Evaluation of baseflow modelling structure in monthly water balance models using 443 Australian catchments

Shujie Cheng [a, b, c], Lei Cheng [a, b, c, *], Pan Liu [a, b, c], Lu Zhang [d], Chongyu Xu [a, e], Lihua Xiong [a, b, c], Jun Xia [a, b, c]

[a] State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan 430072, China

[b] Hubei Provincial Collaborative Innovation Center for Water Resources Security, Wuhan 430072, China

[c] Hubei Provincial Key Lab of Water System Science for Sponge City Construction, Wuhan University, Wuhan, Hubei, China

[d] CSIRO Land and Water, Black Mountain, Canberra, ACT 2601, Australia

[e] Department of Geosciences, University of Oslo, PO Box 1047 Blindern, 0316 Oslo, Norway

* Correspondence: lei.cheng@whu.edu.cn

Abstract: It is critical for monthly water balance models (MWBMs) to achieve realistic hydrological modelling of total flow and its components (*i.e.* quick flow and baseflow) in practical application. Various methods have been developed to improve the performances of the three flow components by focusing on calibration procedures. However, the understanding of runoff partitioning structure in MWBMs for better performances is still very limited, especially whether the storage-discharge relationship is linear or nonlinear at monthly time scale. In this study, model structures for baseflow simulation in 5 widely used MWBMs are reviewed and modified from a linear storage-discharge relationship to a nonlinear exponential storage-discharge relationship to achieve realistic baseflow simulation in 443 catchments from Australia with diverse hydro-climatic conditions. The performances of original and modified models are evaluated and compared through four assessment criteria including Nash-Sutcliffe efficiency (NSE), logarithmic form of NSE (NSE(log)), Pearson correlation coefficient (r) and Bias (B). Basically, the original models with linear storage-discharge relationship perform satisfactorily in simulating total streamflow and quick flow, but degrade remarkably for simulating baseflow with an underestimation of −60±36% in all study catchments. The modified MWBMs with nonlinear storage-discharge relationship significantly outperform the original ones for simulating both total streamflow and baseflow. The assessment criteria NSE, NSE(log), r and B of total streamflow improve in 82±4.0% (mean ± 1 standard deviation of 5 MWBMs), 72±4.7%, 76±4.5% and 51±2.4% study catchments, respectively. The NSE(log) and r of baseflow simulated using the modified MWBMs have improved in 68±4.6% and 83±4.1% catchments with median improvement of 0.17±0.03 and 0.14±0.03, respectively. It suggests that the exponential nonlinear storage-discharge relationship is more capable for MWBMs to capture storage-discharge dynamics than the linear one at monthly time scale. This study highlights that, at monthly time scale, the nonlinearity in catchment storage-discharge relationship is a very important factor for MWBMs performance and more studies are required to reveal catchment monthly runoff generation mechanisms.

Keywords: monthly water balance model; baseflow mechanisms; runoff partitioning structure; storage-discharge relationship

1 Introduction

Monthly water balance models (MWBMs) are important tools for effective water resource management as they have low input requirement, simple model structure and easy to calibrate (Nasseri et al., 2014; Dakhlaoui et al., 2017). Good performance and robustness of MWBMs are fundamental for water resources assessment (Xu and Singh, 1998), streamflow forecasting (Alley, 1984; Schar et al., 2004), climate change impact assessment (Gleick, 1987; Bastola et al., 2011; Chen et al., 2011), and snowmelt runoff simulation (Xu et al., 1996; Rezaeianzadeh et al., 2013). Lumped MWBMs tend to oversimplify the complexity of hydrological processes, which casts doubt on their capacity to predict seasonal flows under various climate conditions (Dakhlaoui et al., 2017; Hamel et al., 2017). In the majority of widely used MWBMs, such as the Dynamic Water Balance Model (DWBM) (Zhang et al., 2008), Belgium Model (VUB) (Vandewiele et al., 1992), Time Variant Gain Model (TVGM) (Xia et al., 1997), WatBal Model (WM) (Leaf and Brink, 1973) and Schaake Model (SM) (Schaake, 1990), runoff generation process consists of quick flow generation and baseflow generation, referred as runoff partitioning structure. To improve the performance of MWBMs for simulating total streamflow, Bai et al. (2015) modified the evapotranspiration equations but the performances of total streamflow have no significant improvement. The improvement of total flow performance in MWBMs should focus on runoff generation mechanisms rather than actual evapotranspiration process (Vandewiele et al., 1992; Bai et al., 2015). However, studies on the deficiencies in MWBMs are limited and it is important to assess the model structure for runoff generations, especially the baseflow that is of critical importance for water resource management and ecosystem health.

Evaluation of hydrological behaviours extracted from total streamflow can guide model improvements in a meaningful way (Gupta et al., 2008; Yilmaz et al., 2008) and achieve realistic hydrological simulation (Duan et al., 2006; McMillan, 2020). For MWBMs with runoff partitioning structure, performances of quick flow and baseflow provide new insight of internal model behaviour, which can be directly used to detect the deficiency of runoff partitioning structure (Shafii et al., 2019). The ideal structure of MWBMs is supposed to achieve realistic representation of the real world, namely keeping acceptably accurate simulations of not only total streamflow but also quick flow and baseflow (Gupta et al., 2008; Euser et al., 2013; Khatami et al., 2019). However, good performance of total streamflow does not necessarily mean internal model processes are correct as it may be achieved under insufficient parameterization constraints and improper conceptualization of hydrological processes in real-world systems (Hrachowitz et al., 2014). Dynamics in quick flow and baseflow can be improperly simulated due to the weaknesses in calibration procedures (Beven, 1993; Bai et al., 2018) and structural inadequacy (Shafii et al., 2017). Many approaches have been proposed to improve calibration procedures such as multi-objective optimization framework (Shafii and Tolson, 2015; Kelleher et al., 2017; He et al., 2018; Larabi et al., 2018; Schuite et al., 2019), temporal variation of parameters

(Deng et al., 2018; Xiong et al., 2019) and alternative calibration criteria (Gupta et al., 2009; Larabi et al., 2018; Fowler et al., 2018a). In these studies, observed signals of quick flow and baseflow have been incorporated into multi-objective optimization framework, which results in reliable performance of quick flow and baseflow. However, studies to achieve realistic simulation of total streamflow, quick flow and baseflow through evaluating and developing runoff partitioning structure in MWBMs are still very limited, especially the baseflow modelling structure (Westra et al., 2014; Fowler et al., 2018b).

Although most MWBMs use a linear storage-discharge relationship to describe storage-discharge dynamics, the storage-discharge relationship at monthly time scale is still unclear. Catchment storage-discharge relationship is traditionally established at event or daily time scales in previous studies and is rarely investigated at monthly time scale. At event or daily time scale, there has been an on-going discussion for decades that whether the storage-discharge relationship is linear or nonlinear (Moore, 1997; Wittenberg, 1999; Lee, 2007). Various linear and nonlinear storage-discharge relationships have been developed (Stoelzle et al., 2015) via mathematical derivation (Duffy, 1996), recession analysis (Chapman, 1999; Aksoy and Wittenberg, 2011; Cheng et al., 2016) and hydrological modelling analysis (Fenicia et al., 2006; Markovic and Koch, 2015). At short time scale, the linear storage-discharge relationship has solid physical basis, which conceptualize moisture storage as a straight-sided bucket with a hole at the bottom (Beven, 2001). The moisture storage at short time scale is assumed to be replenished by previous rainfall events and the recharge from the current rainfall event is typically neglected (Buttle, 1994; Wittenberg, 1999). However, at monthly time scale, the mechanism of baseflow generation is different because the recharge to soil water storage from precipitation at the current month must be considered (Lindstrom et al., 1997; Hrachowitz et al., 2014). Therefore, the linear storage-discharge relationships based on straight-sided bucket conceptualization in MWBMs have to be carefully investigated.

To evaluate the baseflow modelling structure in monthly water balance models, 5 widely used MWBMs (the DWBM, VUB, TVGM, WM and SM) with both quick flow and baseflow generation processes are selected. The 5 selected MWBMs all adopt a linear storage-discharge relationship to describe baseflow generation mechanism at monthly time scale. Observed hydroclimatic data from 443 catchments across Australia with a wide range of climatic and physiographical conditions are collected to test the performance of models with two different types of catchment storage-discharge relationships. First, the performances of these 5 MWBMs in their original form (*i.e.* with a linear storage-discharge relationship) are assessed in terms of their capability for simulating total streamflow, quick flow and baseflow. Then, the linear storage-discharge relationships in all selected MWBMs are replaced with a nonlinear exponential relationship proposed by Peters and Aulenbach (2011) (hereafter denoted as PA11). The performances of modified MWBMs are evaluated for simulating total flow and baseflow. The primary objectives of this study are: (1) to diagnose the performance in runoff partitioning structure of 5 widely applied MWBMs with both quick flow and baseflow generation processes; (2) to examine the influences of nonlinear storage-discharge relationship on the capability of MWBMs for simulating total streamflow; (3) to examine the ability of nonlinear storage-discharge relationship for MWBMs to achieve realistic hydrological modelling performance in terms of baseflow.

2 Study catchments and data

Daily streamflow of 443 un-nested catchments in Australia with minimum human interferences (without dams, intensive irrigation and land use change) are collected to test the performance of MWBMs (Figure 1). All these catchments are part of the Australia unregulated catchment dataset (Zhang et al., 2013). The collected streamflow, precipitation and potential evapotranspiration data span over the period of 1975-2012. All the catchments have a minimum length of 20-year records with at least 10-year continuous records and less than 10% missing daily data in total. The drainage area ranges from the order of 10 to 10000 km^2. Based on the Köppen-Geiger climate classification map produced by Kottek et al. (2006), the 443 catchments cover all the 5 distinct climatic zones in Australia including tropics, arid, equiseasonal-hot, equiseasonal-warm and winter rainfall dominant (see Figure 1). The number of the catchments within tropics, arid, equiseasonal-hot, equiseasonal-warm and winter rainfall climate zones is 56, 50, 105, 171 and 61, respectively. The average precipitation of all catchments is 958±421 (mean ± standard deviation), potential evapotranspiration is 1411±294, aridity index is 1.76±1.01, runoff coefficient is 0.19±0.15 and baseflow index is 0.28±0.15. The coefficient of variation (CV) of monthly precipitation, defined as the ratio of standard deviation (σ) to mean (μ) monthly precipitations, is 0.89±0.31. The CV of monthly runoff, representing the integrated effects of geological and climatic characteristics on catchments runoff, is 2.36±1.33 (ranges see Table 1).

3 Methodology

3.1 Separation of quick flow and baseflow

Daily observed total streamflow is separated into daily quick flow and daily baseflow using the Lyne-Hollick (denoted as LH) method (Lyne and Hollick, 1979). The LH method is adopted in this study not only because it has been widely applied worldwide (Ahiablame et al., 2013), but also due the reason that it yields practically equivalent results as other complex physical methods (Cheng et al., 2012; Zhang et al., 2017). The principle of this separation method is based on signal processing theory. According to the high frequency characteristic of quick flow, the filter equation for quick flow is expressed as:

$$Q_{d(i)} = f_1 Q_{d(i-1)} + \frac{1+f_1}{2}(Q_{(i)} - Q_{(i-1)}) \quad (1)$$

where Q is total streamflow (mm d^{-1}); Q_d is quick flow (mm d^{-1}); i is the index of time step; and f_1 is the filter parameter (unit of d^{-1}), which is also called the recession constant. Baseflow Q_b can be calculated subsequently by:

$$Q_{b(i)} = \begin{cases} Q_{(i)} - Q_{d(i)}, & Q_{(i)} > Q_{d(i)} \\ 0 \quad , & Q_{(i)} \le Q_{d(i)} \end{cases} \tag{2}$$

Here the digital filter is applied in a traditional way, *i.e.* baseflow is separated from total flow with three passes (forward, backward and forward) and the filter parameter (recession constant) f_1 is set to 0.925 as suggested by Nathan and McMahon (1990). Separated daily quick flow and baseflow are aggregated to monthly values and are taken as the observed monthly quick flow and baseflow for evaluating model performance.

3.2 Descriptions of the MWBMs

In this study, 5 widely applied monthly water balance models with both quick flow and baseflow generation process are chosen to assess the runoff partitioning structure in these MWBMs. They are the Dynamic Water Balance Model (DWBM) (Zhang et al., 2008), the Belgium Model (VUB) (Vandewiele et al., 1992), the Time Variant Gain Model (TVGM) (Xia et al., 2005), the WatBal Model (WM) (Leaf and Brink, 1973) and the Schaake Model (SM) (Schaake, 1990). The general water balance equation of all these models can be expressed as:

$$(S(t) - S(t-1))/\Delta t = P(t) - E_a(t) - Q_d(t) - Q_b(t) \tag{3}$$

where $S(t\text{-}1)$ and $S(t)$ are the soil moisture storage (unit of mm) at the beginning and end of the time interval t, respectively; $P(t)$, $E_a(t)$, $Q_d(t)$, $Q_b(t)$ are precipitation, actual evapotranspiration, quick flow and baseflow, respectively. Unit of $P(t)$, $E_a(t)$, $Q_d(t)$ and $Q_b(t)$ is mm month^{-1}. For a given time step (*i.e.* month), Δt is equal to

1. Basically, all the selected 5 MWBMs have similar conceptual structure for estimating actual evapotranspiration (E_a), quick flow (Q_d) and baseflow (Q_b). The only differences in model structure are the number of water storages and whether equations for estimating different components of water budget are linear or not. The structure of 5 MWBMs are shown in Figure 2. Equations for simulating actual evapotranspiration, quick flow and baseflow of the 5 models are summarized in Table 2. The symbols w_1-w_{17} (see Figure 2 and Table 2) represent serial numbers of the equations of 5 original MWBMs. Detailed descriptions of all the 5 models are provided in the Appendix. Major similarities and differences are briefly summarized here.

