration under different climatic conditions are shown in **Figure 5** and **Table 3**. **Figure 5(a)** shows the percentage of model calibration tests that have a NSE value exceeding a given NSE value. Similarly, **Figure 5(b-c)** are corresponding plots of the modified index of agreement (d_1), the water balance error (*WBE*), respectively. It can be seen that the SIMHYD model was well calibrated under both dry and wet conditions. The average value is greater than 0.70 for NSE, 0.73 for d_1. The average water balance error is 14% and 11% for the dry and wet calibration periods. Compared with the SIMHYD model, the DWBM model showed slightly poorer results. The average value for the DWBM model is greater than 0.57 for NSE, 0.65 for d_1. The average water balance error is 22% and 17% for the dry and wet calibration periods.

The plots show that both models were better calibrated under wet periods than under dry ones, with higher values of NSE and d_1 and lower values of WBE in the wet calibration periods. For example, under the dry conditions, average NSE was 0.70 and 0.57 for the SIMHYD and the DWBM model. Under the wet conditions, average NSE was 0.76 and 0.65 respectively for the two models. In **Figure 5(a)**, a larger NSE value means a better performance, whereas in **Figure 5(c)**, a smaller percentage WBE value is better. It can be noted that all the results became worse when the calibration periods became drier, indicating a higher sensitivity of the models to dry climatic conditions. The results also indicated that the errors in the simulated runoff were increased under drier climatic conditions.

It can be seen from Table 3 that under dry and wet calibration periods, the median NSE values are, for the SIMHYD model, 0.70 and 0.77, respectively, and for the DWBM model, 0.58 and 0.66. The median d_1 values showed similar patterns under dry and wet calibration conditions. The median percentile of the WBE values are 13% and 8% for the SIMHYD model under dry and wet calibration periods respectively, and 15% and 12% for the DWBM model. All these results indicate that the two models can be calibrated satisfactorily for most of the tests, although the calibration results of the DWBM model are slightly poorer compared with those of the SIMHYD model. The average NSE values calibrated under the wet periods are higher – i.e. better – by 0.06 (SIMHYD model) and 0.08 (DWBM model) than those calibrated under dry periods. The average WBE values calibrated under wet periods are lower – again better – by 3% (SIMHYD model) and 5% (DWBM model) than those calibrated under the dry period.

[Figure 5 and Table 3 about here]

4.2 Comparisons of Model Validation using Different Calibration Periods

Validation runs were conducted for 60, 120, 60, and 120 tests for the **dry**/dry, **dry**/wet, **wet**/dry, and **wet**/wet groups, respectively. The model validation results are summarized in **Figure 6 and Table 4**. As expected, the validation results are slightly poorer than the calibration results, with the averaged NSE values in the model validation generally being 0.1 to 0.2 lower than those in the model calibration and percentage water balance error being 2 to 7% higher.

Comparing the validation results of the **dry**/dry, **dry**/wet, **wet**/dry, and **wet**/wet groups in **Figure 6**, it can be noted both the SIMHYD and DWBM models gave similar patterns. The results for the **wet**/wet are better than those of the **dry**/wet – this means that the models performed better during a wet period when they are calibrated against a wet period, compared to when they are calibrated against a dry period. These results suggest, not unexpectedly, that if a hydrological model is intended to simulate streamflow for a wet climate period then it should be calibrated on a wet segment of the historic record. They also show that hydrological models will, in general, perform better when calibrated in a wet period than when calibrated in the dry period.

Table 4 summarizes the 25th percentile, median, 75th percentile, and average values of NSE, d_1 and WBE in the validation periods. The results from the **dry**/dry test are slightly better than the results from the **wet**/dry test in terms of NSE, d_1 and WBE. The results indicate, again reasonably, that the hydrological models perform better in a dry period when calibrated in a dry period rather than in a wet period.

[Figure 6 and Table 4 about here]

4.3 Parameter Uncertainty under Climatic Nonstationarity

As described in section 3.2, assemblies of the 100 best parameter sets were selected from Monte Carlo simulation under different calibration conditions. **Table 5** shows the percentage of the catchments in which the model parameter distributions for a dry and wet period were significantly different (p<0.01). For each model, the parameters are ranked from the most sensitive to calibration conditions to least sensitive. For the SIMHYD model, the most sensitive parameters were SUB, SMSC, SQ, and CRAK, each of which significantly affected 50% or more of the catchments. The other three parameters, K, COEFF, and INSC had smaller effects, with INSC (having an effect in only 10% of catchments) being the most insensitive to choice of dry and wet calibration periods.

[Table 5 about here]

In order to further examine the effects of climatic conditions on the results, we grouped the 30 study catchments into two climatic types: 16 water-limited catchments with an index of dryness (E_p/P) greater than 1, and 14 energy-limited catchments with an index of dryness less than 1. It can be noted that all parameters performed differently in water-limited and energy-limited catchments, in particular SUB, SMSC, and CRAK.

For the DWBM model, the parameters α_1 and S_{max} exhibited different effects on runoff under the dry and wet calibration periods as 67% and 63% of the catchments showed statistically different results at the 0.01 level. At the other extreme, the parameter α_2 displayed an apparent insensitivity to the calibration periods (just 23% of catchments were affected). The parameter α_2 represents evapotranspiration efficiency and it behaves similarly to the parameter *w* of *Zhang et al*. (2001) and (2004), which was shown to be mostly correlated with vegetation cover. The parameter *d* was more sensitive to the choice of the calibration period for the water-limited catchments than for the energy-limited catchments. It is interesting to note that all the parameters behaved differently under the water-limited and energy-limited conditions, except perhaps for parameter α_2.

The above results indicate that some of the model parameters are sensitive to calibration conditions and the others are relative robust. An important question is how the sensitive parameters vary between the different calibration periods. **Figures 7** and **8** show the distributions of the optimized parameters of the two models under the dry and wet conditions in two selected catchments. The catchment 110003 has summer-dominant rainfall and catchment 401210 is winter-dominant. For the SIMHYD model, some parameters exhibited different distributions in the dry and wet calibration periods. For example, the parameter SUB tends to be more likely at a higher value in the dry periods than in the wet periods. However, the results did not reveal any systematic trends in the other parameters. For the DWBM model, the most likely value for the parameter α_1 was higher in the dry period than in the wet period for catchment 110003 and vice versa for catchment 401210 (**Figure 8**). The parameter

S_{max} showed different distributions in the dry and wet periods and these distributions vary across the catchments.

