Comparison of performance of twelve monthly water balance models in different climatic catchments of China

Peng Bai, Xiaomang Liu*, Kang Liang, Changming Liu

Key Laboratory of Water Cycle & Related Land Surface Processes, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

A R T I C L E   I N F O

Article history:
Received 13 September 2014
Received in revised form 18 August 2015
Accepted 5 September 2015
Available online 15 September 2015
This manuscript was handled by Konstantine P. Georgakakos, Editor-in-Chief, with the assistance of Emmanouil N. Anagnostou, Associate Editor

Keywords:
Monthly water balance model
Model comparison
Model selection

A B S T R A C T

Multi-model comparison can provide useful information for model selection and improvement. In this study, twelve monthly water balance models with different structures and various degree of complexity are compared in 153 catchments with different climatic conditions in China. Generally, the GR5M model has the best performance, followed by the GR2M and WBM model. We investigate the relations between model performance and catchment characteristics and find that the climatic characteristic of a catchment is the most important factor impacting model performance. The models have better performance in wet catchments than in dry catchments. Large differences of model performance exist in dry catchments and model users should pay attention to model selection in dry catchments. In addition, we analyze the model performances among different models and conclude that increasing the model complexity does not guarantee a better model performance. Simple models can achieve comparable or even better performance than complex models. For the monthly simulation of hydrological processes, a two-parameter model is sufficient to achieve a good result. Moreover, by comparing the impacts of evapotranspiration simulation and runoff generation simulation on model performance, we find that evapotranspiration simulation has limited influence on the model performance. We suggest model builders focus on runoff generation process rather than evapotranspiration process to improve the performance of a monthly hydrological model.

1. Introduction

Hydrological models use different mathematical formulas to conceptualize processes of hydrologic cycle and are commonly used for simulating and predicting various hydrological processes (Vrugt et al., 2005; Viney et al., 2009). Currently, numerous hydrological models have been developed for different time scales (Mouelhi et al., 2006 monthly and yearly). Among them, the monthly water balance model (MWBM) offers simple yet refined methods to describe hydrological processes and has low input requirement, well-behaving conceptual platform and simple model calibration (Nasseri et al., 2014). For most of the MWBMs, runoff can be simulated using only monthly precipitation and potential evapotranspiration, and the number of model parameters ranges from two to five. Hence, these models are more parsimonious than daily or hourly models for estimating runoff at monthly or yearly time scales and are widely used for various purposes, e.g., seasonal streamflow forecasting (Alley, 1985; Schär et al., 2004), climate change and/or human activity impact assessment (Gleick, 1987; Jiang et al., 2007; Li et al., 2012; Liu et al., 2013) and snow-melt runoff simulation (Xu et al., 1996; Racoviteanu et al., 2013).

The first MWBM was developed in the 1940s by Thornthwaite (1948) and was subsequently revised by Thornthwaite and Mather (1955). Thereafter, different MWBMs were developed based on the framework of the Thornthwaite model. In 1965, Palmer (1965) proposed a two-layer soil moisture storage model based on a meteorological drought index. This model assumes that soil moisture in the lower layer cannot move to the upper layer until all of the available soil moisture in the upper layer has been exhausted. In 1973, Pitman (1973) developed a MWBM with twelve parameters to describe the hydrological processes in South Africa. Since then, new model functions such as reservoir sub-model, wetland sub-model and groundwater recharge sub-model have been successively added to Pitman model (Hughes, 2004; Hughes et al., 2013). This model including more than 20 parameters is likely the most complex model among the existing MWBMs (Hughes, 2013). In 1981, Thomas (1981) proposed a four-parameter “abcd” water balance model based on Thornthwaite’s (1948) conceptual framework, while this model incorporates a more realistic representation of the infiltration process (Martinez and Gupta, 2010). As the concern regarding

* Corresponding author. Tel.: +86 10 64889083.
E-mail addresses: lixiuxfmx@163.com, liuxm@igsnrr.ac.cn (X. Liu).