Four of the 5 model (*i.e.* VUB, TVGM, WM and SM) have only one soil water storage to estimate Q_d, E_a and Q_b. Only the DWBM has two soil water storages (*i.e.* upper soil water storage (S) and lower groundwater storage (G)), and soil water in upper storage can recharge to the lower groundwater storage. In the DWBM, Q_d and E_a are generated from the upper soil water storage, while Q_b is generated from lower groundwater storage.

Table 3 summarizes whether equations for estimating quick flow, baseflow and evapotranspiration of different MWBMs are linear or not. For quick runoff (Q_d), it is simulated as a nonlinear function of precipitation and amount of soil water in all 5 models. With respect to baseflow (Q_b), all the selected models estimate Q_b using a linear storage-discharge relationship. Regarding to the actual evapotranspiration (E_a), all selected models estimate E_a as a function of soil moisture and potential evapotranspiration. The SM and WM adopt a simple linear function to estimate monthly E_a, while other models use nonlinear functions.

3.3 Modification of baseflow generation mechanism

Table 4 shows the modification of the linear storage-discharge relationship in 5 original MWBMs to a nonlinear storage-discharge relationship proposed by Peters and Aulenbach (2011) (hereafter denoted as PA11) with parameterization and equations for estimating E_a, Q_d and S are all kept unchanged. The PA11 can be written as:

$$W(t) = S(t-1) + P(t) \tag{22}$$

$$Q_b(t) = e^{(W(t)-b)/m} \tag{23}$$

where $Q_b(t)$ is baseflow at time step t; $W(t)$ is the available water to generate baseflow; $S(t-1)$ is catchment soil moisture storage at time step $t-1$; $P(t)$ is precipitation at time step t; b and m are constants to be calibrated. Parameter m determines the nonlinear variability between $W(t)$ and $Q_b(t)$. Parameter b mainly influences the magnitude of $Q_b(t)$.

In the PA11, soil moisture storage S includes both shallow soil water storage and groundwater storage as defined by Aulenbach and Peters (2018) and thus all MWBMs are supposed to have only one moisture storage. Four of the five study MWBMs (except the DWBM) have only one water storage and thus the storage structure of these four models are kept the same. In both original and modified forms of these four models, moisture storage supplies water for E_a, Q_d and Q_b. Storage structure of the DWBM model is changed to replace the linear storage-discharge relationship to a nonlinear one. The original DWBM with two water storages is restructured to one storage to incorporate the PA11. For the original DWBM, upper storage (*i.e.* soil storage *S*) supplies water for actual evapotranspiration (E_a) and discharge (*R*) to the lower storage (*i.e.* groundwater storage *G*), from which baseflow (Q_b) is generated. While in the modified DWBM (denoted as $\mathrm{DWBM_{mod}}$), both E_a and Q_b are generated from the same united soil storage. Meanwhile, in all the models, equations for estimation Q_d, E_a and *S* are all kept unchanged, which are shown in Table 2.

3.4 Parameter estimation and model evaluation

In this study, parameters are calibrated using an automatic optimization technique, Genetic Algorithm (GA) (Grefenstette et al., 1986). Five criteria are selected to assess model performance including the Nash-Sutcliffe efficiency (NSE, (Nash and Sutcliffe, 1970)), logarithmic form of NSE (NSE(log)), Pearson correlation coefficient (*r*), Bias Score (*BS*) (Wang et al., 2011) and Bias (*B*). The objective function (F_{opt}), which is used to optimize parameter sets, combines four criteria (NSE, NSE(log), *r* and *BS*) that can minimize both systematic (*e.g.* *BS*) and dynamic error (*e.g.* NSE and NSE(log)) between the simulated and observed high and low flows (Krause et al., 2005). The mathematic formulations of the five criteria and F_{opt} are as follows:

$$\text{NSE} = 1 - \frac{\sum_{t=1}^{n}(Q_{sim}(t) - Q_{obs}(t))^2}{\sum_{t=1}^{n}(Q_{obs}(t) - \overline{Q_{obs}})^2} \tag{24}$$

$$\text{NSE(log)} = 1 - \frac{\sum_{t=1}^{n}(\ln(Q_{sim}(t)) - \ln(Q_{obs}(t)))^2}{\sum_{t=1}^{n}(\ln(Q_{obs}(t)) - \overline{\ln(Q_{obs})})^2} \tag{25}$$

$$r = \frac{\sum_{t=1}^{n}(Q_{sim}(t) - \overline{Q_{sim}})(Q_{obs}(t) - \overline{Q_{obs}})}{\sqrt{\sum_{t=1}^{n}(Q_{sim}(t) - \overline{Q_{sim}})^2 \sum_{t=1}^{n}(Q_{obs}(t) - \overline{Q_{obs}})^2}} \tag{26}$$

$$BS = 1 - (\max\left(\frac{\overline{Q_{sim}}}{\overline{Q_{obs}}}, \frac{\overline{Q_{obs}}}{\overline{Q_{sim}}}\right) - 1)^2 \tag{27}$$

$$B = 1 - \text{abs}(\frac{\overline{Q_{sim}} - \overline{Q_{obs}}}{\overline{Q_{obs}}}) \tag{28}$$

$$F_{opt} = (NSE + NSE(log) + r + BS)/4 \tag{29}$$

$$F_{avg} = (NSE + NSE(log) + r + B)/4 \tag{30}$$

where $Q_{sim}(t)$ and $Q_{obs}(t)$ are the simulated and observed flow at time step *t*, respectively; variables with overbar denote average value; *n* is the number of months during the study period.

In this study, parameters of both original and modified models are calibrated against observed total streamflow only by maximizing the value of F_{opt}. Separated quick runoff and baseflow are not used to calibrate parameters but are used to assess the capability of original and modified MWBMs for simulating different flow components. Model performances for simulating total flow, baseflow and quick flow are evaluated by NSE, NSE(log), *r*, *B* and F_{avg}. The *BS* is much more sensitive to very poorly simulated flows than *B*. The *BS* is used in F_{opt} for parameter optimization to guarantee much more suitable parameters for baseflow simulation that can be identified. When calibrated against total flow only, total volume of baseflow may not be well simulated in a few catchments, which can result in large negative values of F_{opt} and makes comparison and visualization of results of all the catchments very difficult. Therefore, bias (B), which can also measure the systematic error as the BS, is chosen to calculate performance index F_{avg} for evaluation.

The NSE, NSE(log), BS and B can vary from $-\infty$ to 1.0 and r can vary from −1.0 to 1.0. The closer the NSE, NSE(log), r, BS and B approach 1.0, the better the model performs. The NSE=1.0 or NSE(log)=1.0 means simulated flows are exactly the same as observed flows in every time step. The r=1 means the predicted flows show a complete linear relationship with the observed flows. The BS=1.0 or B=1.0 means the volume of simulated and observed flows are the same and there is no systematic error. For the evaluation of baseflow performance, logarithmic form of NSE (NSE(log)) and correlation coefficient (r) are more suitable than NSE and B because baseflow is typically a few orders of magnitude smaller than total flow and quick flow, which will be discussed in section 5.1.

4 Results

4.1 Performances of the original MWBMs

Performances of the 5 MWBMs in their original forms for estimating total flow (Q), quick flow (Q_d) and baseflow (Q_b) in all the 443 catchments are shown in Figure 3. Basically, all the MWBMs perform satisfactorily in simulating total streamflow and quick flow. However, all the models perform poorly in simulating baseflow.

As for the performance of total flow (Q), the median F_{avg} of all the models is larger than 0.68 with a range of 0.68 ~ 0.77 (see Table 5). The DWBM and VUB have the best performance with median F_{avg}=0.77, followed by the TVGM (0.71), SM (0.71), and WM (0.68). The inter-quantile range (IQR, *i.e.* range between 75th and the 25th percentiles) of F_{avg} varies from 0.11 to 0.19. The VUB model is the most robust model with an IQR of 0.11, followed by the DWBM (0.12), SM (0.14), WM (0.18), and TVGM (0.19). The median F_{avg} of all the five MWBMs is quite far from the perfect match between the observed and simulated total flow, *i.e.* both $F_{avg} = 1.0$ and $IQR = 0.0$.

Regarding to the performance of quick flow (Q_d), the median F_{avg} of all the models is higher than 0.52 with a range of 0.52 ~ 0.63. The VUB has the best performance with median F_{avg}=0.63, followed by the SM (0.57), DWBM (0.56), TVGM (0.55), and WM (0.52). The IQR of F_{avg} is smaller than 0.27 with a range from 0.15 to 0.27. The TVGM is the most robust model with an IQR of 0.15, followed by the WM (0.16), VUB (0.17), SM (0.19), and DWBM (0.27). The median F_{avg} and IQR of quick flow are roughly as good as those of total flow for all the 5 MWBMs, which means the parameterization schemes of the quick flow are accurate in all these models.

With respect to the performance of baseflow (Q_b), the median F_{avg} of all the models is smaller than 0.39 with a range of 0.11 ~ 0.39. The SM has the best performance with median F_{avg}=0.39, followed by the DWBM (0.30), VUB (0.12), WM (0.12), and TVGM (0.11). The IQR of F_{avg} ranges from 0.21 to 0.42. The VUB is the most robust model with IQR equal to 0.21, followed by the SM (0.27), TVGM (0.33), DWBM (0.39), and WM (0.42). For baseflow, the median F_{avg} is almost three times smaller and the IQR is about twice wider than those of total flow.

Comparison of observed and simulated monthly baseflow by all the 5 MWBMs in their original forms in the 443 catchments over the study period are shown in Figure 4. The baseflow is significantly underestimated by all the models about $-60 \pm 36\%$. The median Pearson correlation coefficient (r) between baseflow estimated by the 5 original MWBMs and observed baseflow is smaller than 0.62 with a range of 0.48 ~ 0.62. These results indicate the linear storage-discharge relationships in all the 5 MWBMs are not appropriate for baseflow simulation. Therefore, model structure for simulating baseflow in these MWBMs has to be modified to improve model performances of both baseflow and total streamflow.

4.2 Performances of the modified MWBMs in simulating total streamflow

Figure 5 shows total flow performances of the 5 MWBMs together in their original and modified forms across 443 study catchments. The modified models outperform original models clearly in terms of the NSE (Figure 5a) and NSE(log) (Figure 5b), marginally in terms of the r (Figure 5c) and B (Figure 5d). Performances of the modified models in terms of the objective values of 4 evaluation indices using box-plot are compared with original models in Figure 3 as well in all the study catchments. Figure 6 shows the changes in model performance between modified and original MWBMs individually in simulating total streamflow. The modified MWBMs outperform the original models on total streamflow in terms of the percentages of catchments that have a better performance (Figure 6a), median increased value (Figure 6b) and change in IQR (Figure 6c) of all study catchments in four different criteria (*i.e.* NSE, NSE(log), *r* and *B*).

For the criterion of NSE, all modified MWBMs have higher NSE in most study catchments. All the models show smaller IQR compared with the original models, except for the VUB model (see Table 6). The modified models have higher NSE in 82±4.0% catchments with a range of 72% ~ 93%. The WM has the largest proportion of catchments that is improved (93%), followed by the VUB (85%), TVGM (82%), DWBM (77%) and SM (72%). The median improved NSE for all the study catchments is 0.03±0.007 with a range of 0.01 ~ 0.05. The WM has the largest median improved NSE (0.05), followed by the SM (0.03), DWBM (0.03), TVGM (0.02), and VUB (0.01). The IQR of NSE reduces about 0.02±0.02 with a range of −0.05 to 0.06. The DWBM has the largest reduction of IQR (0.06), followed by the SM (0.05), WM (0.03), TVGM (0.02) and VUB (−0.05).

All the 5 modified models have higher NSE(log) in most study catchments and different change of IQR compared with those of the original models. Compared with original models, modified models are better in simulating total streamflow in 72±4.7%

catchments with a range of 61% ~ 81% in term of NSE(log). The TVGM has the largest proportion of catchments that is improved (81%), followed by the DWBM (79%), WM (77%), SM (63%) and VUB (61%). The median improved NSE(log) for all the study catchments is 0.03±0.008 with a range of 0.01 ~ 0.05. The DWBM has the largest median improved NSE(log) (0.05), followed by the WM (0.04), TVGM (0.04), SM (0.02) and VUB (0.01). The IQR of NSE(log) has reduced about 0.002±0.02 with a range of −0.04 to 0.06. The DWBM has the largest reduction of IQR (0.06), followed by the SM (0.03), WM (0.00), TVGM (−0.04) and VUB (−0.04).

For the criterion of *r*, the 5 modified models also have marginal higher *r* in most study catchments and smaller IQR. The modified models perform better for simulating total streamflow in 76±4.5% catchments with a range of 61% ~ 86% in terms of *r*. The WM has the largest proportion of catchments that is improved (86%), followed by the VUB (79%), DWBM (79%), TVGM (77%) and SM (61%). The median improved *r* for all the study catchments is 0.01±0.002 with a range of 0.00 ~ 0.01. Expect for the SM (0.00), the median improved *r* of other models values 0.01. The IQR of *r* has reduced about 0.02±0.01 with a range of 0.00 to 0.04. The DWBM has the largest reduction of IQR (0.04), followed by the SM (0.02), WM (0.02), TVGM (0.01) and VUB (0.00).

For the criterion of *B*, the 5 modified models have marginal improvement of *B* and marginal reduction of IQR. The modified models have improvement in 51±2.4%

study catchments. The median improved B of the 5 modified model is 0.002±0.002 and mean reduction of IQR is 0.004±0.003.

In summary, compared to original models, the NSE, NSE(log), r and B of modified models are better for the simulation of total streamflow in 82±4.0%, 72±4.7%, 76%±4.5% and 51±2.4% study catchments, respectively. The median improved NSE, NSE(log), r and B are 0.03±0.007, 0.03±0.008, 0.01±0.002 and 0.002±0.002, respectively. The IQR of NSE, NSE(log), r and B have reduced about 0.02±0.02, 0.002±0.02, 0.02±0.01, 0.004±0.003, respectively. Increase in model performance and decrease in IQR suggest that MWBMs became more reliable and robust by replacing the linear storage-discharge relationship with an exponential nonlinear (*i.e.* PA11) relationship.