[Figures 7 and 8 about here]

5 Discussion

Streamflow of a catchment is influenced by a number of factors, most noticeably rainfall and antecedent soil moisture. During dry periods, catchments are generally characterized by small runoff events and lower runoff to rainfall ratios with higher percentage error in both rainfall and runoff. In this case, rainfall-runoff models become very sensitive to both rainfall and parameter optimization. Also, dry periods may not contain enough high flows to adequately calibrate model parameters responsible for simulating high flows (*Gan et al*., 1997). Apart from rainfall amount, spatial variability of rainfall can also affect runoff. *Smith et al.* (2004) showed that improved runoff simulations can be obtained from distributed versus lumped rainfall-runoff models in catchments with considerable rainfall variability. Spatial variability of rainfall was also found to be the dominant control on runoff production (*Segond et al*., 2007). In this study, spatially averaged rainfall was used in both model calibration and validation. This is likely to affect the model results and it is expected that the rainfall variability effect will be greater in dry periods than in wet periods.

It has been widely acknowledged that spatial variability of antecedent soil moisture conditions plays an important role in runoff generation (*Grayson and Blöschl*, 2000). *Minet et al*. (2011) investigated the effect of spatial soil moisture variability on runoff simulations using a distributed hydrologic model and showed that model results are sensitive to soil moisture spatial variability, especially in dry conditions. At catchment scales, soil moisture exhibit larger heterogeneity under dry conditions than wet conditions and this means errors associated with dry period runoff simulations are likely to be greater as runoff generation exhibits non-linear threshold behaviour. In this study, the differences in average annual rainfall between the wet and dry periods ranged from 10 to 47% of the long-term average rainfall and are comparable with percentage change in man annual rainfall for 2030 relative to 1990 from 15 GCMs for the Murray Darling Basin in Australia (*Chiew* et al., 2008).

The results of this study indicate that calibration periods can cause significant shifts in model parameter distributions. Some model parameters are relatively sensitive to the choice of calibration periods, while the others are fairly insensitive. As well as the impact of calibration periods on parameter distributions, whether catchments are water-limited or energy-limited also needs to be taken into consideration. For the SIMHYD model, the most sensitive parameters are SUB, SMSC, and CRAK. The parameter SUB is used to estimate interflow and it can be an important parameter in some catchments (*Chiew and McMahon*, 1994). However, it is difficult to estimate this parameter *a priori* as it is poorly correlated with any catchment characteristics (*Chiew and McMahon*, 1994). The soil moisture store capacity (SMSC) affects many processes such as infiltration and evapotranspiration and it is determined by soil properties and vegetation characteristics (e.g. rooting depth). Accurate estimation of this parameter is essential to achieving satisfactory model performance. The parameter CRAK determines groundwater recharge/baseflow and is highly correlated with soil types. For the DWBM model, the most sensitive parameters are α_1 and S_{max}, and *d*, representing catchment rainfall retention efficiency, maximum storage capacity, and the recession constant, respectively (*Zhang* et al. 2008). In a way, these parameters are similar to those sensitive parameters in SIMHYD in terms of their functional controls on water balance components. *Merz et al* (2011) applied a semi-distributed conceptual rainfall-runoff model to 273 catchments in Austria and showed that the parameters of the soil moisture accounting schemes exhibited strong dependence on calibration conditions, consistent with the results of the current study. This also suggests that parameters related to soil moisture accounting are likely to change with calibration conditions. The fact that these parameters are sensitive to the choice of calibration period (i.e. dry *vs* wet) also indicates that large uncertainty may be associated with these parameters and cares need to be exercised when transferring the parameters to conditions different from the calibration.

These findings have major implications for studies of climate change impact on streamflow. When a hydrological model calibrated for a given climatic condition (e.g. wet periods) is used to simulate runoff of different climatic conditions (e.g. dry periods), transfer of some model parameters (i.e. sensitive parameters) may result in large errors in simulated runoff. One may argue that the sensitive model parameters should be updated by functionally relating them with climatic variables such as rainfall (*Merz* et al., 2011). This could potentially reduce uncertainty and lead to more accurate predictions. However, some of the parameters are poorly related to catchment characteristics (e.g. rainfall) and the problem is further complicated by the fact that not every parameter is well identified and different parameter values can result in equal model performance, i.e. equifinality (*Beven*, 1993). It has also been recognized that model calibration tends to compensate model structural errors (*Merz* et al., 2011, *Wagner* et al., 2003), making it difficult to understand how model parameters vary with calibration conditions (*Wagener* et al., 2010).

The differential split-sample test can be considered as the first step in addressing the issue of parameter transferability under non-stationary conditions. Monte Carlo simulation provided an effective and pragmatic approach to exploring uncertainty in hydrological model parameters. The performance of rainfall-runoff models is related to catchment characteristics such as climate, topography, soil, vegetation, catchment shape, geology, drainage network. In such a complex situation, it is hard to pinpoint the source of parameter uncertainty, but the results of this study showed that calibration periods and catchment climatic conditions are both important factors that can result in uncertainty in model performance.

Credibility of a hydrological model has traditionally been tested using streamflow data from a validation period that is similar to calibration period. The assumption is that the model will be used under conditions similar to those of the calibration. However, when dealing with impact of climate change on streamflow, the assumption is not generally valid and the model needs to be tested under conditions different from those of the calibration. For this purpose, the two hydrological models were evaluated using differential split-sample test (Klemes, 1986). When using a dry period for calibration and a wet period for validation, the models produced more accurate estimates of streamflow (i.e. higher NSE and lower bias) compared with estimates produced using a wet period for calibration and a dry period for validation (see Table 4). Similar results have been reported by Vaze et al. (2010) and the finding can be partly explained by the fact that hydrological models generally perform better in wet periods than in dry periods (Vaze et al., 2010; Gallart et al., 2007, Perrin et al. 2007; Lidén and Harlin, 2000, Gan et al., 1997; Hughes, 1997).

A closer examination of model errors reveals that when the model parameters, calibrated on a dry period, were used to simulate runoff during a wet period, the mean of the simulated runoff was usually underestimated; conversely, when model parameters, calibrated on a wet period, were used to simulate dry period runoff, the mean simulated runoff was overestimated, consistent with the findings of *Gan et al*. (1997). *Vaze et al.* (2010) also showed that when hydrological models were calibrated using long period of record and tested for sub-periods with above long-term average rainfall, the model performed well. However, performance of the models starts to deteriorate when tested for sub-periods with below long-term average rainfall.

Traditionally, one would use a sufficiently long period of records for model calibration to ensure proper presentation of climate/streamflow variability and to achieve stable model parameters. If the model is to be used under stationary conditions, it is generally recommended that the whole record should be divided into two segments, one for calibration and the other for validation. However, if a model is to be used under non-stationary conditions, its parameters should be transferable. In other words, the parameters should be estimated so that the model gives accurate estimates of streamflow outside the climatic conditions encountered in calibration period. In this case, one should identify two periods with different climatic conditions (e.g. a dry period and wet period) from the whole record and apply the so-called differential split-sample test (*Klemes*, 1986). One another approach to this problem is to examine how other catchments behave under these different climatic conditions, i.e. trading space for time (*Singh* et al., 2011).