http://dx.doi.org/10.1016/j.jhydrol.2015.09.015
0022-1694/© 2015 Elsevier B.V. All rights reserved.
climate change began to increase in the 1990s, additional MWBMs were developed for evaluating the impacts of climate change on hydrological processes. During this period, some representative models were successively developed, such as the Belgium model (Vandewiele and Xu, 1992), GR2M model (Makhlof and Michel, 1994), MWB-6 model (Xu et al., 1996), Xiong model (Xiong and Guo, 1999) and DWBM model (Zhang et al., 2008). Besides the conceptual MWBMs, the artificial intelligence methods are also important tools to simulate monthly rainfall–runoff processes (Komorník et al., 2006; Shu and Ouarda, 2008; Wang et al., 2009; Yilmaz et al., 2011). In some studies, artificial intelligence models exhibited better performances than conceptual MWBMs (Hsu et al., 1995; Shamseldin, 1999; Machado et al., 2011; Rezaeianzadeh et al., 2013). However, artificial intelligence models have also been criticized for their lack of explanation capability, over-parameterization and over-fitting (Kaastra and Boyd, 1996; Gaume and Gosset, 2003; de Vos and Rientjes, 2005).

With the existence of numerous hydrological models, model users may require help to select a suitable model for a specific hydrological practice. To provide scientific guidance on the application of hydrological models, several model comparisons have been conducted with various types of hydrological models, such as flood forecasting models (WMO, 1975; Toth et al., 2000; Chau et al., 2005), snowmelt runoff models (WMO, 1986; Gurtz et al., 2003; Kumar Pokhrel et al., 2014) and distributed hydrological models (Yang et al., 2000; Reed et al., 2004; Smith et al., 2012). These studies mainly focus on the hourly and daily hydrological models. For the comparison of monthly hydrological models, Vandewiele and Xu (1992) compared a set of MWBMs in 79 catchments with areas less than 4000 km² and found that their new proposed models presented better performance than the existing models; Makhlof and Michel (1994) compared a two-parameter MWBM with four widely used models in 91 French catchments with area between 315 and 5560 km² and concluded that the simple two-parameter model has comparable performance with the four models; Jiang et al. (2007) applied six MWBMs in a humid catchment of China and found that all the models have similar performance in spite of a wide range of model complexity.

Here, we intend to extend these previous comparative studies by testing twelve MWBMs on a large set of 153 catchments in China with different climatic conditions. The main objective of this study is to investigate differences in model performance and provide valuable information for model selection and improvement. The paper is structured as follows. Section 2 briefly describes the data used and parameter calibration. Section 3 presents the data used and the methodology, followed by results and discussion in Section 4. Finally, the main findings are summarized in Section 4.

2. Models used and parameter calibration

2.1. Model descriptions

For a catchment, the general water balance equation at the monthly time scale can be written as:

\[ P(t) = S(t + 1) - S(t) + E_a(t) + R(t) + I_{deep}(t) - \Delta O(t) \quad (1) \]

where \( S(t) \) and \( S(t + 1) \) represent the soil moisture storage at the beginning and end of the time interval \( t \), respectively. \( P \) represents the precipitation, \( E_a \) represents the actual evapotranspiration and \( R \) represents the runoff at the outlet of the watershed. \( I_{deep} \) is the infiltration loss to deep aquifier and \( \Delta O \) is the water recharge from neighboring basins. Among these variables, \( S \), \( E_a \) and \( R \) are the three basic variables included in most of the MWBMs (Xu and Singh, 2004; Jiang et al., 2007). \( I_{deep} \) and \( \Delta O \) are rarely considered in MWBMs, with the exception of the SFB3 model considering \( I_{deep} \) (Boughton, 1984) and the GR2M mode considering \( \Delta O \) (Mouelhi et al., 2006).

This water balance equation describes the storage, transformation and movement of water at watershed scale with simple concepts. Generally, complex models are inclined to employ more storage or nonlinear formulas to describe these hydrological processes and have more model parameters. For example, the Pitman model (Pitman, 1973) including three types of storage (canopy, soil moisture and groundwater) and four nonlinear formulas has more than 20 parameters. The calibration of parameters becomes more difficult as the number of parameters increases. However, an inadequate complexity often results in over-parameterization (Ye et al., 1997a; Perrin et al., 2001). Therefore, the models with too many parameters (e.g., Pitman model) are excluded from this comparative study. Through an extensive literature review, twelve MWBMs are selected for the model comparison (Table 1). These models cover a relatively wide range of complexities with the parameter number ranging from two to five. The detailed structural characteristics of the selected models are summarized in Fig. 1. The main expressions for estimating the actual evapotranspiration and runoff are summarized in Table 2.