4.3 Performance of modified models in simulating baseflow

Figure 7 shows the comparison of baseflow performance of the 5 MWBMs together between their original and modified forms across 443 study catchments. The modified models outperform original models clearly on the NSE (when NSE>0) (Figure 7a), NSE(log) (Figure 7b) and r (Figure 7c). Both original and modified MWBMs have poor NSE with nearly 70% catchments smaller than 0. Figure 8 shows the changes in model performances between modified and original MWBMs for simulating baseflow individually. Basically, the modified MWBMs outperform the original models in terms of NSE (log) and r, but underperform original models in terms of NSE (all catchments) and *B*. All 5 modified MWBMs have much higher *r* and NSE(log) in 83±4.1% and 68±4.6% study catchments comparing with those of the original models, respectively (Figure 8a, Table 7). The median improved *r* and NSE(log) of all study catchments are 0.14±0.03 and 0.17±0.03, respectively (Figure 8b). Change in IQR of *r* is marginal (0.03±0.04, Figure 8c). The IQR of NSE(log) has reduced significantly about 0.12±0.08. For the criteria of NSE and *B*, simulated baseflow using modified models is better than that using original ones in about half study catchments. For NSE, the modified MWBMs perform better in 41±7.2% but worse in 59±7.2% catchments. For *B*, the modified MWBMs perform better and worse in 46±8.3% and 54±8.3% catchments, respectively. The median improved NSE and *B* of all study catchments are −0.08±0.05 and −0.04±0.06, respectively. The change of IQR of NSE and *B* are 0.99±0.57 and 0.27±0.06, respectively.

The increased NSE(log) and *r* suggest general improvement of baseflow simulation using nonlinear baseflow modelling structure because NSE(log) is more suitable than NSE to evaluate the performance of baseflow and *r* is the most direct criterion to evaluate whether the storage-discharge relationship is exponential nonlinear or not. According to the equation (31), the much higher r (*i.e.* higher A in Eq.31) of the modified MWBMs provides a precondition of model structure for MWBMs to have higher NSE on baseflow simulation. It directly indicates that the nonlinear relationship (*i.e.* PA11) is better than the linear relationship to capture catchment storage-discharge dynamics at monthly time scale.

5 Discussion

5.1 Characteristics of different criterion and their suitability for evaluation the performance of baseflow

Every criterion has advantages and disadvantages in quantifying the agreement between observed and simulated flows. The Nash-Sutcliffe efficiency (NSE) proposed by Nash and Sutcliffe (1970) is a widely used assessment criterion. The NSE is largely a dynamic indicator and it is very sensitive to high flows and is insensitive to low flows because of its squared formulation (Legates and McCabe Jr., 1999). To compensate the disadvantage of NSE, NSE(log) is used to give more weights on low flows in the performance assessment. Pearson correlation coefficient (*r*) measures the co-variability of the simulated and observed flows, which describes how much of the dispersion in observed flows is explained by the simulated flows. The *BS* and *B* are employed to measure symmetric error between simulated and observed flows.

The four different criteria (*i.e.* NSE, NSE(log), *r* and *B*) provide useful insight into basic characteristics of simulation performance. For the evaluation of baseflow performance, logarithmic form of NSE (*i.e.* NSE(log)) and correlation coefficient (*r*) are more important than NSE and *B*. The NSE(log) is more suitable than NSE for baseflow evaluation because baseflow is typically a few orders of magnitude smaller than the quick flow that is generated during the heavy rainfall events. The *r* is the most straightforward criterion among these four selected criteria to indicate the storage-discharge relationship is linear or nonlinear because *r* measures the degree of linear association between the observed and simulated baseflow. Thus r is the most powerful criterion to evaluate baseflow generation structure in this study. Moreover, higher value of linear correlation coefficient (r) is the precondition for higher value of Nash-Sutcliffe efficiency (NSE) because NSE can be decomposed three components as advised by Murphy (1988):

$$NSE = A - M - N = r^2 - \left[r - \left(\frac{\sigma_s}{\sigma_0}\right)\right]^2 - [(\mu_s - \mu_0)/\sigma_0)]^2 \qquad (31)$$

where r is the linear correlation coefficient; (μ_s, σ_s) and (μ_0, σ_0) represent the first two statistical moments (means and standard deviations) of simulated and observed sequences, respectively. The quantity A measures the strength of the linear relationship between simulated and observed values, M measures the conditional bias, and N measures the unconditional bias. Higher value of NSE depends on higher A, as well as lower M and N. That is to say, higher NSE is achieved by both higher r and lower bias. In this study, the much high r of modified MWBMs for simulating baseflow provide precondition for higher value of Nash-Sutcliffe efficiency (NSE).

5.2 Different control of the two parameters in the PA11 method on baseflow simulation

The capability of PA11 can be evidenced by comparison of the variability and magnitude of simulated and observed baseflow, which can be measured by r and B, respectively. The variability and magnitude of baseflow are controlled by different parameters in the PA11 approach. The PA11 method (equation (23)) can be reformulated as:

$$Q_b(t) = e^{\left(\frac{W(t)}{m}\right)} / e^{b/m} \tag{32}$$

In terms of equation (32), it can be found that the value of simulated baseflow is determined by two parts. One is the nonlinearity between $W(t)$ and $Q_b(t)$ (*i.e.* $e^{\left(\frac{W(t)}{m}\right)}$). The other is the magnitude of $Q_b(t)$ (*i.e.* $e^{b/m}$). The first part represents the nonlinear structure between $W(t)$ and $Q_b(t)$, and the second part only includes parameters m and b. Linear correlation (r) between observed and simulated baseflow is only controlled by the first part. Thus the criterion r is the most direct criterion to evaluate whether the storage-discharge relationship is exponential nonlinear or not, which is controlled only by m. In other words, the ability of the nonlinear baseflow modelling structure is only controlled by parameter m and directly measured by criterion r. While NSE(log), NSE and B are determined by both parts, which is controlled by both m and b.

Take DWBM as an example, DWBM_{mod} (*i.e.* modified DWBM) can capture the variability of baseflow, but it underestimates apparently the magnitude of baseflow. Figure 9 shows observed and simulated baseflow sequences for catchment 238204 (Figure 9a) and catchment 108002 (Figure 9b). The baseflow simulated by DWBM_{mod} has the same variability with the observed baseflow, *i.e.* both increase and decrease simultaneously and peak at the same time, but they have different magnitudes.

The differences in magnitude can be attributed two reasons. One is the uncertainty of observed baseflow derived from the LH method. The other one is the poorly calibrated parameter *b*, which determines the magnitude of simulated baseflow. The magnitude of baseflow is much smaller than that of total and direct flows, thus the magnitude of baseflow is easily poorly simulated with poorly calibrated parameter *b*. Figure 10 shows the comparison of baseflow derived from LH method used in this study with the other two baseflow separation methods, *i.e.* the United Kingdom Institute of Hydrology (UKIH) method (Richards, 1994) and the Chapman-Maxwell (CM) method (Chapman and Maxwell, 1996). Baseflows derived from three digital filter methods have the same temporal variability, but have different magnitudes. Considering that the variability of baseflow has been well captured by $DWBM_{mod}$, the difference in magnitude can be further reduced by adjusting the parameter *b*. In $DWBM_{mod}$, under-estimation of baseflow in catchment 238204 means *b* is overestimated. Over-estimation of baseflow in catchment 108002 means *b* is underestimated. The poorly calibrated *b* in $DWBM_{mod}$ is adjusted (hereafter denoted as Adjusted-*b*-$DWBM_{mod}$) through minimizing the criterion *B*. The details of adjustment of parameter *b* is shown in Table 8. As shown in Figure 11, the magnitude difference of baseflow decreased in Adjusted-*b*-$DWBM_{mod}$ with higher NSE, NSE (log) and *B* than those of both DWBM and $DWBM_{mod}$.

However, in this study, further calibration of baseflow parameters (*b*) against the separated (or "observed") baseflow to get a better performance of baseflow is not considered. The calibration procedure adapted in this study is to calibrate both original and modified models against total flow only due to lack of direct measurement of baseflow. Separated slow component from hydrographs using widely used baseflow separation method (such as the LH method used in this study) may not be strictly considered as baseflow (Klaus and McDonnell, 2013; Pelletier and Andreassian, 2020). The baseflow modelling structure of modified MWBMs can catch the variability of baseflow from all three digital filter methods, the magnitude difference of baseflow can attribute to the uncertainty of baseflow separation method and calibration process. Here we just demonstrate the superiority of nonlinear baseflow modelling structure to capture storage baseflow dynamics at monthly time scale. More studies on the calibration of baseflow parameters still required in the future to improve the performance of MWBMs. But it is beyond the scope of this study.

5.3 Considering model consistency for structure evaluation

The structure of the five models consists of several components, representing different hydrological processes. The evaluation of baseflow performance can be referred as “model consistency” evaluation, defined as the ability of a model structure to adequately reproduce several hydrological signatures simultaneously while using the same set of parameter values (Euser et al., 2013). The consistency is considered important for evaluating model structure because consistency can achieve the realistic representation of the real world and reduce equifinality (McMillan, 2020). The improved performance of baseflow using modified MWBMs is resulted from more reasonable baseflow modelling structure with only one more parameter rather than overparameterization or equifinality and thus overparameterization is not evaluated here.

In this study, the general improvement of baseflow performance indicated that the nonlinear storage-baseflow relationship can improve the consistency of MWBMs. The more realistic modelling in the modified MWBMs can achieve the least uncertainty in simulating not only total streamflow (Kumar, 2011) but also baseflow. During past few decades, both quantity and quality of baseflow have received increased attention (Arnold et al., 1995) such as sustaining aquatic habitats (Poff et al., 1997; Fan et al., 2013) and dynamics of chemicals in watersheds (Shafii et al., 2019). Accurate simulation of baseflow in MWBMs will extend the capability and application of MWBMs. Thus, improvement of MWBMs structure should consider the consistency of several hydrological signatures to achieve realism of hydrological processes instead of focusing on total streamflow only.

5.4 Nonlinear exponential storage-discharge relationship for baseflow estimation

The rationality and physical basis of the nonlinear exponential storage-discharge relationship to describe baseflow process have been proved by previous studies through reservoir conceptualization and recession analysis (Brutsaert and Nieber, 1977; Stoelzle et al., 2015; Nippgen et al., 2016). The baseflow process is complex and nonlinear due to the joint control of hydroclimatic conditions and geological characteristics on baseflow generation (Maneta et al., 2018). The nonlinearity of baseflow process is widely observed in catchment storage-discharge relationship. At lower baseflows, large changes in soil moisture are related to relatively small change in baseflow; while at higher baseflows, small changes in soil moisture result in relatively large changes in baseflow (Nippgen et al., 2016). Based on reservoir conceptualization, the nonlinear storage-discharge relationship is usually described by combination of several linear reservoirs or single nonlinear reservoir (Stoelzle et al., 2015). For the case of single nonlinear reservoir, baseflow is typically estimated using a power function (Harman and Sivapalan, 2009) or an exponential function (Beven and Kirkby, 1979). Peters and Aulenbach (2011) proposed the PA11 model in virtue of observed soil moisture using the exponential function. Aulenbach and Peters (2018) showed that the exponential function (*i.e.* the PA11) can well describe the storage-discharge dynamics with a high coefficient of determination (adjusted $R^2 = 0.96$) using observed soil moisture and estimated baseflow from Eckhardt filter method (Eckhardt, 2005). The nonlinear storage-discharge relationship can also be derived from recession analysis. In Kirchner (2009), the recession curve is described by a power law function $-\frac{dQ}{dt} = aQ^b$ based on the fundamental works of Horton (1941) and Brutsaert and Nieber (1977). The storage-discharge relationship can be linear, power, exponential, or more than exponential when b=1, b<2, b=2, b>2, respectively. Patnaik et al. (2018) found the median *b* of the recession curve of 358 catchments in the United States nearly equals to 2, which indicates that the storage-discharge relationship is exponential in most catchments. In this study, the nonlinear exponential storage-discharge relationship in modified MWBMs improve model performance in terms of both total flow and baseflow compared with the linear storage-discharge relationship in original MWBMs. Therefore, the nonlinear exponential storage-discharge relationship may have stronger physical basis and is more universal than linear storage-discharge relationship.

5.5 Monthly versus daily models for baseflow simulation

Within-month variability of hydrological variables and the storage response to daily rainfall events are two main factors that lead to different baseflow generation mechanisms at daily and monthly time scale. The two factors need to be considered in the nonlinear or linear forms of the storage-discharge relationship. Baseflow is possible to be measured at daily and/or hourly time scale through rigours flow recession analysis (Cheng et al., 2016) and tracer-based methods (Gonzales et al., 2009), while it is difficult to be measured at monthly time scale. Based on observed hydrological variables, the nonlinear and linear storage-discharge relationships can be derived from reservoir conceptualization and recession analysis at short time scale (*i.e.* daily and hourly). For the MWBMs investigated in this study, catchment storage-discharge relationships at short time scale are directly adopted without considering the within-month variability in climate forcing variables and rainfall-storage responses. From modelling perspective, Wang et al. (2011) compared the monthly and daily models for the simulation of monthly total runoff and reported that the monthly models have not been disadvantaged for not using the within-month temporal sequences of the forcing variables. The other factor, *i.e.* storage response to rainfall events at current month, is ignored by original MWBMs and leads to apparent lag in the peak time of baseflow. Increased correlation between observed and simulated baseflow using modified MWBMs is probably resulted from adding precipitation to storage at the current month for baseflow generation. No obvious correlation has been found between increased model performance on baseflow and catchments properties such as aridity index, elevation, slope, soil properties, *etc*. From a modelling perspective, monthly storage-baseflow relationship is investigated in this study and results indicate that the nonlinear relationship is more effectively to capture the variability of monthly baseflow at most catchments. However, further studies are still required to advance our capability in simulating baseflow across various spatial and time scales.