6 Conclusions

Potentially large uncertainties arise when predicting hydrological responses to future climate change – due to factors such as the choice of emission scenario, GCM, downscaling technique, hydrological model, optimization technique, and the way the model is calibrated. It is therefore important to develop reliable ways to calibrate hydrological models under present-day conditions. This study compared hydrological model performances under nonstationarity by using the differential split-sample test and two conceptual rainfall–runoff models, DWBM and SIMHYD, applied to 30 catchments in Australia. Monte Carlo simulation was used to explore parameter stability and transferability in the context of historic climate variability.

Apart from quality of the input data (e.g. rainfall) and model structure, performance of a hydrological model is also dependent on how it is calibrated. If a hydrological model is intended to simulate runoff for a wet climate scenario then it should be calibrated on a wet segment of the historic record. Conversely, if it is intended to simulate runoff for a dry climate scenario then it should be calibrated on a dry segment of the historic record. We also found that when using a dry period for calibration and a wet period for validation, the models produced more accurate estimates of streamflow compared with estimates produced using a wet period for calibration and a dry period for validation. In other words, transferring model parameter values obtained from dry periods to wet periods will result in smaller errors in streamflow estimation than transferring model parameter values obtained from wet periods to dry periods. The soil related model parameters are more sensitive to the choice of calibration period than other parameters and large uncertainty may be introduced when transferring the soil related parameters to conditions different from the calibration. Our research has implications for hydrological modellers looking to estimate future runoff and we hope this study will stimulate further research into the selection of calibration data.

Acknowledgement

This study was supported by the National Basic Research Program of China (2010CB951102), the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (51021066), Open Foundation of State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering (2011490511), the Foundation of China Institute of Water Resources and Hydropower Research (1232) and the Regional Water Theme in the Water for a Healthy Country Flagship. We thank Andrew Bell, Enli Wang and anonymous reviewers for their helpful comments on a draft of the paper.

References

Andreassian, V., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., Loumagne, C., Oudin, L., Mathevet, T., Romas, M.-H., Valery, A., 2009. Crash tests for a standardized evaluation of hydrological models. Hydrol. Earth Syst. Sci. 13, 1757–1764.

Beven K J. 1993. Prophecy, reality and uncertainty in distributed hydrological modelling, *Advances in Water Resources*, **16**: 41-51.

Boorman D B, Sefton C E M. 1997. Recognising the uncertainty in the quantification of the effects of climate change on hydrological response, *Climatic Change*, **35**: 415-434.

Boyer, C, Chaumont, D, Chartier, I, and Roy, A.G. 2010. Impact of climate change on the hydrology of St. Lawrence tributaries, *Journal of Hydrology*, **384**, 65-83.

Budyko, M.I., 1958. The Heat Balance of the Earth's Surface. US Department of Commerce, Washington, DC.

Calder, I.R., 1998. Water use by forests, limits and controls, *Tree Physiol*, **18**: 625-631.

Chiew F H S, Whetton P H, McMahon T A, Pittock A B. 1995. Simulation of the impacts of climate change on runoff and soil moisture in Australian catchments, *Journal of Hydrology*, **167**: 121-147.

Chiew F H S, T A McMahon. 2002. Global ENSO-streamflow teleconnection, streamflow forecasting and interannual variability, *Hydrological Sciences Journal – Journal Des Sciences Hydrologiques*, **47**: 505-522.

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

Chiew F H S, Teng J, Kirono D, Frost A J, Bathols J M, Vaze J, Viney N R, Young W J, Hennessy K J and Cai W J. 2008. Climate data for hydrologic scenario modelling across the Murray-Darling Basin. A report to the Australian Government from the CSIRO Murray-Darling Basin Sustainable Yields Project. Water for a Healthy Country Flagship. CSIRO. 42 pp.

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

Fu, B P , 1981. On the calculation of the evaporation from land surface, Sci. Atmos. Sin., 23-31 (in Chinese).

Gallart, F., Latron, J., Liorens, P., and Beven, K. 2007. Using internal catchment information to reduce the uncertainty of discharge and baseflow predictions. *Advances in Water Resources*, **30**, 808-823.

Gan T Y, Dlamini E M, Biftu G F. 1997. Effects of model complexity and structure, data quality, and objective functions on hydrologic modelling, *Journal of Hydrology*, **192**: 81-103.

Grayson, R B and Blöschl G. 2000. Spatial Patterns in Catchment Hydrology: Observations and Modelling. Cambridge University Press. 404pp.

Henriksen H J, Troldborg L, Nyegaard P, Sonnenborg T O, Refsgaard J C, Madsen B. 2003. Methodology for construction, calibration and validation of a national hydrological model for Denmark, *Journal of Hydrology*, **280**: 52-71.

Hogue S T, Gupta H, Sorooshian S. 2006. A ‘user-friendly’ approach to parameter estimation in hydrologic models, *Journal of Hydrology*, **320**: 202-217.

Hughes, D.A. 1997. Southern African ‘FRIEND’-the Application of Rainfall-runoff Models in the SADC Region, Report to the Water Research Commission by the Institute for Water Research, Rhodes University, WRC Report No. 235/1/97, Pretoria, South Africa. 69 pp.

IPCC: Climate Change 2007: The Physical Basis, Contributions of Working Group 1 to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M. and Miller, H.L. (eds.). Cambridge University Press, Cambridge,United Kingdom and New York, USA, 996 pp, 2007.

Jeffrey S J, Carter J O, Moodie K B, Beswick A R. 2001. Using spatial interpolation to construct a comprehensive archive of Australian climate data, *Environmental Modelling & Software*, **16**: 309-330.

Klemes V. 1986. Operational testing of hydrological simulation models, *Hydrological Sciences Journal*, **31**: 13-24.

Lidén, R. Harlin, J. 2000. Analysis of conceptual rainfall-runoff modeling performance in different climates. *Journal of Hydrology*, **238**, 231-247.

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

Milly P C D, Betancourt J, Falkenmark M, Hirsch R M, Kundzewicz Z W, Lettenmaier D P, Stouffer R J. 2008. Stationarity is dead: whither water management? *Science*, **319**: 573-574.

Minet J, Laloy E, Lambot S, and Vanclooster M. 2011. Effect of high-resolution spatial soil moisture variability on simulated runoff response using a distributed hydrologic model, *Hydrology and Earth System Sciences*, **15**: 1323-1338.

Minville M, Brissette F, Leconte R. 2008. Uncertainty of the impact of climate change on the hydrology of a nordic watershed, *Journal of Hydrology*, **358**: 70-83.

Monomoy G, O’Connor, K.M. 2007. Comparative assessment of six automatic optimization techniques for calibration of a conceptual rainfall–runoff model, *Hydrological Sciences Journal – Journal Des Sciences Hydrologiques*, **52**(3): 432-449.

Nash J E, J V Sutcliffe. 1970. River forcasting using conceptual models, 1. A discussion of principles. *Journal of Hydrology*, **10**: 280-290.

Niel H, Paturel J E, Servat E. 2003. Study of parameter stability of a lumped hydrologic model in a context of climatic variability, *Journal of Hydrology*, **278**: 213-230.