Although the selected models have a similar conceptual framework to describe the hydrological processes, the main mathematical equations simulating the hydrological processes are different. Among the twelve models, six models have single moisture storage, the others have two moisture storage. Moreover, complex models consider more runoff components than simple models. The models with more than three parameters have at least two runoff components: surface runoff and groundwater runoff, while all the two-parameter models consider runoff as a single component (Fig. 1).

The actual evapotranspiration is controlled by both water and energy availabilities. In general, the soil moisture storage and potential evapotranspiration are the most commonly used water-limited and energy-limited conditions for monthly hydrological models, respectively. In all of the selected models, except for XM, the actual evapotranspiration is estimated as a function of the potential evapotranspiration and soil moisture storage. However, obvious differences can be identified in the calculations of actual evapotranspiration (Table 2). Some models (e.g., the SFB3, WM and SM model) adopt a simple linear function to calculate the actual evapotranspiration, while the other models (e.g., the

| Model abbreviation | Derived from | No. of parameters | No. of storages | No. of runoff components |
|---------------------|--------------|-------------------|----------------|--------------------------|
| TM                  | Thornthwaite and Mather (1955) | 2 | 2 | 1 |
| XM                  | Xiong and Guo (1999) | 2 | 1 | 1 |
| GR2M                | Mouelhi et al. (2006) | 2 | 2 | 1 |
| VUB                 | Vandewiele and Xu (1992) | 3 | 1 | 2 |
| SFB3                | Boughton (1984) | 3 | 2 | 2 |
| WM                  | Wang et al. (2013) | 3 | 1 | 2 |
| DWBM                | Zhang et al. (2008) | 4 | 1 | 2 |
| abcD                | Thomas (1984) | 4 | 2 | 2 |
| WBM                 | Leaf et al. (1973) | 4 | 1 | 4 |
| GRSM                | Mouelhi et al. (2006) | 5 | 2 | 2 |
| SM                  | Schaake and Liu (1988) | 5 | 1 | 2 |
| TVCM                | Xia et al. (1997) and Wang et al. (2009a) | 5 | 1 | 2 |
GR2M, DWBM and abcd model) involve relatively complex nonlinear equations to calculate the actual evapotranspiration. In the XM model, precipitation is used instead of soil moisture storage as the water-limiting condition to restrict the actual evapotranspiration. Significant differences in the simulation of runoff generation exist among the models (Table 2). Some models (e.g., the TM, XM and GR2M model) hypothesize that runoff is directly derived from water storage instead of precipitation. Precipitation is used to satisfy evapotranspiration demand and replenish the water storage. In other models (e.g., the WM, DWBM and SM model), the surface (fast) runoff is treated as a portion of precipitation, and the soil moisture content is the adjustor to determine the portion. The groundwater flow (slow runoff or base flow) in most of the models is considered a linear function of storage (Table 2).

2.2. Model assessment methodology

Based on the Split Sample Test method proposed by Klemeš (1986), the available data for each catchment is split into two sub-periods with an equal length, and each of the two sub-periods is used in turn for calibration and validation. The first three years of each sub-period is used for model warm-up. The King–Gupta Efficiency (KGE) (Gupta et al., 2009) is employed as an objective function to calibrate the model parameters in this study. It has been demonstrated that KGE provides better model calibration than the widely used Nash–Sutcliffe Efficiency (Martinec and Rango, 1989; Pechlivanidis et al., 2010; Pechlivanidis et al., 2014). The function of KGE incorporates three basic assessment criteria: linear correlation (γ), variability (α) and bias ratio (β), and is defined as:

![Fig. 1. Conceptual representations for twelve monthly water balance models. A list of symbols is given in the notation section.](image-url)
KGE = 1 - \sqrt{\left(1 - r^2\right)^2 + (1 - x)^2 + (1 - \beta)^2} \tag{2}

The value of KGE ranges from negative infinity (poor model) to 1 (perfect model). The Particle Swarm Optimization (PSO) method is used for optimizing the model parameters. The PSO method proposed by Eberhart and Kennedy (1995) has been proven to be an effective calibration method and is widely applied in parameter optimization of hydrological models (Gill et al., 2006; Zhang and Chiew, 2009; Luo et al., 2012).