6 Conclusions

In this study, the performance of linear storage-discharge relationship in 5 widely used monthly water balance models (MWBMs) is diagnosed and evaluated using observed daily hydrological data from 443 catchments across Australia with distinct hydro-climatic conditions. A nonlinear exponential storage-discharge relationship (*i.e.* the PA11) is employed to replace the linear one in the study MWBMs to improve monthly baseflow modelling accuracy and to achieve realistic hydrological modelling at monthly time scale. The main findings are summarized as follows:

(1). Baseflow simulated by 5 original MWBMs are remarkably underestimated and unable to explain the dispersion of observed baseflow. The poor performance of baseflow suggests the linear baseflow generation mechanism may not be suitable for monthly water balance models.

(2). Modified MWBMs with nonlinear baseflow modelling structure outperform the original ones in simulating total flow. On average, the criteria NSE, NSE(log), r and B of modified models are improved in 82±4.0%, 72±4.7%, 76%±4.5% and 51±2.4% of study catchments, respectively.

(3). The modified MWBMs improve baseflow performance significantly with better NSE(log) and r in 68±4.6% and 83±4.1% study catchments with median improvement of 0.17±0.03 and 0.14±0.03, respectively.

These results suggest that the modified MWBMs with the nonlinear storage-discharge relationship is more capable than the original MWBMs with the linear storage-discharge relationship to capture the dynamics in monthly baseflow component.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (41890822; 51879193; 51861125102); the National Key Research and Development Program of China (2018YFC0407202; 2017YFC1502503); and the Overseas Expertise Introduction Project for Discipline Innovation (111 Project) funded by Ministry of Education and State Administration of Foreign Experts Affairs P.R. China (B18037). We thank Dr. Y. Zhang for providing the collation of streamflow data of Australian unregulated catchments.

Appendix A. Model description

A.1. Dynamic Water Balance Model (DWBM)

The DWBM was proposed by Zhang et al. (2008) based on the Budyko framework. The model structure is presented in Figure 2a. The DWBM conceptualizes a catchment as a system of two storages, *i.e.* soil water storage and groundwater storage. Rainfall in time step *t* is partitioned into quick flow (Q_d) and the sum of the other water balance components. The $Q_d(t)$ in the DWBM is calculated as:

$$Q_d(t) = P(t) - X(t) \tag{4}$$

where $X(t)$ is called catchment rainfall retention, calculated as $X(t) = P(t)F(\frac{PET(t)+S_{max}-S(t-1)}{P(t)}, a_1)$. The parameter S_{max} is soil water storage capacity. a_1 is a parameter that influences retention efficiency. The form of $F(\frac{PET(t)+S_{max}-S(t-1)}{P(t)}, a_1)$ is generalized from the equation $1 + \frac{PET(t)+S_{max}-S(t-1)}{P(t)} - [1 + (\frac{PET(t)+S_{max}-S(t-1)}{P(t)})^{a_1}]^{1/a_1}$, which is a classical Budyko framework proposed by (Fu, 1981). The form of *F*() used following is the same. Water availability of a catchment can be defined as $W(t) = X(t) + S(t-1)$. The $W(t)$ is the amount of rainfall retained in the catchment for actual evapotranspiration, soil moisture and groundwater recharge. Namely, $W(t) = E_a(t) + S(t) + R(t)$. The $E_a(t)$ is estimated as:

$$E_a(t) = W(t) \times F(\frac{PET(t)}{W(t)}, a_2) \tag{5}$$

where a_2 is a parameter that influences evapotranspiration efficiency. Groundwater recharge (R) is also generated from $W(t)$ and is calculated as:

$$R(t) = W(t) - Y(t) \tag{6}$$

where $Y(t)$ is called evapotranspiration opportunity, calculated as $Y(t) = W(t)F(\frac{PET(t)+S_{max}}{W(t)}, a_2)$. Groundwater discharge in the DWBM is treated as linear reservoir and $Q_b(t)$ is calculated as:

$$Q_b(t) = dG(t-1) \tag{7}$$

where the parameter d represents the baseflow generation efficiency. Groundwater balance can be modeled as $G(t) = (1-d)G(t-1) + R(t)$. In total, there are 4 parameters in the DWBM to be calibrated including S_{max}, a_1, a_2 and d. The unit of S_{max} is mm and the unit of d is month^{-1}.

A.2. Belgium Model (VUB)

The VUB was proposed by Vandewiele et al. (1992). The model structure is presented in Figure 2b. In this model, actual evapotranspiration (E_a) is computed as:

$$E_a(t) = \min\left[PET(t) \times \left(1 - {x_1}^{\frac{W(t)}{PET(t)}}\right), W(t)\right] \tag{8}$$

where the x_1 is a non-negative parameter which represents evapotranspiration resistance of the river basin; $W(t)$ is available water for E_a and is estimated as $W(t) = P(t) + S(t-1)$. Simulated monthly total flow of the VUB is the sum of quick flow (Q_d) and baseflow (Q_b). The $Q_d(t)$ is calculated as a function of soil moisture and effective precipitation as:

$$Q_d(t) = x_3 S(t-1) \times P_e(t) \quad (9)$$

$$P_e(t) = P(t) - PET(t) \times (1 - e^{\frac{-P(t)}{PET(t)}}) \quad (10)$$

where x_3 is a parameter, representing the fraction of precipitation that is immediately transformed into Q_d during the same rainfall event; $P_e(t)$ is the effective precipitation. The Q_b in month *t* is calculated as:

$$Q_b(t) = x_2 S(t-1) \quad (11)$$

where x_2 is a parameter, representing the fraction of stored soil water that is discharged as baseflow. In total, 3 parameters in the VUB are to be calibrated including x_1, x_2 and x_3. The unit of x_2 and x_3 is month^{-1}.

A.3. Time Variant Gain Model (TVGM)

The theory of the TVGM was first proposed by Xia et al. (1997) and then developed later by Xia et al. (2005). The model structure is presented in Figure 2c. As for the actual evapotranspiration (E_a), it can be expressed as a function of soil moisture and potential evapotranspiration as:

$$E_a(t) = PET(t) \times (S(t-1)/S_{max})^{\gamma} \quad (12)$$

where γ is a parameter, representing the nonlinear relationship between E_a and relative soil moisture. Quick flow (Q_d) in month *t* is calculated as:

$$Q_d(t) = g_1(S(t-1)/S_{max})^{g_2} \times P(t) \quad (13)$$

where S_{max} is saturated soil moisture; g_1 and g_2 are two empirical coefficients. As for the subsurface runoff generation model, the soil moisture at time step t is calculated by combining the water balance equation and the dynamic storage-discharge function. And the baseflow (Q_b) is calculated by a linear function of the soil moisture at time steps $t-1$ and t:

$$Q_b(t) = k_r\,(S(t-1) + S(t))/2 \quad (14)$$

where $S(t-1)$ and $S(t)$ are the soil moisture at time $t-1$ and t, respectively; k_r is an empirical coefficient related to baseflow generation. In total, there are 5 parameters in the TVGM to be calibrated including S_{max}, γ, g_1, g_2 and k_r. The unit of S_{max} is mm and the unit of k_r is month^{-1}.

A.4. WatBal Model (WM)

The WM was originally developed by Leaf and Brink (1973) and was further modified by Wang et al. (2014). The model structure is presented in Figure 2d. $E_a(t)$ is a function of potential evapotranspiration and the relative soil moisture and is estimated as:

$$E_a(t) = PET(t) \times S(t-1)/S_{max} \quad (15)$$

where $S(t-1)$ is the soil moisture storage at the beginning of time step t; S_{max} is the maximum storage capacity. $Q_d(t)$ is calculated as a function of relative storage and precipitation as:

$$Q_d(t) = k_s P(t) \times S(t-1)/S_{max} \quad (16)$$

where k_s the is quick flow coefficient. Q_b in month t is calculated with a linear storage-discharge function as:

$$Q_b(t) = k_g S(t-1) \quad (17)$$

where k_g is a parameter, representing the fraction of stored soil water that discharges as baseflow. In total, there are 3 parameters in the WM to be calibrated including S_{max}, k_s and k_g. The unit of S_{max} is mm and the unit of k_g is month^{-1}.

A.5. Schaake Model (SM)

The SM was firstly developed by Schaake and Liu (1989) and was improved later by Schaake (1990). The model structure is presented in Figure 2e. The uniqueness of the model is to introduce soil moisture deficit (*D*) for estimation of actual evapotranspiration (E_a) and runoff. In the SM, E_a is assumed to deplete the soil water at a potential rate when the storage deficit is zero, whereas E_a is zero when the storage deficit reaches the maximum. For the case storage deficit does not reach the maximum, E_a of month *t* is calculated as:

$$E_a(t) = PET(t) \times \frac{D_{max} - D(t)}{D_{max}} \quad (18)$$

where $D(t)$ is the soil water storage deficit at current time step, and D_{max} is the maximum deficit of soil moisture storage. Quick runoff (Q_d) is calculated as:

$$Q_d(t) = P_e(t)^2/(P_e(t) + D_{max}) \quad (19)$$

$$P_e(t) = P(t) - \theta E_a(t) - zD(t) \tag{20}$$

where $P_e(t)$ is effective precipitation; and θ and z are empirical parameters. Parameter θ represents the proportion of actual evapotranspiration that must be satisfied by current month precipitation before runoff can occur, and parameter z represents the proportion of infiltration that must be satisfied by current month precipitation before runoff can occur. Baseflow (Q_b) is assumed to vary with soil moisture deficit (*i.e.* D) and is calculated as:

$$Q_b(t) = k(G_{max} - D(t)) \tag{21}$$

where k is a parameter representing the proportion of surplus to generate baseflow, and G_{max} is the maximum groundwater storage. In total, there are 5 parameters in the SM to be calibrated including D_{max}, θ, z, k and G_{max}. The unit of D_{max} and G_{max} is mm and the unit of k is month^{-1}. The storage structure of SM is different with the other four. The SM uses only one soil moisture deficit (D) to represent both soil water and groundwater storages and uses two parameters (*i.e.* D_{max} and G_{max}) to represent the capacity of soil water and groundwater storages, respectively. Recharge from soil moisture to groundwater is not allowed in the SM. $Q_d(t)$ is calculated as the function of $-D(t)$, $E_a(t)$ is a function of ($D_{max}-D(t)$), and $Q_b(t)$ is a function of ($G_{max}-D(t)$). The number of water storage is regarded as only one (Jiang et al., 2007) as there is only one current status of moisture storage (*i.e.* D).

References

Ahiablame, L., Chaubey, I., Engel, B., Cherkauer, K. and Merwade, V., 2013. Estimation of annual baseflow at ungauged sites in indiana usa. J. Hydrol, 476: 13-27.

Aksoy, H. and Wittenberg, H., 2011. Nonlinear baseflow recession analysis in watersheds with intermittent streamflow. Hydrological Sciences Journal, 56(2): 226-237.

Alley, W.M., 1984. On the treatment of evapotranspiration, soil moisture accounting, and aquifer recharge in monthly water balance models. Water Resour Res, 20(8): 1137-1149.

Arnold, J.G., Allen, P.M., Muttiah, R. and Bernhardt, G., 1995. Automated base flow separation and recession analysis techniques. Groundwater, 33(6): 1010-1018.

Aulenbach, B.T. and Peters, N.E., 2018. Quantifying climate-related interactions in shallow and deep storage and evapotranspiration in a forested, seasonally water-limited watershed in the southeastern united states. Water Resour Res, 54(4): 3037-3061.

Bai, P., Liu, X. and Liu, C., 2018. Improving hydrological simulations by incorporating grace data for model calibration. J. Hydrol, 557: 291-304.

Bai, P., Liu, X., Liang, K. and Liu, C., 2015. Comparison of performance of twelve monthly water balance models in different climatic catchments of china. J. Hydrol, 529: 1030-1040.

Bastola, S., Murphy, C. and Sweeney, J., 2011. The role of hydrological modelling uncertainties in climate change impact assessments of irish river catchments. Adv Water Resour, 34(5): 562-576.

Beven, K., 1993. Prophecy, reality and uncertainty in distributed hydrological modelling. Adv Water Resour, 16(1): 41-51.

Beven, K.J., 2001. rainfall-runoff modelling - the primer. Wiley, Chichester, UK.

Beven, K.J. and Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology. Hydrological Sciences Bulletin, 24(1): 43-69.

Brutsaert, W. and Nieber, J.L., 1977. Regionalized drought flow hydrographs from a mature glacial plateau. Water Resour Res, 13(3): 637-643.

Buttle, J.M., 1994. Isotope hydrograph separations and rapid delivery of pre-event from drainage basins. Prog Phys Geog, 18(1): 16-41.

Chapman, T., 1999. A comparison of algorithms for stream flow recession and baseflow separation. Hydrol Process, 13(5): 701-714.

Chen, J., Brissette, F.P., Poulin, A. and Leconte, R., 2011. Overall uncertainty study of the hydrological impacts of climate change for a canadian watershed. Water Resour Res, 47: W12509.

Cheng, L., Yaeger, M., Viglione, A., Coopersmith, E., Ye, S. and Sivapalan, M., 2012. Exploring the physical controls of regional patterns of flow duration curves - part 1: insights from statistical analyses. Hydrol Earth Syst Sc, 16(11): 4435-4446.

Cheng, L., Zhang, L. and Brutsaert, W., 2016. Automated selection of pure base flows from regular daily streamflow data: objective algorithm. J. Hydrol Eng, 21(11): 06016008.

Dakhlaoui, H., Ruelland, D., Tramblay, Y. and Bargaoui, Z., 2017. Evaluating the robustness of conceptual rainfall-runoff models under climate variability in northern tunisia. J. Hydrol, 550: 201-217.

Deng, C., Liu, P., Wang, D. and Wang, W., 2018. Temporal variation and scaling of parameters for a monthly hydrologic model. J. Hydrol, 558: 290-300.

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G. and Gupta, H.V. et al., 2006. Model parameter estimation experiment (mopex): an overview of science strategy and major results from the second and third workshops. J. Hydrol, 320(1): 3-17.

Duffy, C.J., 1996. A two-state integral-balance model for soil moisture and groundwater dynamics in complex terrain. Water Resour Res, 32(8): 2421-2434.

Eckhardt, K., 2005. How to construct recursive digital filters for baseflow separation. Hydrol Process, 19(2): 507-515.