Peel M C, Chiew F H S, Western A W, McMahon T A. 2000. Extension of unimpaired monthly stream flow data and regionalization of parameter values to estimate stream flow in ungauged catchments. Report to National Land and Water Resources Audit, Cent. For Environ. Appl. Hydrol., Univ. of Melbourne, Parkville, Vic., Australia.

Perrin C., Oudin L., Anderassian V., Rojas-serna C., Michel C., and Mathevet T. 2007. Impact of limited streamflow data on the efficiency and the parameters of rainfall-runoff models. *Hydrological Sciences Journal*, **52**:1, 131-151.

Rind D, Rosenzweig C, Goldberg R. 1992. Modelling the hydrological cycle in assessments of climate change, *Nature*, **358**: 119-120.

Priestley C H B, Taylor R J. 1972. On the assessment of the surface heat flux and evaporation using large-scale parameters, *Monthly Weather Review*, **100**: 81-92.

Sankarasubramanian, A., Vogel, R.V., 2003. Hydroclimatology of the continental United States, Geophysical Research Letter, **30**(7), 1363. doi:10.1029/2002GL015937.

Segond, M. L., Wheater, H. S., Onof, C., 2007. The significance of spatial rainfall representation for flood runoff estimation: A numerical evaluation based on the Lee catchment, UK, *Journal of Hydrolog*y, **347**(1-2): 116-131.

Siriwardena L, Finlayson B L, McMahon T A. 2006. The impact of land use change on catchment hydrology in large catchments: The Comet River, Central Queensland, Australia, *Journal of Hydrology*, **326**: 199-214.

Smith M B , Koren V I, Zhang Z Y, Reed S M, Pan J J and Moreda F. 2004. Runoff response to spatial variability in precipitation: an analysis of observed data, *Journal of Hydrology*, **298**: 267-286.

Raupach M R. 2000. Equilibrium evaporation and the convective boundary layer, *Boundary-Layer Meteorology*, 96: 107-141.

Raupach M R, J M Kirby, D J Barrett, P R Briggs, H Lu and L Zhang. 2001. Balances of water, carbon, nitrogen and phosphorus in Australian landscapes: 2. Model formulation and testing, Tech. Rep. 41/01, CSRIO Land and Water, Canberra, ACT, Australia.

Refsgaard J C, Storm B. 1996. Construction, calibration and validation of hydrological models. In Abbott, M. B., and Refsgaard, J. C. (ed.) *Distributed Hydrological Modelling*, Kluwer Academic Publishers, Netherlands, 50 pp.

Seibert, J., 2003. Reliability of model predictions outside calibration conditions, Nordic Hydrology, 34: 477-492.

Singh, R., Wageber, T., Vab Werkhoveb, K., Mann, M., and Crane, R. (2011). A trading-space-for time approach to probabilistic continuous streamflow predictions in a changing climate. Hydrol. Earth Syst. Sci. Discuss., 8, 6385-6417.

Uhlenbrook S, Seibert J, Leibundgut C, Rodhe A. 1999. Prediction uncertainty of conceptual rainfall–runoff models caused by problems in identifying model parameters and structure, *Hydrological Sciences – Journal des Sciences Hydrologiques*, **44**(5): 779-797.

Vaze J, Post D A, Chiew F H S, Perraud J M, Viney N R, Teng J. 2010. Climate non-stationarity – Validity of calibrated rainfall–runoff models for use in climate change studies, *Journal of Hydrology*, **394**: 447-457.

Viney N, Vaze J, Chiew F, Perraud J. 2008. Regionalisation of runoff generation across the Murray–Darling Basin using an ensemble of two rainfall–runoff models. Paper presented at Water Down Under 2008, April 2008, Adelaide: Engineers Australia.

Wagener, T., N. McIntyre, M. J. Lees, H. S. Wheater, and H. V. Gupta. 2003. Towards reduced uncertainty in conceptual rainfall-runoff modeling: Dynamic identifiability analysis, Hydrol. Processes, **17**, 455-476.

Wagener, T., M. Sivapalan, P. A. Troch, B. L. McGlynn, C. J. Harman, H. V. Gupta, P. Kumar, P. S. C. Rao, N. B. Basu, and J. S. Wilson. 2010. The future of hydrology: An evolving science for a changing world, Water Resour. Res., **46**, W05301, doi:10.1029/2009WR008906.

Widen-Nilsson E, Gong L, Halldin S, Xu C Y. 2009. Model performance and parameter behavior for varying time aggregations and evaluation criteria in the WASMOD-M global water balance model. *Water Resources Research*, **45**: W05418, doi:10.1029/2007WR006695.

Wilby R L. 2005. Uncertainty in water resource model parameters used for climate change impact assessment, *Hydrological Processes*, **19**: 3201-3219.

Xu C Y. 1999. Operational testing of a water balance model for predicting climate change impacts, *Agricultural and Forest Meteorology*, **98**(9): 295-304.

Zhang L, Dawes W R, Walker G R. 2001. Response of mean annual evapotranspiration to vegetation changes at catchment scale. *Water Resources Research*, **37**: 701-708.

Zhang L, Hickel K, Dawes W R, Chiew F H S, Western A W, Briggs P R. 2004. A rational function approach for estimating mean annual evapotranspiration, *Water Resources Research*, **40**: W02502, doi:10.1029/2003WR002710.

Zhang L, Potter N, Zhang Y Q, Hickel K, Shao Q X. 2008. Water balance modeling over variable time scales based on the Budyko framework: model development and testing, *Journal of Hydrology*, **360**: 117-131.

Zhang Y Q, Chiew F H S, Zhang L, Leuning R, Cleugh H A. 2008. Estimating catchment evaporation and runoff using MODIS leaf area index and the Penman–Monteith equation, *Water Resources Research*, **44**: W10420, doi:10.1029/2007WR006563.

Zhang Y Q, Chiew F H S, Zhang L, Li H X. 2009. Use of remotely sensed actual evapotranspiration to improve rainfall–runoff modelling in southeast Australia, *Journal of Hydrometeorology*, **10**, 969-980. doi: 10.1175/2009JHM1061.1.

Table and Figure Captions

Table 1 Ranges of parameter values in DWBM (/ indicates dimensionless).

Table 2 Ranges of parameters in the SIMHYD model (/ indicates dimensionless).

Table 3 Summary results of the model calibration under different climatic conditions (*i.e.* dry and wet periods).

Table 4 Summary results of the model validation when calibrated under different climatic conditions.

Table 5 Percent of the catchments in which the model parameter distributions for a dry and wet calibration period were significantly different (p<0.01) under Monte Carlo simulation. Also shown are the results for water-limited ($E_p/P>1$) and energy-limited ($E_p/P<1$) catchments. For each model, the parameters are ranked from the most sensitive to calibration conditions to least sensitive.

Figure 1 Structure of the lumped dynamic water balance model (DWBM).