The average Nash–Sutcliffe Efficiency calculated on root squared streamflows (NSE_{sq}) (Chiew and McMahon, 1994) and Water Balance Error (WBE) in two validation sub-periods are used as model assessment criteria:

\[
\text{NSE}_{\text{sq}} = 1 - \frac{\sum_{i=1}^{n} \left( Q_{\text{sim},i} - Q_{\text{obs},i} \right)^2}{\sum_{i=1}^{n} \left( Q_{\text{obs},i} - \bar{Q}_{\text{obs}} \right)^2} \tag{3}
\]

\[
\text{WBE} = 1 - \frac{\sum_{i=1}^{n} Q_{\text{sim},i}}{\sum_{i=1}^{n} \left| Q_{\text{obs},i} \right|} \tag{4}
\]

where \( Q_{\text{obs}} \) and \( Q_{\text{sim}} \) are the observed and simulated streamflow, respectively. \( \sqrt{Q_{\text{obs}}} \) is the square root of observed streamflow. NSE_{sq} emphasizes the overall agreement between observed and simulated streamflow, and the optimal value of NSE_{sq} is 1. The WBE ranges from negative infinity to positive infinity, and the zero value means perfect agreement.

### 2.3. Test catchments and data source

Previous studies indicate that catchment characteristics can strongly affect model performance (Martinec and Rango, 1989; Perrin et al., 2008; Vaze et al., 2010; Merz et al., 2011; Coron et al., 2012). To evaluate the impacts of catchment characteristics on model performance, 153 unimpaired catchments (with limited anthropogenic influence) in China are selected to compare model performances in this study (Fig. 2). The catchments cover a broad range of climatic conditions with mean annual precipitation ranging from 123 to 2411 mm, runoff coefficient from 0.02 to 0.89 and aridity index (\( P/\text{PET} \)) from 0.40 to 1.96 (Table 3). The catchments were selected to have limited snow influence, since the models were tested without snowmelt module.

To investigate the influences of catchment characteristics on model performance, the catchments are divided into different types based on the catchment area, aridity index and the coefficient of variation (CV) of monthly runoff series (Table 4). The CV is defined as the ratio of the standard deviation \( \sigma \) to the mean \( \mu \): \( CV = \sigma/\mu \), which represents the magnitude of variability in relation to the mean values of runoff (Everitt and Skrondal, 2010). High CV means high variability in flows and vice versa. It shows the integrated effects of catchment geological and climatic characteristics on runoff and is mainly affected by rainfall variability and the capacity of catchment water storage. In general, catchments with higher rainfall variability or larger catchment area tend to have higher CV of monthly runoff (Fountain and Tangborn, 1985; Blöschl and Sivapalan, 1997).

### Table 2

| Model | Parameters | Equation of actual evapotranspiration | Equation of runoff generation |
|-------|------------|--------------------------------------|-------------------------------|
| TM    | \( S_{\text{max},i} \) | \[
\begin{align*}
E_{\text{obs}}(t) &= -\text{PET}(t)\quad (P(t) > \text{PET}(t)) \\
E_{\text{sim}}(t) &= P(t)\quad (P(t) < \text{PET}(t))
\end{align*}
\] | \[
R(t) = (1 - \bar{z}) \times [Q(t - 1) + \Delta Q]
\] |
| XM    | \( S_{\text{max},c} \) | \[
E_{\text{obs}}(t) = c \text{PET}(t) \quad \text{tan}(P(t)/\text{PET}(t))
\] | \[
R(t) = S(t) \quad \text{tan}(S(t)/\text{PET}(t))
\] |
| GR2M  | \( x_1, x_2 \) | \[
S_1 = (S + (\phi_1)) / (1 + \phi_1 S/x_1) \\
S_2 = (S + (\phi_2)) / (1 + \phi_2 S/x_1)
\] | \[
R_i = R + (1 - x) P_x
\] |
| VUB   | \( x_1, x_2, x_3 \) | \[
E_{\text{obs}}(t) = \min \left[ \frac{P(t) \times (1 - \gamma)}{\gamma} \right]
\] | \[
R_i = R + (1 - x) P_x
\] |
| SFB3  | \( S, F, B \) | \[
E_{\text{obs}}(t) = \min \left[ \frac{P(t) \times \gamma}{\gamma} \right]
\] | \[
R_i = R + (1 - x) P_x
\] |
| WM    | \( a, b, c, d \) | \[
E_{\text{obs}}(t) = \max \left( W(t) - (1 - e^{-\alpha t}), 0 \right)
\] | \[
R_i = R + (1 - x) P_x
\] |
| DWBM  | \( x_1, x_2, d \) | \[
E_{\text{obs}}(t) = \max \left( W(t) - (1 - e^{-\alpha t}), 0 \right)
\] | \[
R_i = R + (1 - x) P_x
\] |
| WBM   | \( a, b, c, d \) | \[
E_{\text{obs}}(t) = \frac{P(t) \times (1 - \gamma)}{\gamma}
\] | \[
R_i = R + (1 - x) P_x
\] |
| GR5M  | \( x_1, x_2, x_3, x_4, x_5 \) | \[
E_{\text{obs}}(t) = \frac{P(t) \times (1 - \gamma)}{\gamma}
\] | \[
R_i = R + (1 - x) P_x
\] |
potential evapotranspiration are calculated by the Thiessen polygon method based on the available meteorological stations in and around the catchment. The monthly streamflow records are collected from the Hydrological Bureau of the Ministry of Water Resources of China.