Euser, T., Winsemius, H.C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S. and Savenije, H.H.G., 2013. A framework to assess the realism of model structures using hydrological signatures. Hydrol. Earth Syst. Sci., 17(5): 1893-1912.

Fan, Y., Li, H. and Miguez-Macho, G., 2013. Global patterns of groundwater table depth. Science, 339(6122): 940.

Fenicia, F., Savenije, H.H.G., Matgen, P., Pfister, L. and Abebe, A., 2006. Is the groundwater reservoir linear? Learning from data in hydrological modelling. Hydrol. Earth Syst. Sci., 10(1): 139-150.

Fowler, K., Coxon, G., Freer, J., Peel, M., Wagener, T. and Western, A. et al., 2018b. Simulating runoff under changing climatic conditions: a framework for model improvement. Water Resour Res, 54(12): 9812-9832.

Fowler, K., Peel, M., Western, A. and Zhang, L., 2018a. Improved rainfall-runoff calibration for drying climate: choice of objective function. Water Resour Res, 54(5): 3392-3408.

Fu, 1981. On the calculation of the evaporation from land surface. SCIENTIA ATMOSPHERICA SINICA, 5(1): 23-31.

Gleick, P.H., 1987. The development and testing of a water balance model for climate impact assessment: modeling the sacramento basin. Water Resour Res, 23(6): 1049-1061.

Gonzales, A.L., Nonner, J., Heijkers, J. and Uhlenbrook, S., 2009. Comparison of different base flow separation methods in a lowland catchment. Hydrol. Earth Syst. Sci., 13: 2055-2068.

Grefenstette, J.J., Member and Ieee, 1986. Optimization of control parameters for genetic algorithms. IEEE Transactions on Systems, Man, and Cybernetics, 16(1): 122-128.

Gupta, H.V., Kling, H., Yilmaz, K.K. and Martinez, G.F., 2009. Decomposition of the mean squared error and nse performance criteria: implications for improving hydrological modelling. J. Hydrol, 377(1): 80-91.

Gupta, H.V., Wagener, T. and Liu, Y., 2008. Reconciling theory with observations: elements of a diagnostic approach to model evaluation. Hydrol Process, 22(18): 3802-3813.

Hamel, P., Guswa, A.J., Sahl, J., Zhang, L. and Abebe, A., 2017. Predicting dry-season flows with a monthly rainfall-runoff model: performance for gauged and ungauged catchments. Hydrol Process, 31(22): 3844-3858.

Harman, C. and Sivapalan, M., 2009. A similarity framework to assess controls on shallow subsurface flow dynamics in hillslopes. Water Resour Res, 45: W01417.

He, Z., Vorogushyn, S., Unger-Shayesteh, K., Gafurov, A., Kalashnikova, O. and Omorova, E. et al., 2018. The value of hydrograph partitioning curves for calibrating hydrological models in glacierized basins. Water Resour Res, 54(3): 2336-2361.

Horton, R.E., 1941. Virtual channel-inflow graphs. Eos Transactions American Geophysical Union, 22(3): 811-820.

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S. and Nijzink, R. et al., 2014. Process consistency in models: the importance of system signatures, expert knowledge, and process complexity. Water Resour Res, 50(9): 7445-7469.

Jiang, T., Chen, Y.D., Xu, C., Chen, X., Chen, X. and Singh, V.P., 2007. Comparison of hydrological impacts of climate change simulated by six hydrological models in the dongjiang basin, south china. J. Hydrol, 336(3-4): 316-333.

Kelleher, C., McGlynn, B. and Wagener, T., 2017. Characterizing and reducing equifinality by constraining a distributed catchment model with regional signatures, local observations, and process understanding. Hydrol Earth Syst Sc, 21(7): 3325-3352.

Khatami, S., Peel, M.C., Peterson, T.J. and Western, A.W., 2019. Equifinality and flux mapping: a new approach to model evaluation and process representation under uncertainty. Water Resour Res, 55(11): 8922-8941.

Kirchner, J.W., 2009. Catchments as simple dynamical systems: catchment characterization, rainfall-runoff modeling, and doing hydrology backward. Water Resour Res, 45: W02429.

Kottek, M., Grieser, J., Beck, C., Rudolf, B. and Rubel, F., 2006. World map of the köppen-geiger climate classification updated. Meteorol Z., 15(3): 259-263.

Krause, P., Boyle, D.P. and Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Advances in Geosciences, 5: 89-97.

Kumar, P., 2011. Typology of hydrologic predictability. Water Resour Res, 47: W00H05.

Larabi, S., St-Hilaire, A., Chebana, F. and Latraverse, M., 2018. Multi-criteria process-based calibration using functional data analysis to improve hydrological model realism. Water Resour Manag, 32(1): 195-211.

Leaf, C. and Brink, G., 1973. Computer simulation of snowmelt with a colorado subalpine watershed. For. Serv. Res. Pap., RM-99.

Lee, D., 2007. Testing a conceptual hillslope recession model based on the storage-discharge relationship with the richards equation. Hydrol Process, 21(23): 3155-3161.

Legates, D.R. and McCabe Jr., G.J., 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour Res, 35(1): 233-241.

Lindstrom, G., Johansson, B., Persson, M., Gardelin, M. and Bergstrom, S., 1997. Development and test of the distributed hbv-96 hydrological model. J. Hydrol, 201(1-4): 272-288.

Lyne, V. and Hollick, M., 1979. Stochastic time-variable rainfall-runoff modelling, Proceedings of the Hydrology and Water Resources Symposium. Institute of Engineers Australia National Conference, Perth.

Maneta, M.P., Soulsby, C., Kuppel, S. and Tetzlaff, D., 2018. Conceptualizing catchment storage dynamics and nonlinearities. Hydrol Process, 32: 3299-3303.

Markovic, D. and Koch, M., 2015. Stream response to precipitation variability: a spectral view based on analysis and modelling of hydrological cycle components. Hydrol Process, 29(7): 1806-1816.

McMillan, H., 2020. Linking hydrologic signatures to hydrologic processes: a review. Hydrol Process, 34: 1393-1409.

Moore, R.D., 1997. Storage-outflow modelling of streamflow recessions, with application to a shallow-soil forested catchment. J. Hydrol, 198(1): 260-270.

Murphy, A.H., 1988. Skill scores based on the mean square error and their relationships to the correlation coefficient. Mon Weather Rev, 116(12): 2417-2424.

Nash, J.E. and Sutcliffe, J.V., 1970. River flow forecasting through conceptual models, part i - a discussion of principles. J. Hydrol, 10: 282-290.

Nasseri, M., Zahraie, B., Ajami, N.K. and Solomatine, D.P., 2014. Monthly water balance modeling: probabilistic, possibilistic and hybrid methods for model combination and ensemble simulation. J. Hydrol, 511: 675-691.

Nathan, R.J. and McMahon, T.A., 1990. Evaluation of automated techniques for base flow and recession analyses. Water Resour Res, 26(7): 1465-1473.

Nippgen, F., McGlynn, B.L., Emanuel, R.E. and Vose, J.M., 2016. Watershed memory at the coweeta hydrologic laboratory: the effect of past precipitation and storage on hydrologic response. Water Resour Res, 52(3): 1673-1695.

Patnaik, S., Biswal, B., Nagesh Kumar, D. and Sivakumar, B., 2018. Regional variation of recession flow power-law exponent. Hydrol Process, 32(7): 866-872.

Peters, N.E. and Aulenbach, B.T., 2011. Water storage at the panola mountain research watershed, georgia, usa. Hydrol Process, 25: 3878-3889.

Poff, L.R., Allan, J.D., Bain, M.B., Karr, J.R., Prestegaard, K.L. and Richter, B.D. et al., 1997. The natural flow regime. Bioscience, 47(11): 769-784.

Rezaeianzadeh, M., Stein, A., Tabari, H., Abghari, H., Jalalkamali, N. and Hosseinipour, E.Z. et al., 2013. Assessment of a conceptual hydrological model and artificial neural networks for daily outflows forecasting. Int. J. Environ. Sci. Te., 10(6): 1181-1192.

Schaake, J., 1990. From climate to flow. In: waggoner, p.e. (Ed.), Cimate change and us water resourses, New York, 177-206 pp.

Schaake, J.C. and Liu, L.Z., 1989. Development and application of simple water balance models to understand the relationship between climate and water resources. In: kavvas, m.l. (Ed.), New directions for surface water modelling (proceedings of the baltimore symposium, may 1989), 181, 345-352 pp.

Schar, C., Vasilina, L., Pertziger, F. and Dirren, S., 2004. Seasonal runoff forecasting using precipitation from meteorological data assimilation systems. J. Hydrometeorol, 5(5): 959-973.

Schuite, J., Flipo, N., Massei, N., Rivière, A. and Baratelli, F., 2019. Improving the spectral analysis of hydrological signals to efficiently constrain watershed properties. Water Resour Res, 55: 4043-4065.

Shafii, M. and Tolson, B., 2015. Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives. Water Resour Res, 51: 3796-3814.

Shafii, M., Basu, N., Craig, J., L. Schiff, S. and Van Cappellen, P., 2017. A diagnostic approach to constraining flow partitioning in hydrologic models using a multiobjective optimization framework. Water Resour Res, 53: 3279-3301.

Shafii, M., Craig, J.R., Macrae, M.L., English, M.C., Schiff, S.L. and Van Cappellen, P. et al., 2019. Can improved flow partitioning in hydrologic models increase biogeochemical predictability? Water Resour Res, 55(4): 2939-2960.

Stoelzle, M., Weiler, M., Stahl, K., Morhard, A. and Schuetz, T., 2015. Is there a superior conceptual groundwater model structure for baseflow simulation? Hydrol Process, 29(6): 1301-1313.

Vandewiele, G.L., Xu, C. and Ni-Lar-Win, 1992. Methodology and comparative study of monthly water balance models in belgium, china and burma. J. Hydrol, 134(1): 315-347.

Wang, G.Q., Zhang, J.Y., Jin, J.L., Liu, Y.L., He, R.M. and Bao, Z.X. et al., 2014. Regional calibration of a water balance model for estimating stream flow in ungauged areas of the yellow river basin. Quatern Int, 336: 65-72.

Wang, Q.J., Pagano, T.C., Zhou, S.L., Hapuarachchi, H.A.P., Zhang, L. and Robertson, D.E., 2011. Monthly versus daily water balance models in simulating monthly runoff. J. Hydrol, 404(3-4): 166-175.

Westra, S., Thyer, M., Leonard, M., Kavetski, D. and Lambert, M., 2014. A strategy for diagnosing and interpreting hydrological model nonstationarity. Water Resour Res, 50: 5090-5113.

Wittenberg, H., 1999. Baseflow recession and recharge as nonlinear storage processes, 13, 715-726 pp.

Xia, J., Connor, K.M.O. and Kachroo, R.K., 1997. A non-linear perturbation model considering catchment wetness and its application in fiver flow forecasting. J. Hydrol, 200(1-4):

164-178.

Xia, J., Wang, G., Tan, G., Ye, A. and Huang, G.H., 2005. Development of distributed time-variant gain model for nonlinear hydrological systems. Science in China Series D: Earth Sciences, 48(6): 713-723.

Xiong, M., Liu, P., Cheng, L., Deng, C., Gui, Z. and Zhang, X. et al., 2019. Identifying time-varying hydrological model parameters to improve simulation efficiency by the ensemble kalman filter: a joint assimilation of streamflow and actual evapotranspiration. J. Hydrol, 568: 758-768.

Xu, C.Y. and Singh, V.P., 1998. A review on monthly water balance models for water resources investigations. Water Resour Manag, 12(1): 20-50.

Xu, C.Y., Seibert, J. and Halldin, S., 1996. Regional water balance modelling in the nopex area: development and application of monthly water balance models. J. Hydrol, 180(1): 211-236.

Yilmaz, K.K., Gupta, H.V. and Wagener, T., 2008. A process-based diagnostic approach to model evaluation: application to the nws distributed hydrologic model. Water Resour Res, 44: W09417.

Zhang, J., Zhang, Y., Song, J. and Cheng, L., 2017. Evaluating relative merits of four baseflow separation methods in eastern australia. J. Hydrol, 549: 252-263.

Zhang, L., Potter, N., Hickel, K., Zhang, Y. and Shao, Q., 2008. Water balance modeling over variable time scales based on the budyko framework - model development and testing. J. Hydrol, 360(1-4): 117-131.

Zhang, Y., Viney, N., Frost, A., Oke, A., Brooks, M. and Chen, Y. et al., 2013. Collation of australian modeller's streamflow dataset for 780 unregulated australian catchments, water for a healthy country national research flagship, CSIRO, Australia.

Figure

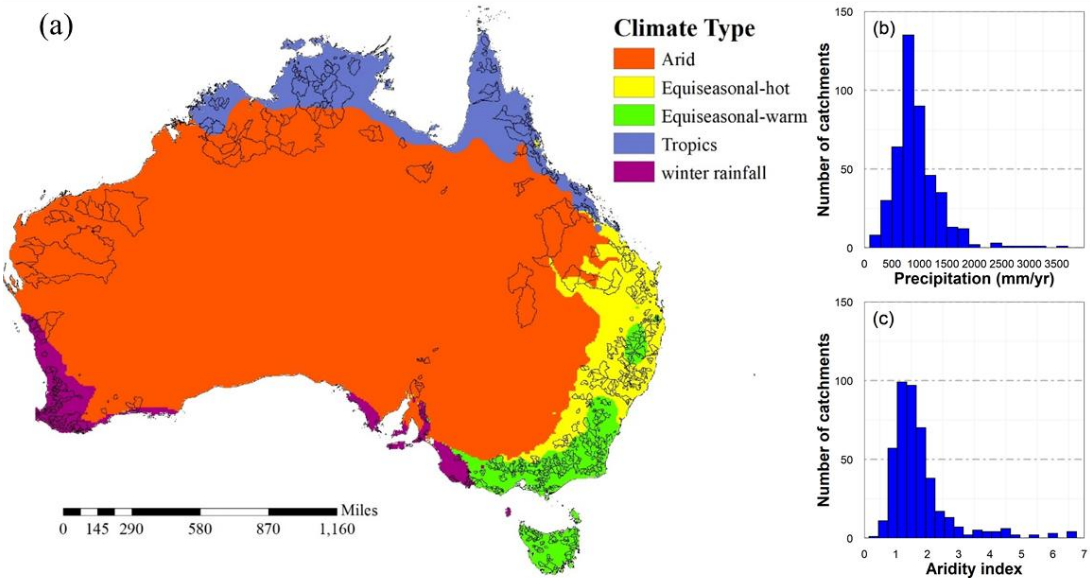

Figure 1. Spatial distribution and catchment characteristics of the 443 unregulated catchments used in this study. The background colour of subplot (a) shows different climatic types based on the Köppen-Geiger classification schemes. Subplots (b) and (c) show the frequency histograms of mean annual precipitation and aridity index, respectively.