Figure 2 Structure of the lumped daily rainfall–runoff model (SIMHYD).

Figure 3 Location map of the 30 catchments used for this study.

Figure 4 Annual historical precipitation of the Corang River catchment showing estimation of 2 wet periods (A) and 2 dry periods (B) to represent different calibration conditions.

Figure 5 (a) Percentage of model calibration tests with a NSE value greater than or equal to a given NSE value. Similarly, **Figure 5 (b-c)** are corresponding plots of the modified index of agreement (d_1), the water balance error (*WBE*), respectively.

Figures 6 (a) and (d) Percentage of model validation tests with a NSE value greater than or equal to a given NSE value. Similarly, **Figures 6 (b) and (e)**, **Figures 6 (c) and (f)** are corresponding plots of the modified index of agreement (d_1), the water balance error (*WBE*), respectively.

Figure 7 Probability density functions for 7 parameters of the SIMHYD model under dry and wet calibration periods in catchments 110003 and 4021210.

Figure 8 Probability density functions for 4 parameters of the DWBM model under dry and wet calibration periods in catchments 110003 and 4021210.

Tables and Figures

Table 1 Ranges of parameter values in DWBM (/ indicates dimensionless).

Parameter	Units	Description	Lower bound	Upper bound
α_1	/	retention efficiency	1	5
α_2	/	evapotranspiration efficiency	1	5
S_{max}	mm	soil water storage capacity	5	500
d	/	baseflow linear regression	0.01	1

Table 2 Ranges of parameter values in the SIMHYD model (/ indicates dimensionless).

Parameter	Units	Description	Lower bound	Upper bound
INSC	mm	interception store capacity	0.5	5.0
COEFF	mm	maximum infiltration loss	50	400
SQ	/	infiltration loss exponent	0	6.0
SMSC	mm	soil moisture store capacity	50	500
SUB	/	constant of proportionality in interflow equation	0	1
CRAK	/	constant of proportionality in groundwater recharge equation	0	1
K	/	baseflow linear regression parameter	0.003	0.3

Table 3 Summary results of the model calibration under different climatic conditions (*i.e.* dry and wet periods).

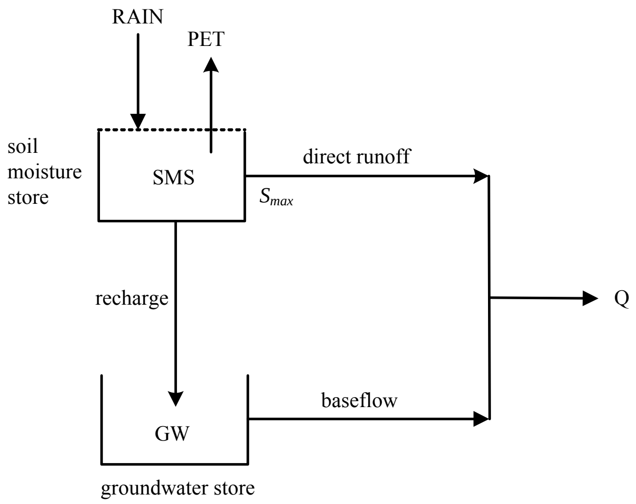
Indicator	SIMHYD calibrated on dry period	SIMHYD calibrated on wet period	DWBM calibrated on dry period	DWBM calibrated on wet period
25th NSE	0.84	0.85	0.71	0.77
Median NSE	0.70	0.77	0.58	0.66
75th NSE	0.61	0.68	0.43	0.54
Average NSE	0.70	0.76	0.57	0.65
25th d_1	0.77	0.79	0.71	0.75
Median d_1	0.72	0.76	0.67	0.71
75th d_1	0.70	0.74	0.61	0.68
Average d_1	0.73	0.76	0.65	0.71
25th WBE	22	16	25	24
Median WBE	13	8	15	12
75th WBE	6	4	9	5
Average WBE	14	11	22	17

Table 4 Summary results of the model validation when calibrated under different climatic conditions.

Model	Indicator	dry/dry	dry/wet	wet/dry	wet/wet
SIMHYD	25th NSE	0.72	0.74	0.68	0.77
	Median NSE	0.55	0.64	0.51	0.69
	75th NSE	0.42	0.44	0.41	0.55
	Average NSE	0.57	0.61	0.54	0.66
	25th d_1	0.74	0.78	0.74	0.78
	Median d_1	0.71	0.74	0.70	0.75
	75th d_1	0.66	0.70	0.63	0.72
	Average d_1	0.69	0.73	0.68	0.74
	25th WBE	34	30	39	23
	Median WBE	20	19	28	13
	75th WBE	14	8	16	7
	Average WBE	24	21	29	17
DWBM	25th NSE	0.56	0.65	0.51	0.72
	Median NSE	0.46	0.48	0.45	0.61
	75th NSE	0.34	0.35	0.30	0.42
	Average NSE	0.48	0.52	0.45	0.59
	25th d_1	0.69	0.73	0.68	0.74
	Median d_1	0.65	0.69	0.63	0.70
	75th d_1	0.58	0.64	0.56	0.66
	Average d_1	0.62	0.68	0.61	0.69
	25th WBE	35	29	53	25
	Median WBE	22	20	33	18
	75th WBE	15	12	18	11
	Average WBE	27	23	36	19

Table 5 Percent of the catchments in which the model parameter distributions for a dry and wet calibration period were significantly different (p<0.01) under Monte Carlo simulation. Also shown are the results for water-limited ($E_p/P>1$) and energy-limited ($E_p/P<1$) catchments. For each model, the parameters are ranked from the most sensitive to calibration conditions to least sensitive.

Model	Parameter	Percent of catchments	Percent of water-limited catchments	Percent of energy-limited catchments
SIMHYD	SUB	63	81	43
	SMSC	60	75	43
	SQ	53	56	50
	CRAK	50	63	36
	K	37	31	43
	COEFF	33	38	29
	INSC	10	13	7
DWBM	α_1	67	81	50
	S_{max}	63	75	50
	d	47	63	29
	α_2	23	25	21