3. Results and discussion

3.1. Model performance assessment

Fig. 3 and Table 5 show the $N_{SE_{sqrt}}$ and $WBE$ of the 12 selected models at various percentiles of catchments. The TM model performs the worst, yielding the lowest $N_{SE_{sqrt}}$ and the largest $WBE$ at various percentiles, followed by the SFB3 model. For the other ten models, they have similar $N_{SE_{sqrt}}$ values at the 10th percentile of catchments, and the maximum difference in $N_{SE_{sqrt}}$ values is 0.13. However, the ten models have large difference in $N_{SE_{sqrt}}$ values at the high percentiles of catchments, e.g., 75th and 90th percentiles. The maximum difference of $N_{SE_{sqrt}}$ values in the 90th percentile is about 0.42, indicating that some models are not very robust when tested on a large number of catchments. The median $N_{SE_{sqrt}}$ values range from 0.30 to 0.50 for the ten models, among which the GRSM model performs the best, followed by the GR2M and WBM model. For the criteria of $WBE$, the two worst performers (the TM and SFB3 model) have large positive deviations with the median $WBE$ values of 0.45 and 0.08, respectively, indicating that the two models underestimate the total runoff volume in most catchments. The other 10 models have similar performances, and the maximum difference in median $WBE$ is 0.10. In addition, comparison results indicate that no single model performs better in all the test catchments than the others. The best performing model, namely the GRSM model, has the best performance on 36

Table 3
Summary of the catchment characteristics in the 153 catchments.

| Catchment characteristics | Songhuajiang basin | Yellow River basin | Pearl River basin and Southeast River basin |
|----------------------------|--------------------|--------------------|--------------------------------------------|
| Number of catchments       | 47                 | 45                 | 61                                         |
| Series length              | 1960–2000          | 1960–1989          | 1960–2000                                  |
| Catchment area (km²)       | 385–65,439         | 282–19,019         | 102–128,938                                |
| Mean annual rainfall (P) (mm) | 381–1025          | 123–634            | 960–2411                                   |
| Mean annual potential evapotranspiration (PET) (mm) | 572–961          | 772–1068           | 793–1137                                   |
| Coefficient of variation of monthly runoff (CV) | 0.82–2.46        | 0.55–2.50          | 0.77–1.65                                  |
| Annual runoff coefficient  | 0.06–0.51          | 0.02–0.38          | 0.25–0.89                                  |
| Arid index (P/PET)         | 0.40–0.63          | 0.30–0.60          | 0.80–2.50                                  |

Table 4
Catchments classification for 153 test catchments based on the catchment area, aridity index and CV; LRV and HRV respectively represent the catchments with low runoff variability, and the catchments with high runoff variability.

| Characteristics | Classification | Range | No. of catchments |
|-----------------|----------------|-------|-------------------|
| Area (km²)      | Small          | <2500 | 47                |
|                 | Medium         | 2500–8000 | 66              |
|                 | Large          | >8000 | 40                |
| Aridity index   | Semi-arid      | 0.2–0.5 | 54                |
|                 | Semi-humid     | 0.5–0.75 | 38               |
|                 | Humid          | >0.75 | 61                |
| CV              | LRV            | <1.0   | 44                |
|                 | Medium         | 1.0–1.5 | 52               |
|                 | HRV            | >1.5   | 57                |
catchments for $\text{NSE}_{\text{opt}}$ and 24 catchments for $\text{WBE}$, which only accounts for 24% and 16% of catchment numbers, respectively. This implies that different models have strengths and weakness in different catchments.