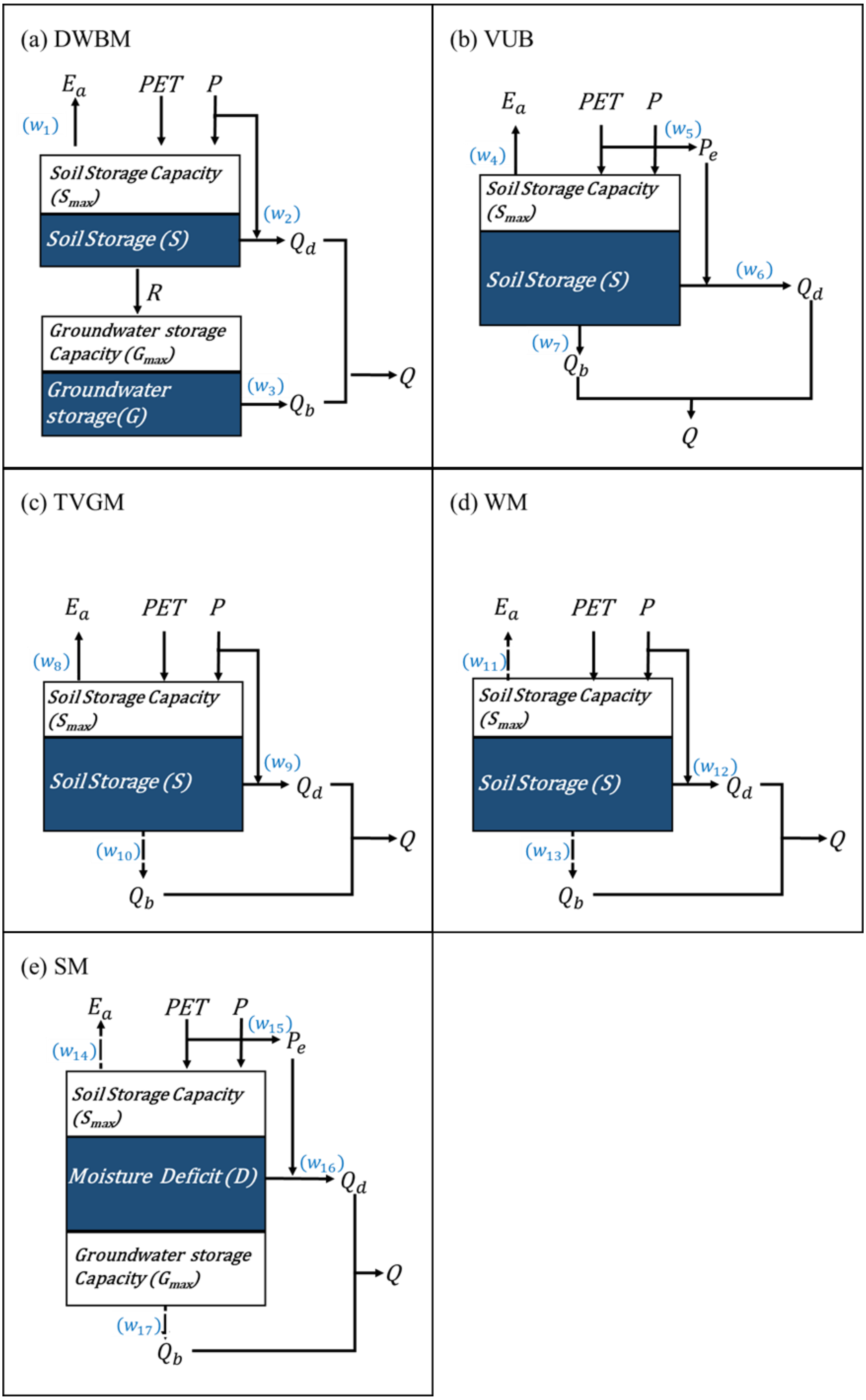

Figure 2. Conceptual representations of the 5 MWBMS with runoff partitioning structure. The meaning of the symbols refers to Table 2.

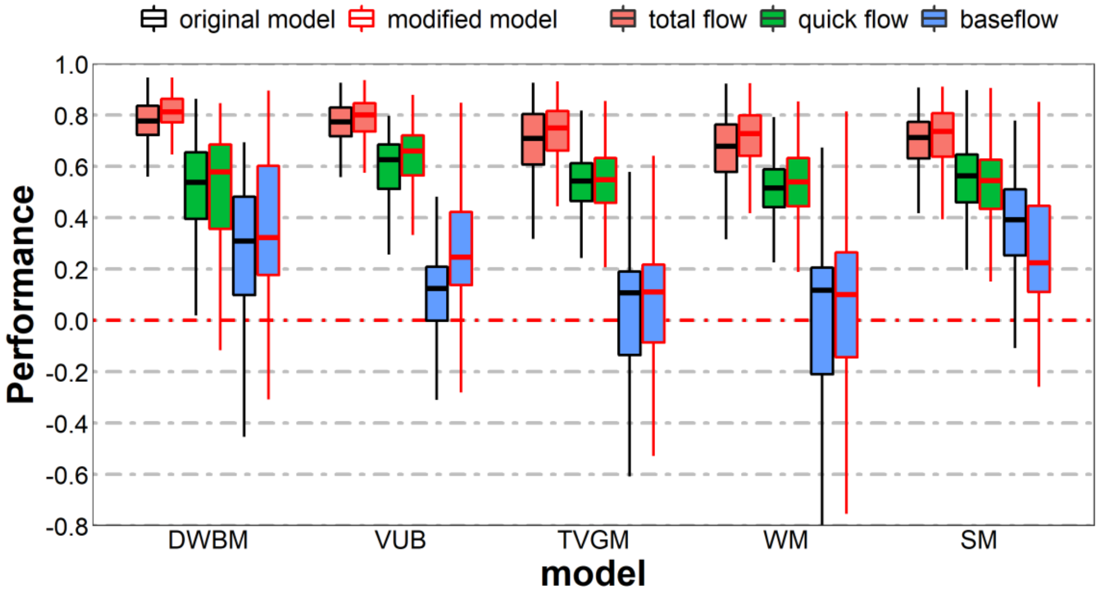

Figure 3. Boxplots showing the performance (value of F_{avg}) of the 5 MWBMs in their original (black line) and modified (red line) forms for estimating total flow (Q, red fill), quick flow (Q_d, green fill) and baseflow (Q_b, blue fill) in all the 443 catchments. Note that the minimum performance of the WM model is not included for a better visualization.

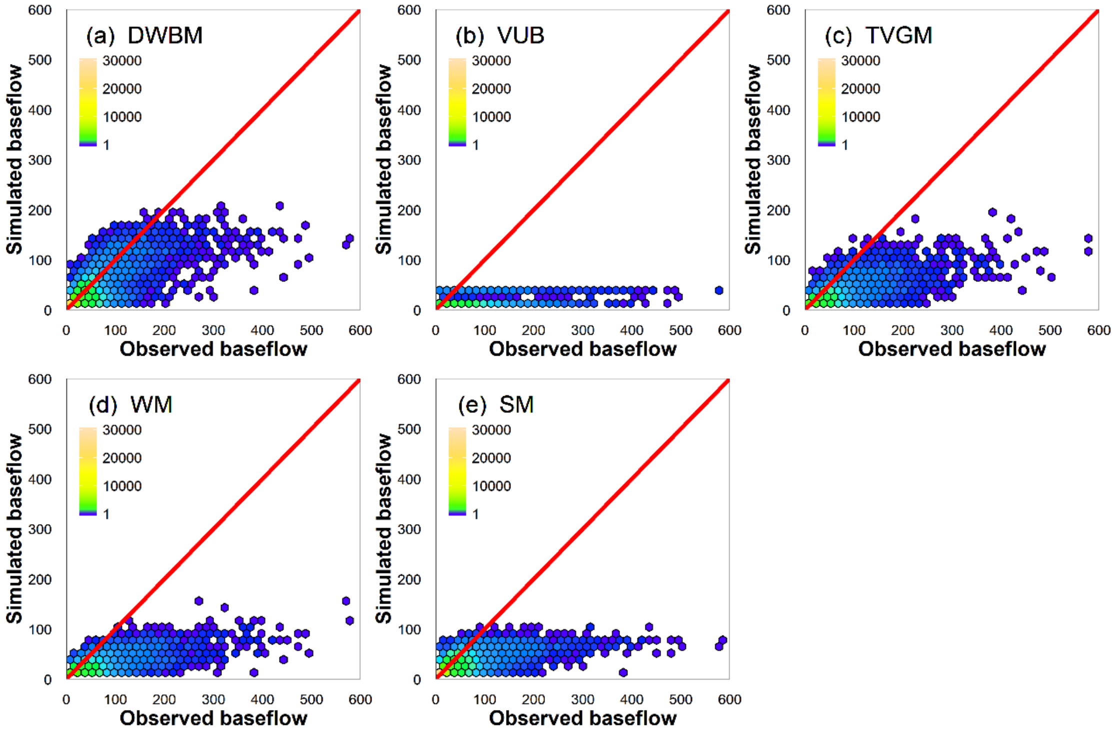

Figure 4. Hexagon binning plots showing comparison of observed and simulated monthly baseflow (mm month^{-1}) by 5 models in their original forms across all the 443 catchments over the period of 1975-2012. Subplots (a)~(e) are the results of DWBM, VUB, TVGM, WM and SM, respectively. The colour ramp of the hexagon in proportion to the counts indicates the density of data points.

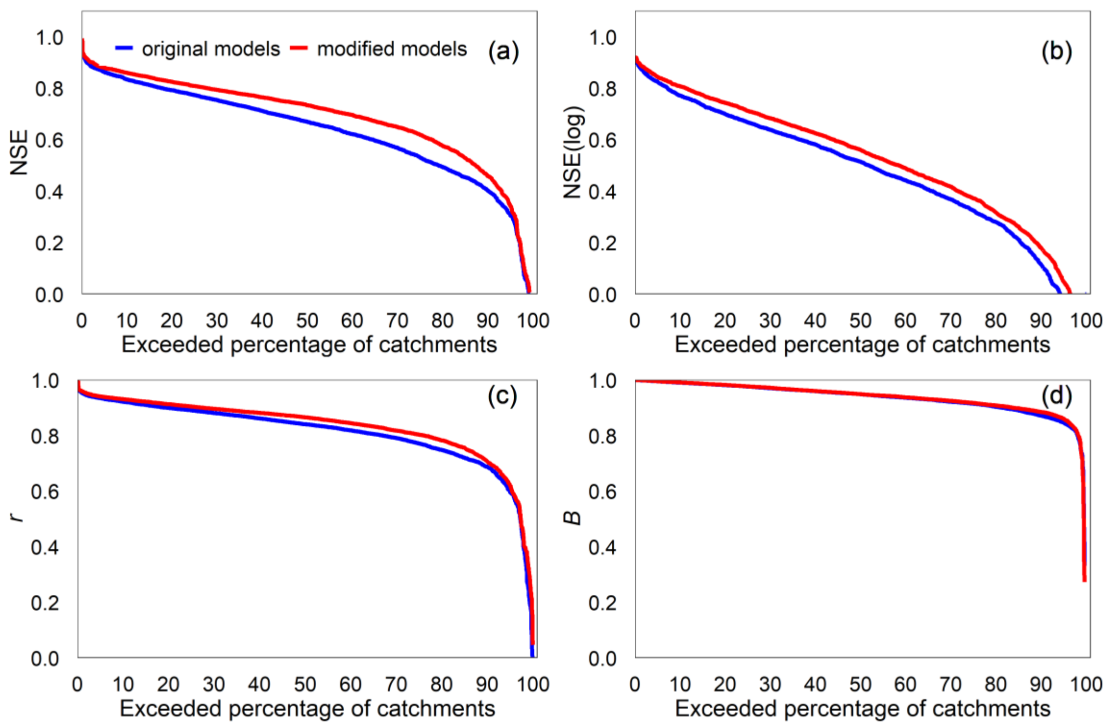

Figure 5. Comparison of total flow performance of the 5 original and modified MWBMs of all the 443 catchments. Subplots show exceeded percentage of catchments that (a) NSE, (b) NSE(log), (c) *r*, and (d) *B*.