Model parameters and description

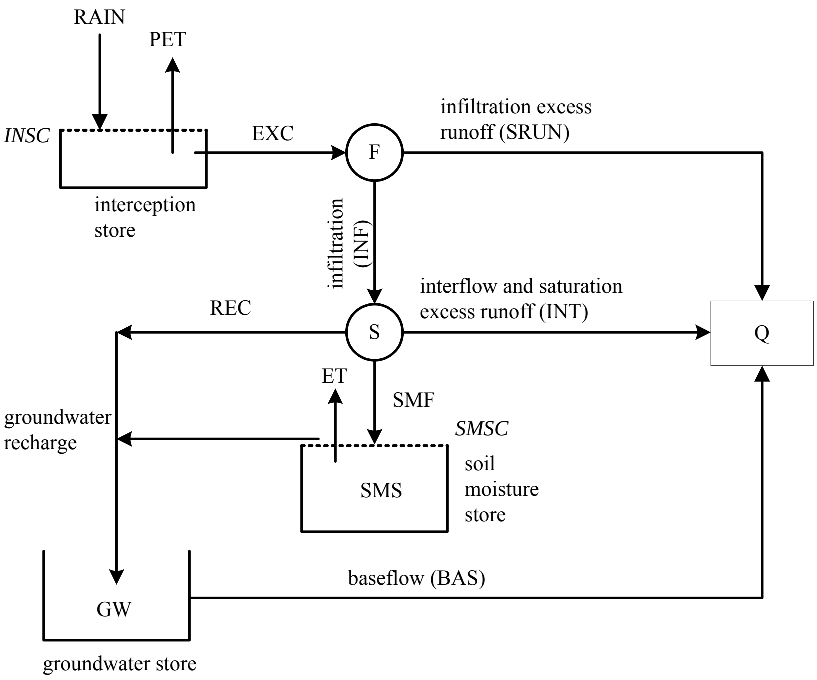
α_1 retention efficiency

α_2 evapotranspiration efficiency

S_{max} soil water storage capacity (mm)

d baseflow linear regression

Figure 1 Structure of the lumped dynamic water balance model (DWBM).



PET = areal potential evapotranspiration (input data)
EXC = RAIN - *INSC*, EXC > 0
INF = lesser of { COEFF exp (-*SQ*×SMS/S*MSC*) , EXC }
SRUN = EXC - INF
INT = *SUB* × SMS/*SMSC* × INF
REC = *CRAK* × SMS/*SMSC* × (INF - INT)
SMF = INF - INT - REC
ET = lesser of { 10 × SMS/*SMSC* , PET }
BAS = *K* × GW

Model parameters and description
INSC interception store capacity (mm)
COEFF maximum infiltration loss (mm)
SQ infiltration loss exponent
SMSC soil moisture store capacity (mm)
SUB constant of proportionality in interflow equation
CRAK constant of proportionality in groundwater recharge equation
K baseflow linear recession parameter

Figure 2 Structure of the lumped daily rainfall–runoff model SIMHYD.

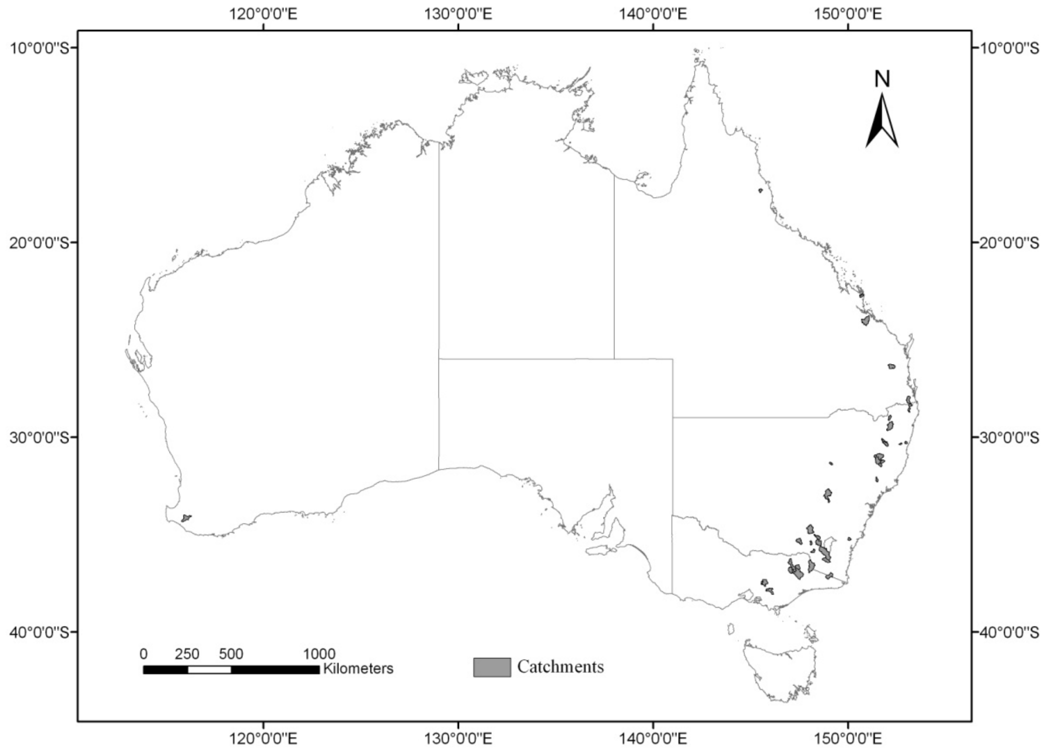


Figure 3 Location map of the 30 catchments used for this study.

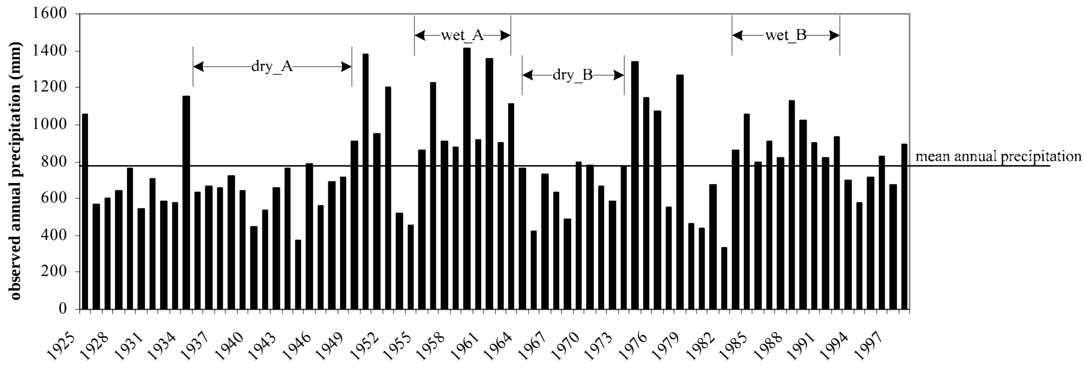


Figure 4 Annual historical precipitation of the Corang River catchment showing estimation of 2 wet periods (A) and 2 dry periods (B) to represent different calibration conditions.