3.2. The relations between model performance and catchment characteristics

Fig. 4 summarizes the model performances classified by three types of catchment characteristics: catchment area, coefficient of variation (CV) of monthly runoff and aridity index, which represent the physical characteristic, runoff response characteristic and climatic characteristic of a catchment, respectively. The slope between two kinds of catchment reflects the sensitivity of model performance to the catchment characteristic: the larger slope means that the model performance is more sensitive to this catchment characteristic. Model performances are most sensitive to aridity index, followed by runoff variability and catchment area. Climatic condition is the most important factors determining model performance among the three catchment characteristics. Runoff simulation in wet catchments is significantly better than that in dry catchments (Fig. 4c and f). The possible reasons of the relative poor model performance in dry catchments are attributed to the high nonlinearity and heterogeneity of rainfall–runoff processes in these regions (Ye et al., 1997b; Atkinson et al., 2002; Millares et al., 2009). Additionally, the maximum differences in $\text{NSE}_{\text{opt}}$ for the ten best performers in the semi-arid, semi-humid and humid catchments are 0.36, 0.22 and 0.17, respectively. The model performance difference in dry catchments is much larger.

![Fig. 3. Model performances ($\text{NSE}_{\text{opt}}$ and WBE) in validation period over the 153 catchments. In all the box and whisker plots, the whiskers represent the 10th and 90th percentiles. The outer edges of the boxes represent the 25th and 75th percentiles; the horizontal lines within the boxes represent the median, i.e., 50th percentile.](image-url)

![Fig. 4. Model average performances in two validation sub-periods in three classes of catchment area (a and d), coefficient of variation of monthly runoff (CV) (b and e) and aridity index (c and f). LRV and HRV represent the catchments with low and high runoff variability, respectively.](image-url)
than that in wet catchments. Therefore, model selection in dry catchments should be more careful than that in wet catchments.

Fig. 4 indicates that model performance in large area catchments is better than that in small area catchments (Fig. 4a and d), which is in agreement with previous studies (WMO, 1975; Atkinson et al., 2002; Merz et al., 2009; van Esse et al., 2013). Merz et al. (2009) explained that a large area catchment usually has more meteorological stations than that in a small area catchment, and spatially variability of input data was averaged out as the catchment area increases. Additionally, Fig. 4b and e shows that the CV of runoff affects the accuracy of a MWBM in runoff simulation. Models perform better on the catchments with low variability in runoff than on the catchments with high variability in runoff. This result has been reported by previous studies, such as that made by Martinez and Rango (1989), Houghton-Carr (1999) and Safeeq et al. (2014).

3.3. The impacts of model complexity on model performance

Model complexity and model structure are two key factors affecting hydrological model performance (Yew Gan et al., 1997; Leplastrier, 2002; Perrin et al., 2003; van Esse et al., 2013). Some model builders expect to improve the model performance by increasing model complexity (Atkinson et al., 2002; Leplastrier, 2002; van Esse et al., 2013; Yao et al., 2014), while others argue that model performance depends mainly on the model structure and mathematical equations used to simulate the rainfall–runoff relationships (Yew Gan et al., 1997; Perrin et al., 2001; Valéry et al., 2014). Selecting a MWBM with an appropriate level of model complexity is still a challenging work. The degree of model complexity is commonly quantified by the number of free model parameters, and more parameters indicate higher model complexity (Chiew et al., 1993; Perrin et al., 2001). Fig. 5 shows the relationships between model performance and the number of model parameters. Increasing model complexity does not guarantee a better performance and simple models can achieve comparable or even better performance than the complex models. This result is consistent with WMO (1975), Michaud and Sorooshian (1994), Yew Gan et al. (1997) and Perrin et al. (2001). According to the good performance of the GR2M model, a two-parameter model is sufficient to achieve a good performance in simulating the monthly runoff. Notably, detailed description of the hydrological process does not necessarily increase model complexity correspondingly. The GR2M model provides a good example to balance the model complexity and model accuracy.

3.4. What types of model structure can achieve a better performance?

As shown in Fig. 1, the simulations of \( E_a \) and runoff generation are the two fundamental components that must be described for both simple and complex monthly hydrological models. The differences in these models are primarily reflected in the mathematical equations describing the hydrological processes (Nasseri et al., 2014).