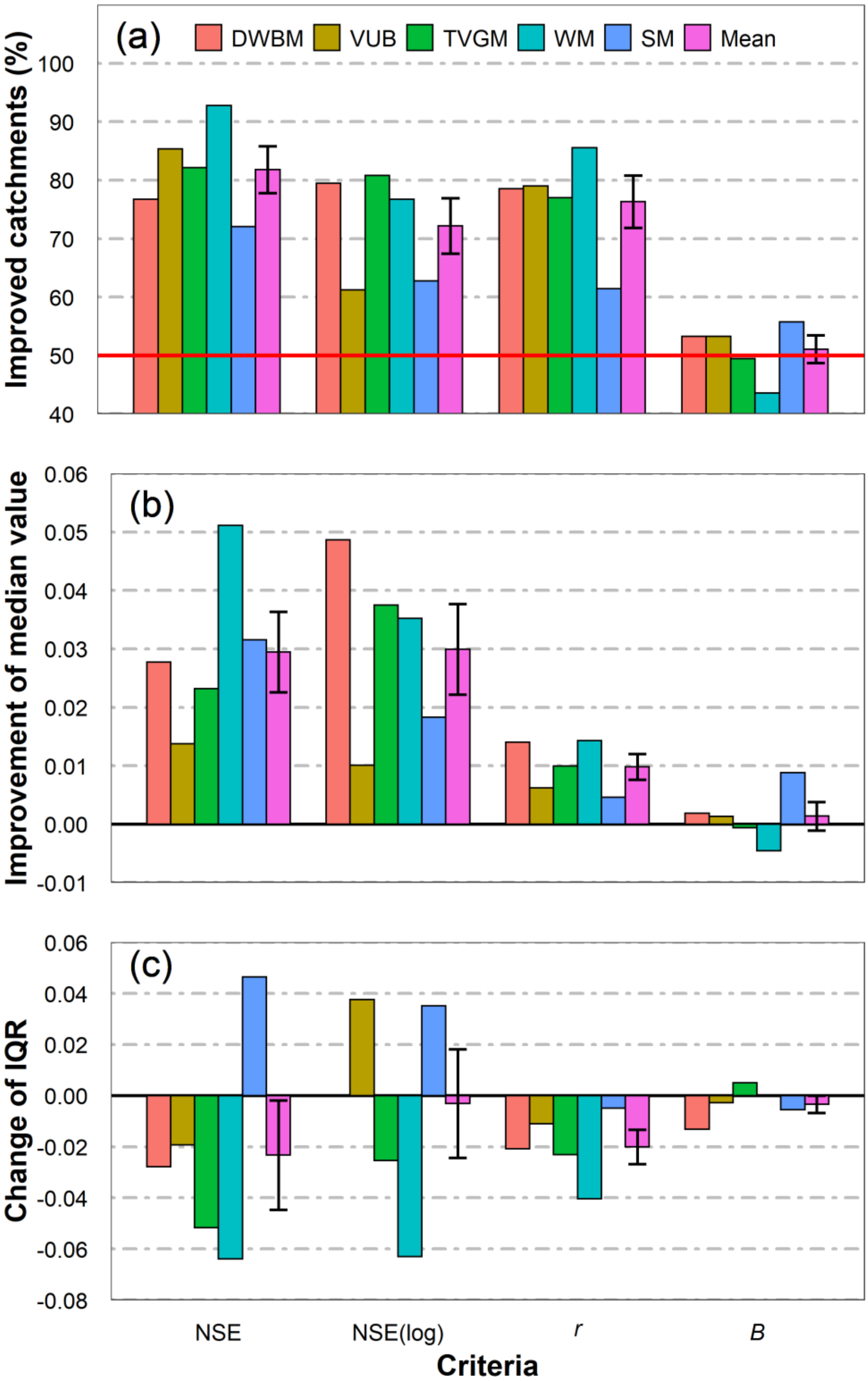

Figure 6. Comparison of total streamflow performance between original and modified models. (a) the percentage of improved catchments, (b) improvement of median value and, (c) change of IQR in terms of NSE, NSE (log), *r* and *B* of all the 443 catchments. The bar and error bar of the mean indicate mean and standard deviation of all the 5 models and all the 443 catchments.

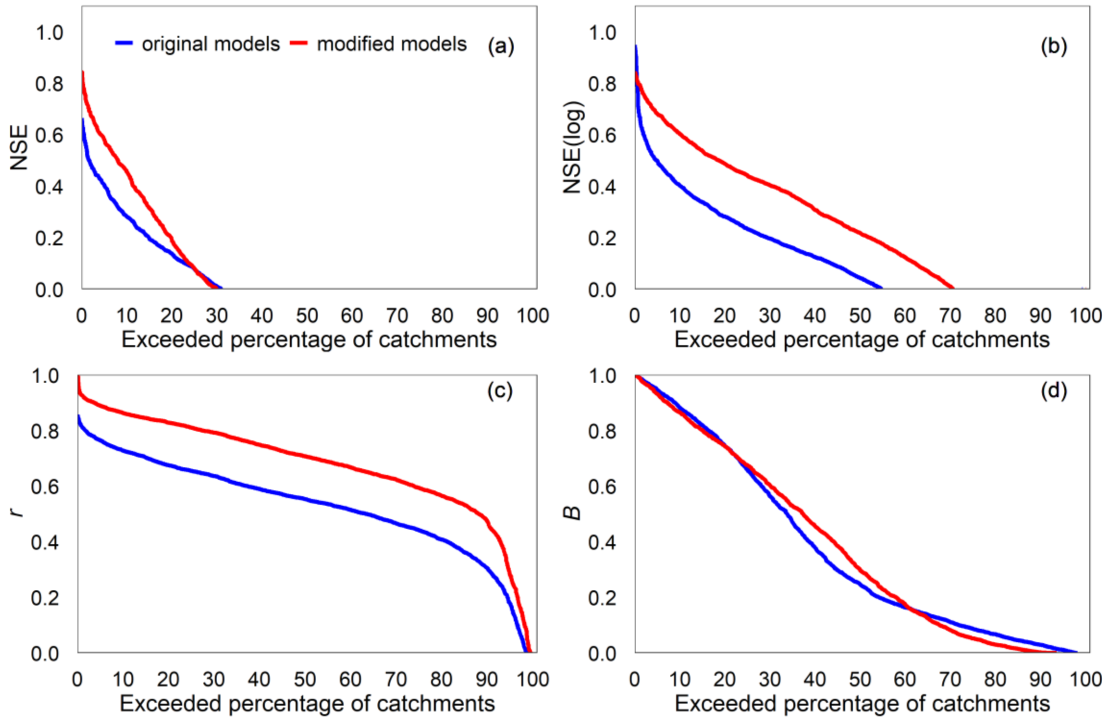

Figure 7. Same as Figure 5 except for baseflow.

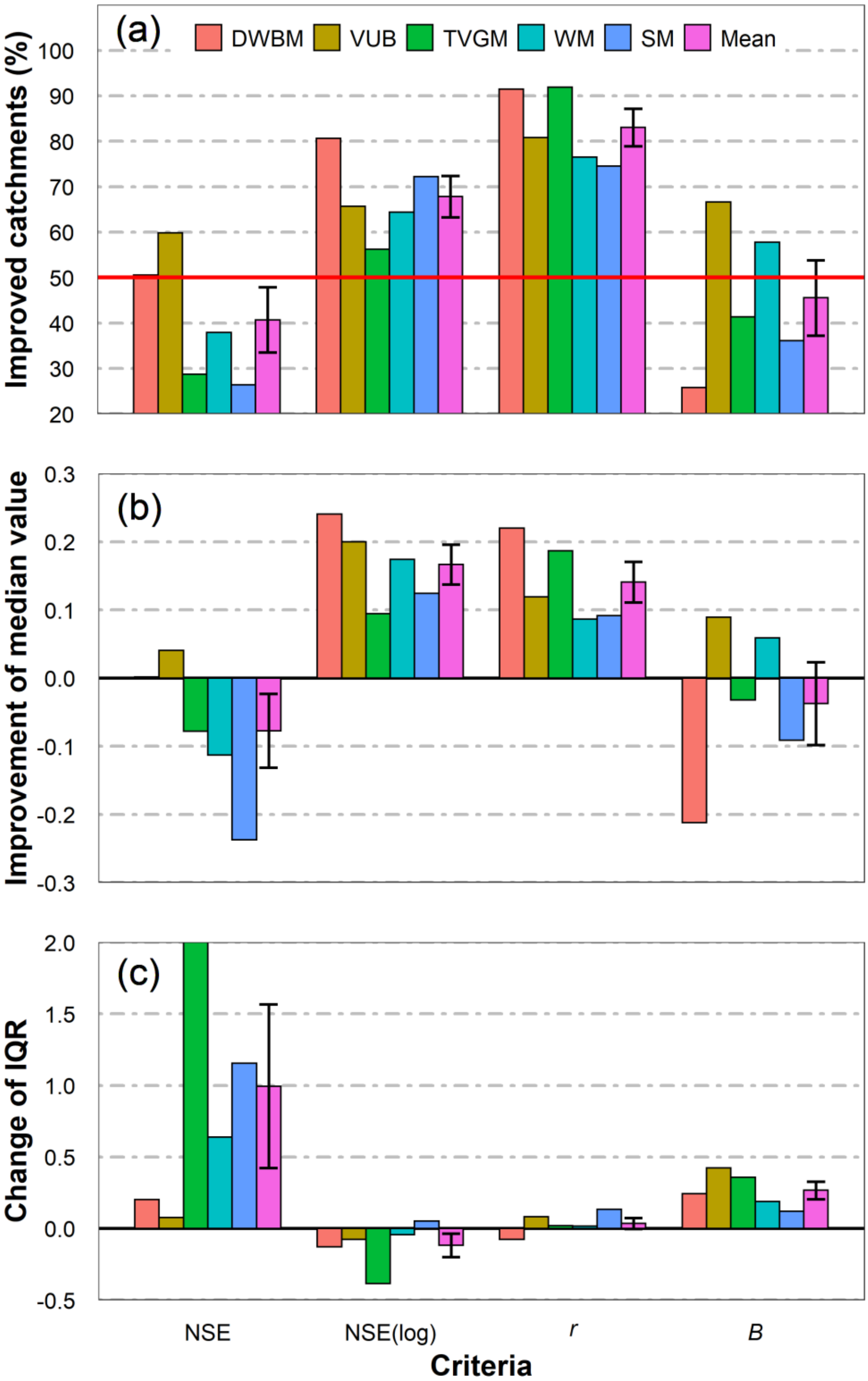

Figure 8. Same as Figure 6 except for baseflow. Note that maximum change of IQR of NSE of the WM model is 2.9 and the y-axis of subplot (c) is truncated to 2.0 for a better visualization.

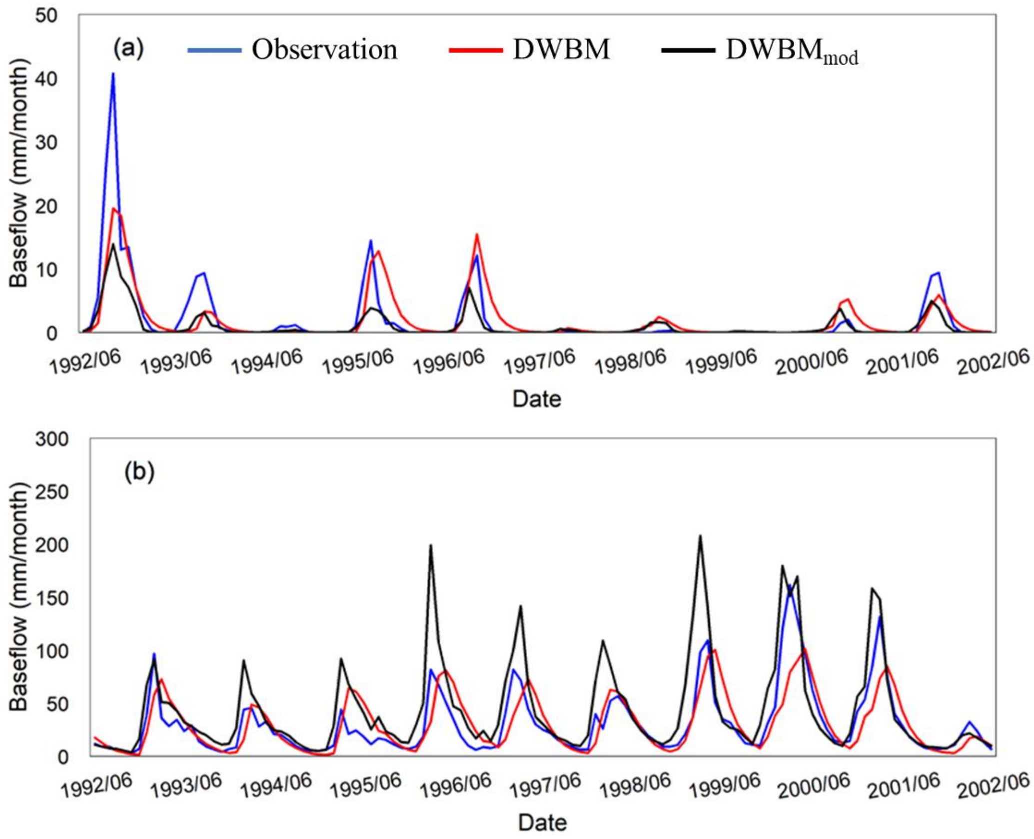

Figure 9. Time series of monthly baseflow (mm month^{-1}) from observation (blue line) and simulated by DWBM (red line) and DWBM$_{mod}$ (black line) in two selected catchments: (a) 238204 and (b) 108002. Note only ten-year records are showed for a better visualization.