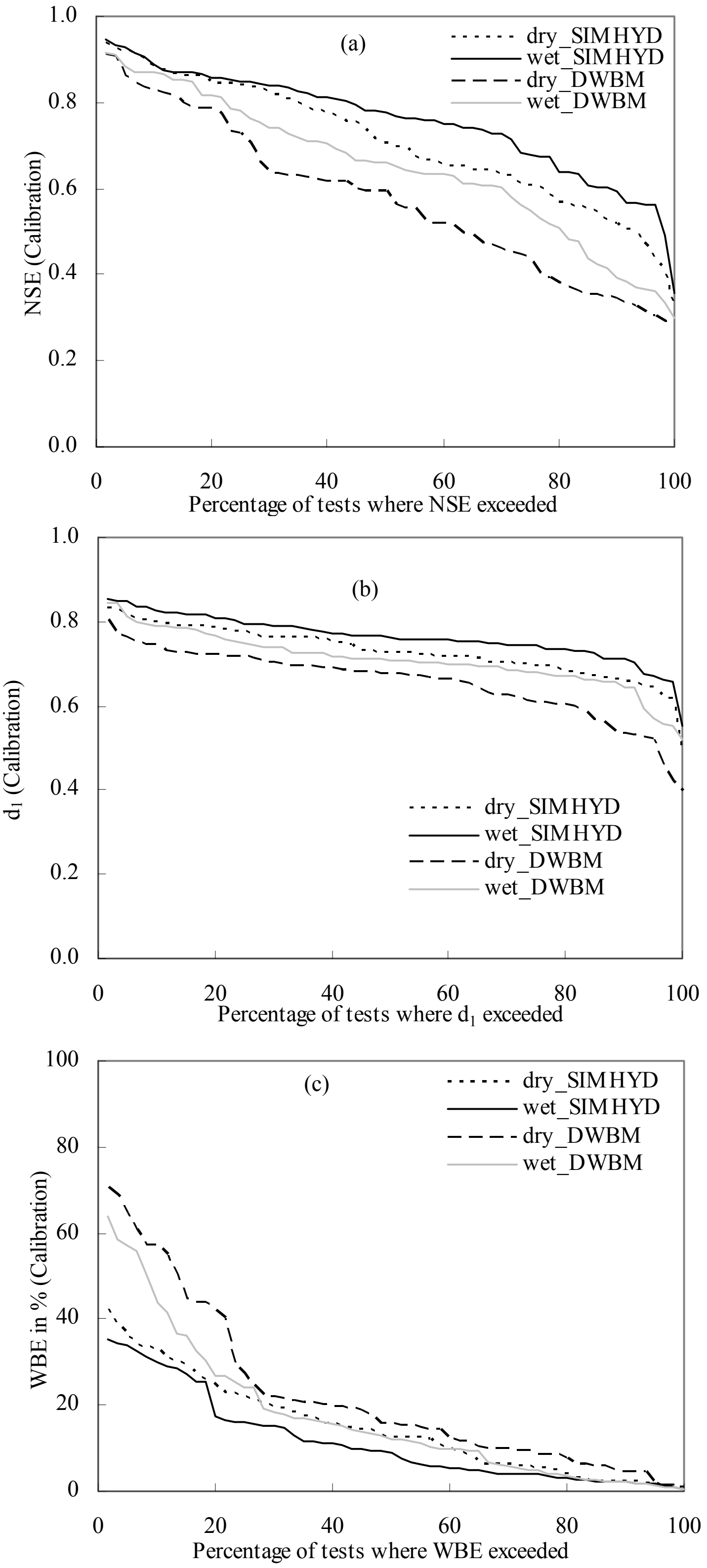
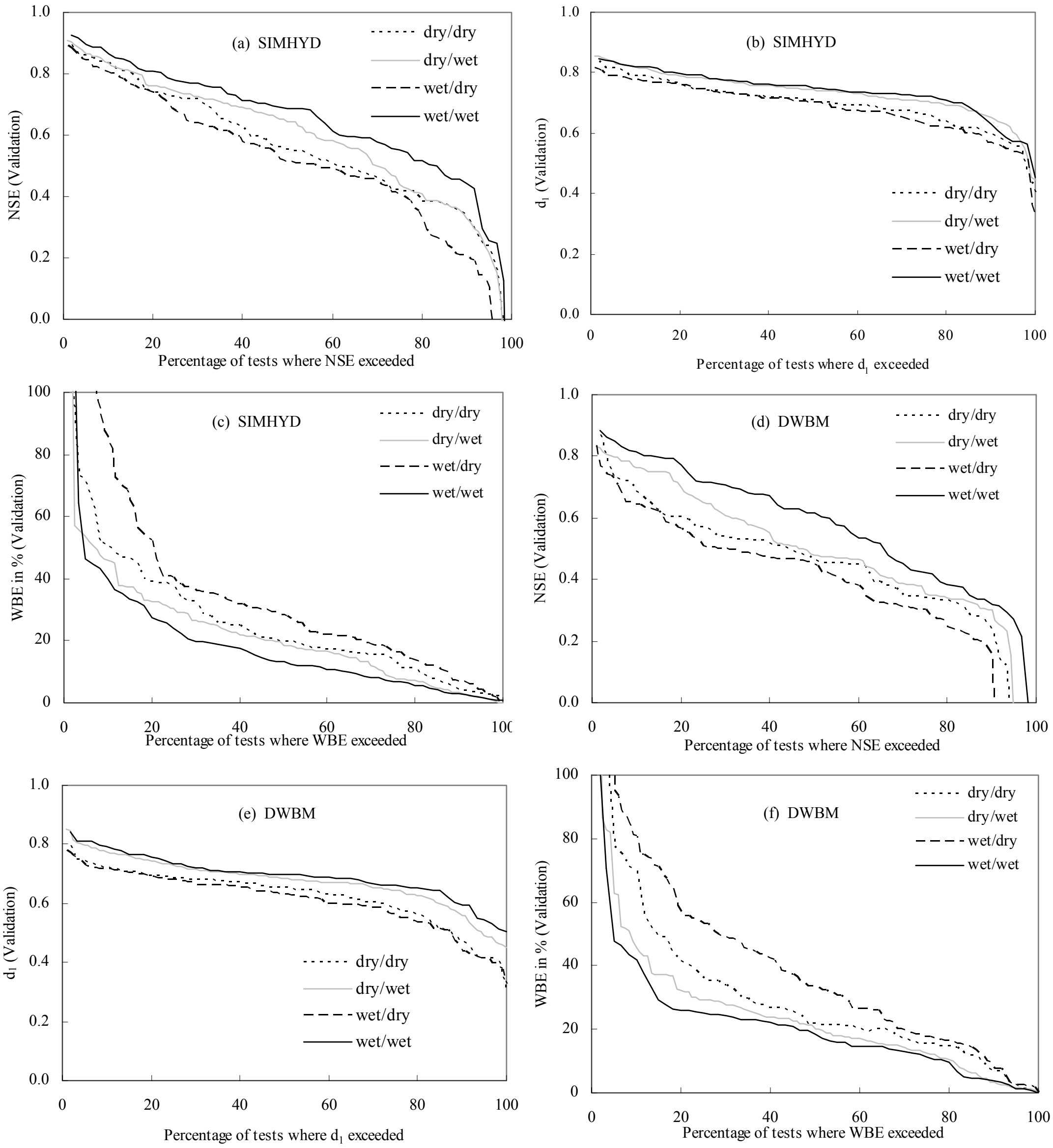


Figure 5 (a) Percentage of model calibration tests with a NSE value greater than or equal to a given NSE value. Similarly, **Figure 5 (b-c)** are corresponding plots of the modified index of agreement (*d_1*) and the water balance error (*WBE*), respectively.