For the simulation of \( E_a \), the models can be classified into two types according to the \( E_a \) simulation equations which are linear or nonlinear. The SFB3, WM and SM models belong to linear models, while the others except TM belong to nonlinear models (Table 2). The TM belongs to neither linear model nor nonlinear model because its \( E_a \) equation is a step function. Generally, the two types of model have comparable performance. The average values of \( NSE_{sqrt} \) for linear and nonlinear models are 0.31 and 0.35 respectively, and the average absolute value of WBE for the linear and nonlinear models are 0.099 and 0.085 respectively (Fig. 5). To provide a better comparison, we tested the sensitivity of model performance to the form of \( E_a \) equation by changing the linear models (SFB3, WM and SM model) into nonlinear models, and by changing the nonlinear models into linear models. Detailed information about models modification in \( E_a \) equations can be found in Appendix A.2. Fig. 6 summarizes the performances of the original and modified models. Model performances do not have significant changes, and the maximum differences in median \( NSE_{sqrt} \) and WBE between the original and modified models are 0.05 and 0.02, respectively. This indicates that the form of the evapotranspiration equation has limited influence on model performance in monthly hydrological model, and model developers should pay more attention to runoff generation process rather than evapotranspiration process. Similar conclusion has been made by Vandewiele and Xu (1992), they fixed the other sub-model and only changed a sub-model to identify the more appropriate model structures. They found that the form of the \( E_a \) equation was less critical to model performance than of the runoff equation. Although the non-linear \( E_a \) function has more solid physical explanations than the linear one to depict evapotranspiration process, the differences in model performance caused by different \( E_a \) equations can be compensated through adjusting model parameters. Lidén and Harlin (2000), Andréassian et al. (2004) and Oudin et al. (2005) also found that parameters adjustment can offset the negative influences of inputs errors on model performance. In addition, the TM model, the worst-performing model, is distinct from the other models in evapotranspiration calculation. The model assumes that actual evapotranspiration is equal to potential evapotranspiration when precipitation is greater than potential evapotranspiration. Considering the variation of rainfall and potential evapotranspiration within a month, water deficit likely occurs even if the monthly rainfall is greater than potential evapotranspiration (Zhang et al., 2008). Therefore, this assumption may overestimate the evapotranspiration and results in a poor model performance.

For the simulation of runoff generation, the models have large differences in the equations, the storage number and runoff
component number. Here, we provide a simple assessment on the links between model structure and performance. The selected models are classed by the number of storages and runoff components respectively. Fig. 7 indicates that the differences in average $\text{NSE}_{\text{sqrt}}$ and $WBE$ between groups of models are less than 0.03 and 0.02, respectively. No significant correlation is identified between model performance and the number of storages and the number of runoff components. This conclusion needs to be further investigated due to the model diversity in describing runoff generation processes.
4. Summary and conclusions

In this study, we compared the performance of twelve monthly water balance models over 153 catchments with different climatic conditions in China. The selected models cover a relatively broad range of complexity. We analyzed the links between model performance and catchment characteristics and discussed the possible reasons for the difference in model performance. The main findings are summarized as follows:

1. The climatic characteristic has the most important influence on the performance of MWBMs compared with catchment area and runoff characteristic. MWBMs have better performance in wet catchments than in dry catchments. Model selection in dry catchments must be more careful due to the large performance difference among different models.

2. Increasing the model complexity does not guarantee a better model performance, which corroborates previous findings in the literature. Two-parameter model is sufficient to achieve good performance for monthly runoff simulation.

3. The form of the evapotranspiration equation has limited influence on the performances of MWBMs, and model improvement could be focused on runoff generation process rather than evapotranspiration process.

Comparison results also highlight the potential complementarities of different types of MWBMs, there is not a model performing better than others in all catchments. Multi-model ensemble would be an efficient way to reduce simulation uncertainty and improve runoff simulation (Nilsson et al., 2006; Li and Sankarasubramanian, 2012; Seiller et al., 2012). Using the multi-model ensemble method to improve monthly runoff simulation should be carried out in further research.

Acknowledgements

This research was supported by the National Science Foundation of China (41201034 and 41330529), the “Strategic Priority Research Program—Climate Change: Carbon Budget and Relevant Issues” of the Chinese Academy of Sciences (Grant No. XDA05090309), the Program for “Bingwei” Excellent Talents in Institute of Geographical Sciences and Natural Resources Research, CAS (Project No. 2013RC202) and Chinese Academy of Sciences Visiting Professorship for Senior International Sciences (Grant No. 2013T220014).