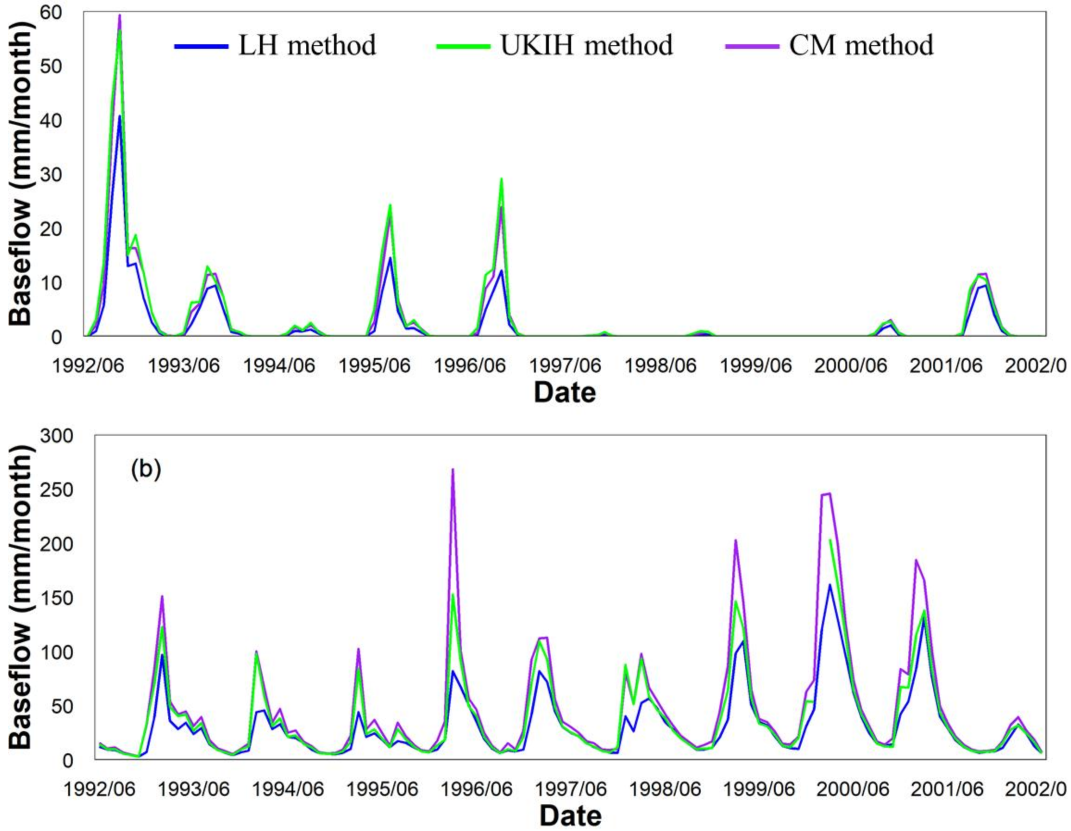

Figure 10. Comparison of baseflow derived from LH method (blue line), UKIH method (green line) and CM method (purple line) in catchments (a) 238204 and (b) 108002.

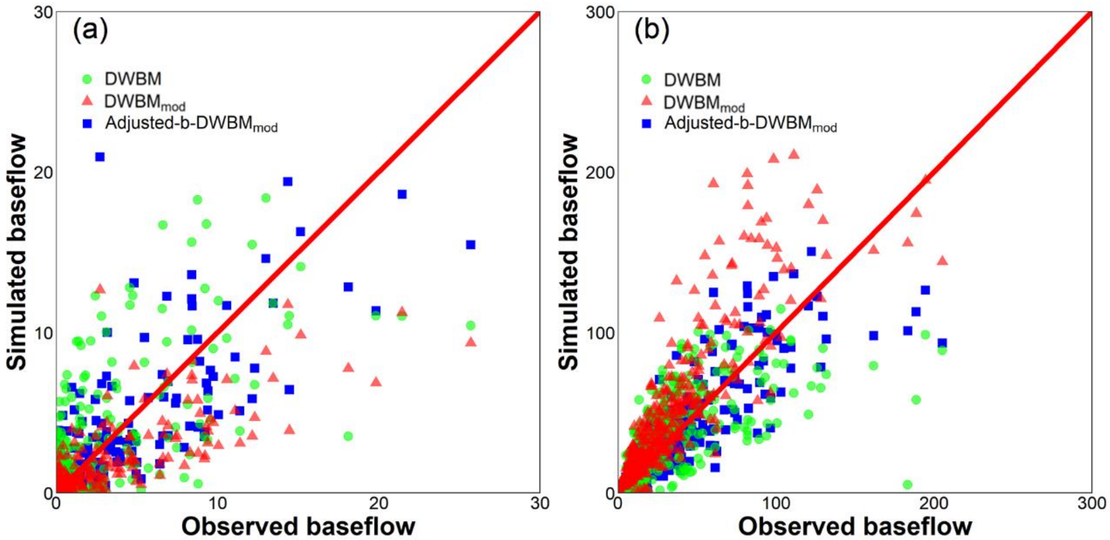

Figure 11. Scatter plots of observed and simulated monthly baseflow (mm month^{-1}) by DWBM (green dots), DWBM$_{mod}$ (red triangles) and Adjusted-*b*-DWBM$_{mod}$ (blue squares) in two selected catchments: (a) 238204 and (b) 108002.

Table 1. Summary of the catchment characteristics in the 443 catchments including tropics, arid, equiseasonal-hot, equiseasonal-warm and winter rainfall dominant.

Catchment characteristics	Total	Arid	Equiseasonal hot	Equiseasonal warm	Winter rainfall	Tropics
Number of catchments	443	50	105	171	61	56
Catchment area (km^2)	48-72902	65-72902	53-15851	51-16953	48-11795	66-47651
Mean annual rainfall (mm)	230-3684	230- 892	547-1791	491-2405	294-1129	760-3684
Mean annual potential evapotranspiration (mm)	921-2238	1214-1988	1190-1819	921-1495	1046-1553	1641-2238
Aridity index	0.39-6.99	2.21-6.99	0.76-2.69	0.39-2.31	1.14-5.28	0.48-2.49
Annual runoff coefficient	0.000-0.961	0.000-0.398	0.025-0.735	0.029-0.861	0.005-0.263	0.106-0.961
Annual baseflow index	0.001-0.792	0.001-0.027	0.032-0.509	0.033-0.792	0.062-0.799	0.061-0.605
CV of monthly precipitation	0.47-1.92	0.73-1.92	0.65-1.13	0.47-1.05	0.57-1.11	1.00-1.57
CV of monthly runoff	0.61-313.65	4.13-107.22	1.32-52.36	0.61-38.51	4.23-313.65	1.14-16.84

Table 2. Equations of the 5 models for simulating actual evaporation, quick flow and baseflow.

Model	Parameters	Equations to simulate actual evapotranspiration	No.	Equations to simulate quick flow	No.	Equations to simulate baseflow	No.
DWBM	S_{max}, a_1, a_2, d	$E_a(t) = W(t) \times F(\frac{PET(t)}{W(t)}, a_2)$	(w_1)	$Q_d(t) = P(t) \times (1 - F(\frac{X_0(t)}{P(t)}, a_1))$	(w_2)	$Q_b(t) = dG(t-1)$	(w_3)
VUB	x_1, x_2, x_3	$E_a(t) = \min\left[PET(t) \times \left(1 - x_1^{\frac{W(t)}{PET(t)}}\right), W(t)\right]$	(w_4)	$P_e(t) = P(t) - PET(t) \times (1 - e^{\frac{-P(t)}{PET(t)}})$ $Q_d(t) = x_3 S(t-1) \times P_e(t)$	(w_5) (w_6)	$Q_b(t) = x_2 S(t-1)$	(w_7)
TVGM	$g_1, g_2, k_r, S_{max}, \gamma$	$E_a(t) = PET(t) \times (S(\mathrm{t}-1)/S_{max})^{\gamma}$	(w_8)	$Q_d(t) = g_1 (S(t-1)/S_{max})^{g_2} \times P(t)$	(w_9)	$Q_b(t) = k_r\,(S(\mathrm{t}-1) + S(\mathrm{t}))/2$	(w_{10})
WM	S_{max}, k_s, k_g	$E_a(t) = PET(t) \times S(t-1)/S_{max}$	(w_{11})	$Q_d(t) = k_s (S(t-1)/S_{max}) \times P(t)$	(w_{12})	$Q_b(t) = k_g S(t-1)$	(w_{13})
SM	$D_{max}, G_{max}, k, z, \theta$	$E_a(t) = PET(t) \times \frac{D_{max} - D(\mathrm{t})}{D_{max}}$	(w_{14})	$P_e(t) = P(t) - \theta E_a(t) - zD(t)$ $Q_d(t) = P_e(t)^2/(P_e(t) + D_{max})$	(w_{15}) (w_{16})	$Q_b(t) = k(G_{max} - D(t))$	(w_{17})

Table 3. Summary of the linear or nonlinear characteristics of actual evapotranspiration, quick flow and baseflow simulating equations of the 5 MWBMs.

Model	Actual evapotranspiration		Quick flow		Baseflow	
	linear	nonlinear	linear	nonlinear	linear	nonlinear
DWBM		√		√	√	
VUB		√		√	√	
TVGM		√		√	√	
WM	√			√	√	
SM	√			√	√	

Table 4. The function for baseflow generation mechanism in the 5 original and modified models.

Original model	Equation for baseflow	Modified model	Equation for baseflow
DWBM	$Q_b(t) = dG(t-1)$	DWBM_{mod}	$Q_b(t) = e^{(W(t)-b)/m}$
VUB	$Q_b(t) = x_2 S(t-1)$	VUB_{mod}	$Q_b(t) = e^{(W(t)-b)/m}$
TVGM	$Q_b(t) = k_r\,(S(\mathrm{t}-1) + S(\mathrm{t}))/2$	TVGM_{mod}	$Q_b(t) = e^{((\mathrm{S}(t-1)+\mathrm{S}(t))/2-b)/m}$
WM	$Q_b(t) = k_g S(t-1)$	WM_{mod}	$Q_b(t) = e^{(W(t)-b)/m}$
SM	$Q_b(t) = k(G_{max} - D(t))$	SM_{mod}	$Q_b(t) = e^{(Dmax-D(t)+P(t)-b)/m}$
Note:	$W(t) = S(t-1) + P(t)$		

Table 5. The value of F_{avg} at 25th, 50th and 75th percentile across 443 catchments of total streamflow (Q_t), quick flow (Q_d) and baseflow (Q_b) simulated by five original models. The IQR is inter-quantile range (*i.e.* range between 75th and the 25th percentiles). The row of "Average" means the average value of F_{avg} of the five models.

Model	F_{avg} of total streamflow				F_{avg} of quick flow				F_{avg} of baseflow			
	25th	50th	75th	IQR	25th	50th	75th	IQR	25th	50th	75th	IQR
DWBM	0.72	0.77	0.84	0.12	0.41	0.56	0.68	0.27	0.09	0.30	0.48	0.39
VUB	0.72	0.77	0.83	0.11	0.52	0.63	0.69	0.17	0.00	0.12	0.21	0.21
TVGM	0.61	0.71	0.8	0.19	0.47	0.55	0.62	0.15	-0.14	0.11	0.19	0.33
WM	0.58	0.68	0.76	0.18	0.44	0.52	0.6	0.16	-0.21	0.12	0.21	0.42
SM	0.63	0.71	0.77	0.14	0.47	0.57	0.66	0.19	0.24	0.39	0.51	0.27
Average	0.65	0.73	0.80	0.15	0.46	0.57	0.65	0.19	0.00	0.21	0.32	0.32

Table 6. Summary of the improved values of different indicators for the total streamflow performance comparing the modified and original models. The last row shows the average (mean ± standard deviation) of all the 5 models.

Model	NSE			NSE(log)			*r*			*B*		
	Proportion (%)	Median	IQR	Proportion (%)	Median	IQR	Proportion (%)	Median	IQR	Proportion (%)	Median	IQR
DWBM	76.75	0.03	−0.06	79.46	0.05	−0.06	78.56	0.01	−0.04	53.27	0.00	0.00
VUB	85.33	0.01	0.05	61.17	0.01	0.04	79.01	0.01	0.00	53.27	0.00	−0.01
TVGM	82.17	0.02	−0.02	80.81	0.04	0.04	76.98	0.01	−0.01	49.44	0.00	0.00
WM	92.78	0.05	−0.03	76.75	0.04	0.00	85.55	0.01	−0.02	43.57	0.00	−0.01
SM	72.01	0.03	−0.05	62.75	0.02	−0.03	61.40	0.00	−0.02	55.76	0.01	0.00
Range	72.01~92.78	0.01~0.05	−0.06~0.05	61.17~80.81	0.01~0.05	−0.06~0.04	61.40~85.55	0.00~0.01	−0.04~0.00	43.57~55.76	0.00~0.01	−0.01~0.00
Average	81.81±3.99	0.03±0.007	−0.02±0.02	72.19±4.73	0.03±0.008	−0.002±0.02	76.30±4.48	0.01±0.002	−0.02±0.01	51.06±2.38	0.002±0.002	−0.004±0.003

Table 7. Same as Table 6 except for baseflow.

Model	NSE			NSE (log)			*r*			*B*		
	Proportion (%)	Median	IQR	Proportion (%)	Median	IQR	Proportion (%)	Median	IQR	Proportion (%)	Median	IQR
DWBM	50.56	0.00	0.64	80.59	0.24	−0.05	91.42	0.22	0.02	25.73	−0.21	0.19
TVGM	28.67	−0.08	0.08	56.21	0.09	−0.08	91.87	0.19	0.08	41.31	−0.03	0.42
VUB	59.82	0.04	1.15	65.69	0.20	0.05	80.81	0.12	0.13	66.59	0.09	0.12
WM	37.92	−0.11	0.20	64.33	0.17	−0.13	76.52	0.09	−0.08	57.79	0.06	0.24
SM	26.41	−0.24	2.90	72.23	0.12	−0.38	74.49	0.09	0.02	36.12	−0.09	0.36
Range	26.41~59.82	−0.24~0.04	0.08~2.90	56.21~80.59	0.09~0.24	−0.38~0.05	74.49~91.42	0.09~0.22	−0.08~0.13	25.73~66.59	−0.21~0.09	0.12~0.42
Average	40.68±7.16	−0.08±0.05	0.99±0.57	67.81±4.57	0.17±0.03	−0.12±0.08	83.02±4.10	0.14±0.03	0.03±0.04	45.51±8.26	−0.04±0.06	0.27±0.06

Table 8. Summary of model parameters and performances of the DWBM, $DWBM_{mod}$ and Adjusted-*b*-$DWBM_{mod}$ in catchment 238204 and 108002.

Station	Model	Parameters		Criteria			
		m	*b*	*r*	NSE	NSE(log)	*B*
	DWBM	/	/	0.713	0.471	0.126	0.811
238204	$DWBM_{mod}$	47.5	233.9	0.840	0.572	0.343	0.612
	Adjusted-*b*- $DWBM_{mod}$	47.5	210	0.840	0.705	0.491	0.988
	DWBM	/	/	0.698	0.468	0.259	0.998
108002	$DWBM_{mod}$	226.8	1.73	0.872	0.307	0.657	0.427
	Adjusted-*b*- $DWBM_{mod}$	226.8	100	0.872	0.753	0.772	0.925