Figures 6 (a) and (d) Percentage of model validation tests with a NSE value greater than or equal to a given NSE value. Similarly, **Figures 6 (b) and (e)**, **Figures 6 (c) and (f)** are corresponding plots of the modified index of agreement (*d_1*), the water balance error (*WBE*), respectively.

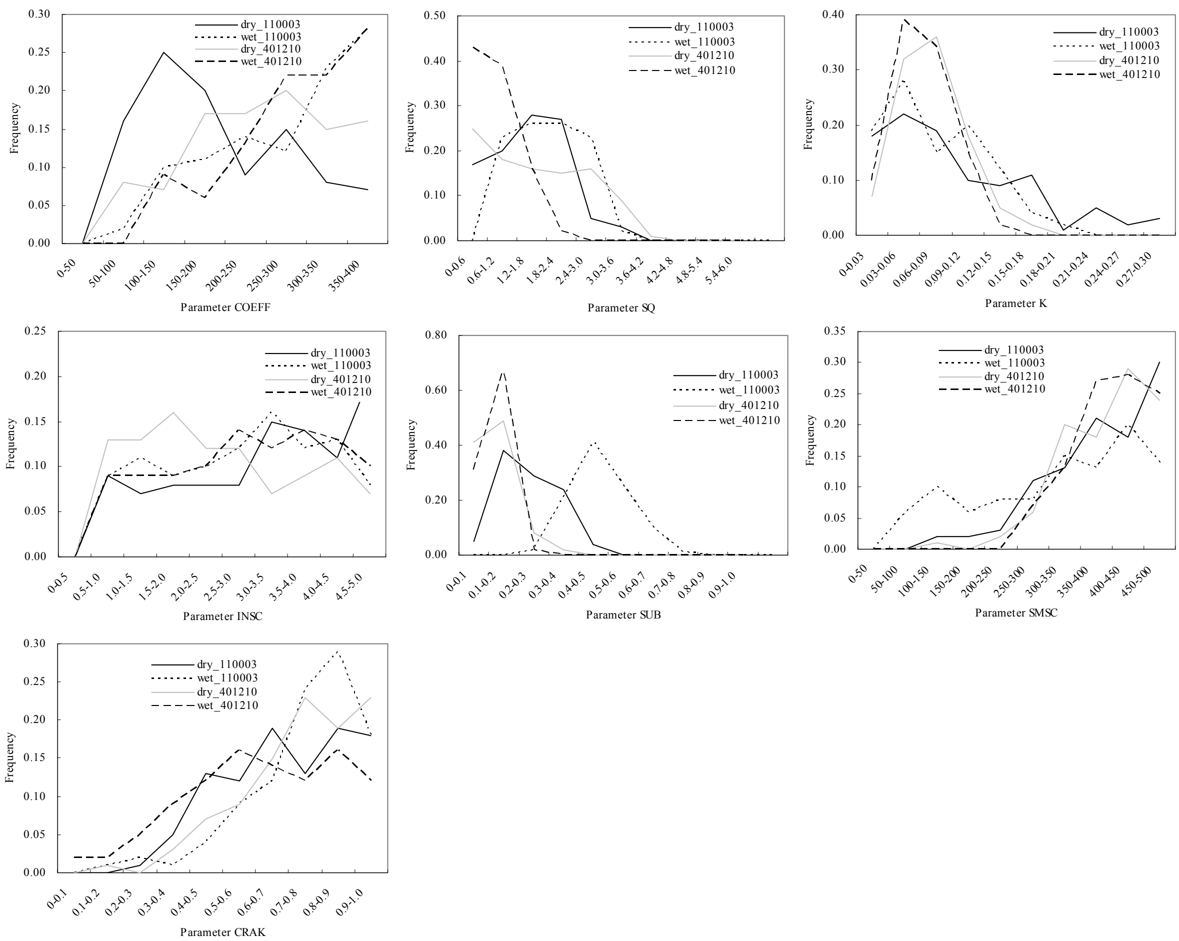


Figure 7 Probability density functions for 7 parameters of the SIMHYD model under dry and wet calibration periods in catchments 110003 and 4021210.

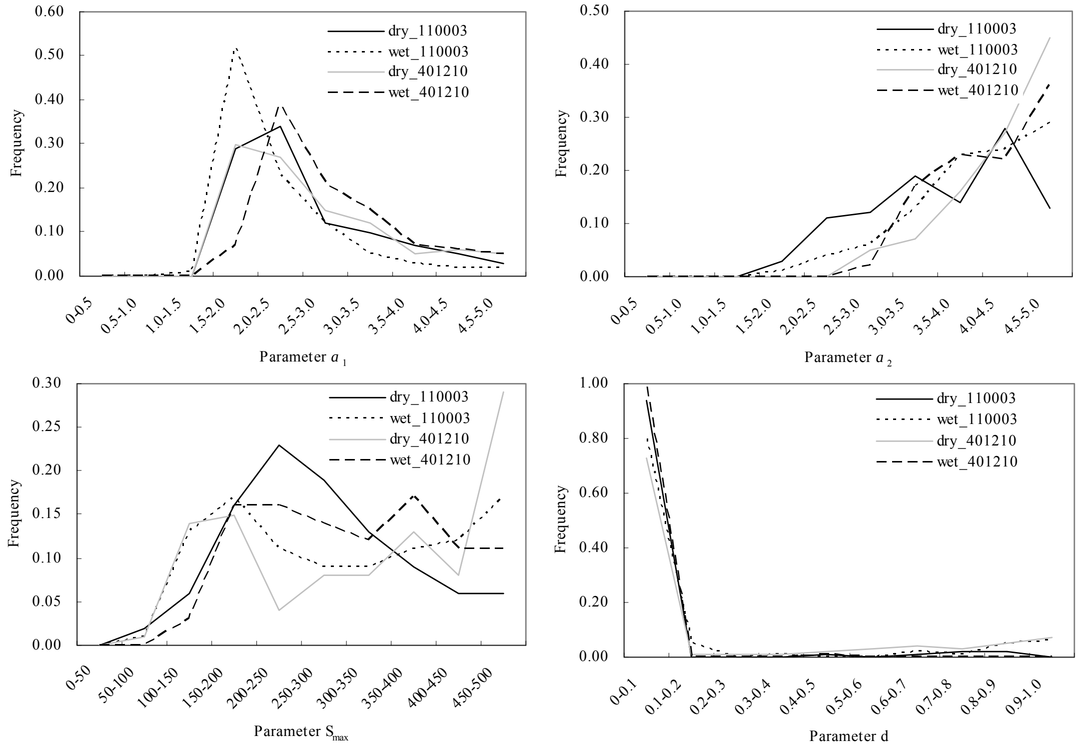


Figure 8 Probability density functions for 4 parameters of the DWBM model under dry and wet calibration periods in catchments 110003 and 4021210.