Appendix A. Notations and modification of $E_a$ equation

A.1. Notations of Table 2

| Notations | Descriptions |
|-----------|-------------|
| $P$ | monthly precipitation (mm) |
| $PET$ | monthly potential evapotranspiration (mm) |
| $E_a$ | actual evapotranspiration (mm) |
| $Q$ | water surplus (mm) |
| $\Delta Q$ | water supply to water surplus storage (mm) |
| $\psi_1$ | the state variable of GR2M/GR5M that equals to $\tanh(P/PET)$ |
| $\psi_2$ | the state variable of GR2M/GR5M that equals to $\tanh(PET/X_1)$ |
| $DPE$ | the lower storage depletion factor that equals to 0.005 |
| $z$ | the relative soil moisture storage |
| $Y$ | the sum of actual evapotranspiration and soil moisture storage (mm) |
| $S$ | soil moisture storage (mm) |
| $RS$ | routing storage (mm) |
| $D$ | soil moisture storage deficit (mm) |
| $W$ | water availability that equals the sum of $P$ and $S$ (mm) |
| $R$ | runoff at the watershed outlet (mm) |
| $R_b$ | surface runoff (mm) |
| $R_g$ | groundwater flow (mm) |
| $R_s$ | base flow (mm) |
| $R_{sub}$ | sub-surface flow (mm) |
| $P_e$ | excess precipitation (mm) |
| $D_f$ | soil moisture storage in the lower layer (mm) |

A.2. Testing sensitivity of model performance to the form of $E_a$ equations; L and NL indicate linear and nonlinear equation of $E_a$ respectively.

| Type | Model | Original expressions | Modified expressions |
|------|-------|----------------------|----------------------|
| Type L to Type NL | SFB3 | $E_a(t) = \min \left\{ \frac{PET(t) \times S(t)}{S_{\text{max}}} \right\}^{PET(t)}$ | $E_a(t) = PET(t) \times \left( \frac{S(t)}{S_{\text{max}}} \right)^{7}$ |
| WM | $E_a(t) = PET(t) \times S(t)/S_{\text{max}}$ | $E_a(t) = PET(t) \times \left( \frac{S(t)}{S_{\text{max}}} \right)^{7}$ |
| SM | $E_a(t) = PET(t) \times \left( \frac{S(t)}{S_{\text{max}}} \right)^{7}$ | $E_a(t) = PET(t) \times \left( \frac{S(t)}{S_{\text{max}}} \right)^{7}$ |
| Type NL to Type L | XM | $E_a(t) = cPET(t) \tanh(P(t)/PET(t))$ | $E_a(t) = cPET(t) \times P(t)/PET(t)$ |
| GR2M | $S_1 = (5 + x_1 \phi_1)/(1 + \phi_1 S/X_1)$ | $P_1 = P(t) \times \tanh(S/X_1)$ |
| $S_2 = S_1(1 - \phi_2)/(1 + \phi_2 - \phi_2 S/X_1)$ | $E_a(t) = P(t) \times S_1 \times \frac{P(t)}{X_1}$ |
| $E_d(t) = S_1 - S_2$ | $E_a(t) = P(t) \times W(t)/W_{\text{max}}$ |
| VUB | $E_a(t) = \min \left\{ \frac{PET(t) \times (1 - A_1 W(t)/PET(t)) \times W(t)}{1 + \frac{PET(t)}{W(t)}} \right\}^{1/3}$ | $E_a(t) = PET(t) \times S(t)/S_{\text{max}}$ |
| abcd | $E_a(t) = Y \times (1 - e^{-PET(t)/b})$ | $E_a(t) = PET(t) \times W(t)/W_{\text{max}}$ |
| DWBM | $E_a(t) = W(t) \times \left( 1 + \frac{PET(t)}{W(t)} - \left[ 1 + \frac{PET(t)}{W(t)} \right]^{z_2} \right)^{1/z_2}$ | $E_a(t) = PET(t) \times W(t)/W_{\text{max}}$ |
| WBM | $E_a(t) = PET(t) \times (5z - 2z^2)/3$ | $E_a(t) = PET(t) \times S(t)/S_{\text{max}}$ |
| GR5M | Same as GR2M | Same as GR2M |
| TVGM | $E_a(t) = PET(t) \times (S(t)/S_{\text{max}})^{7}$ | $E_a(t) = PET(t) \times S(t)/S_{\text{max}}$ |
