Evaluating Regional and Global Hydrological Models against Streamflow and Evapotranspiration Measurements

YONGQIANG ZHANG, HONGXING ZHENG, FRANCIS H. S. CHIEW, AND JORGE PEÑA-ARANCIBIA
CSIRO Land and Water, Canberra, Australian Capital Territory, Australia

XINYAO ZHOU
Key Laboratory of Agricultural Water Resources, Key Laboratory of Agricultural Water Saving, Center for Agricultural Resources Research, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Shijiazhuang, China

(Manuscript received 28 June 2015, in final form 1 November 2015)

ABSTRACT

Land surface and global hydrological models are often used to characterize global water and energy fluxes and stores and to model their future trajectories. This study evaluates estimates of streamflow and evapotranspiration (ET) obtained with a priori parameterization from a land surface model [CSIRO Atmosphere Biosphere Land Exchange (CABLE)] and a global hydrological model (H08) against a global dataset of streamflow from 644 largely unregulated catchments and ET from 98 flux towers and benchmarks their performance against two lumped conceptual daily rainfall–runoff models [modèle du Génie Rural à 4 paramètres Journalier (GR4J) and a simplified version of the HYDROLOG model (SIMHYD)]. The results show that all four models perform poorly in simulating the monthly and annual runoff values, with the rainfall–runoff models outperforming both CABLE and H08. The model biases in runoff are generally reflected as a complementary opposite bias in ET. All models can generally reproduce the observed seasonal and interannual runoff variability. The correlations between the modeled and observed runoff time series are reasonable, with the rainfall–runoff models performing slightly better than CABLE and H08 at the monthly time scale and all four models performing similarly at the annual time scale. The results suggest that while the land surface and global hydrological models cannot adequately simulate the actual runoff time series and long-term average volumes, they can reasonably simulate the monthly and interannual runoff variability and trends and can therefore be reliably used for broadscale or comparative regional and global water and energy balance assessments and simulations of future trajectories. They can be improved through validating the models or calibrating some of the more sensitive and less physically based parameters.

1. Introduction

Understanding the characteristics of water fluxes and storages across regions and globally and their future trajectories is crucial to plan and manage sustainable water use under changing climate and growing population.

Various models, particularly global land surface models and global hydrological models, are used for modeling regional and global water fluxes and storages under present and future climate conditions (Fung et al. 2011; Haddeland et al. 2011; Warszawski et al. 2014). The climate community uses land surface models developed... to provide lower boundary conditions for general circulation models by solving the surface energy balance. A main focus of the land surface models is to estimate actual evapotranspiration (ET), which is a key component of the global energy and water cycle (Jung et al. 2010). The hydrology community uses global hydrological models that focus on solving water balance equations for simulating surface runoff that describes the available catchment water resource (Milly et al. 2005). The land surface and global hydrological models are similar in that they simulate the landscape hydrological processes, run over square grids covering the entire world, and use a priori parameters set to default values or values dependent on climate, vegetation, and soil data layers. The land surface models generally use more process equations and have more parameters. The
global hydrological models generally conceptualize the key hydrological processes, have fewer parameters, and also model water use in river systems (some land surface models also simulate irrigation water use). Some of the most widely used land surface models include Global Water Availability Assessment (GWAVA; Meigh et al. 1999), Hydrology-Tiled ECMWF Scheme for Surface Exchanges over Land (HTESSEL; Balsamo et al. 2009), Joint UK Land Environment Simulator (JULES; Best et al. 2011; Clark et al. 2011), Lund–Potsdam–Jena managed Land (LPJmL; Bondeau et al. 2007), Minimal Advanced Treatments of Surface Interaction and Runoff (MATSIRO; Takata et al. 2003), Max Planck Institute–Hydrology Model (MPI-HM; Hagemann and Dumenil 1997), Organizing Carbon and Hydrology in Dynamic Ecosystems (ORCHIDEE; de Rosnay and Polcher 1998), and CSIRO Atmosphere Biosphere Land Exchange (CABLE; Wang et al. 2011). Some of the most widely used global hydrological models include the Hanasaki et al. (2008) model (H08), PCRaster Global Water Balance (PCR-GLOBWB; Wada et al. 2014), Water–Global Assessment and Prognosis (WaterGAP; Alcamo et al. 2003), and VIC (Liang et al. 1994).

Land surface and global hydrological models are largely used to simulate surface energy and water fluxes across the world with a priori parameters (Gudmundsson et al. 2012; Haddeland et al. 2011), although some studies calibrated several key parameters that control soil water balances and runoff generation (Boone et al. 2004; Cai et al. 2014; Gulden et al. 2007; Lohmann et al. 2004; Mitchell et al. 2004; Niu et al. 2007, 2011; Rodell et al. 2004; Xia et al. 2012a,b; Yang et al. 2011).

In recent years, there have been several studies assessing simulations of land surface and global hydrological models against gauged streamflow (aggregating runoff across catchments or basins). For example, the Water Model Intercomparison Project (WaterMIP) compared 11 land surface and global hydrological models, and initial results suggest that the models can differ significantly in their simulation of energy and water fluxes (Haddeland et al. 2011). Zhou et al. (2012) evaluated 14 land surface and hydrological models against long-term average runoff in 150 large catchments and found that the models had large biases caused by errors in forcing data, model structure, and model parameterization. Gudmundsson et al. (2012) evaluated nine models against monthly runoff observations across Europe and found significant differences among the models, with the models having greater difficulty in modeling small catchments and snow processes. There are few studies that assess simulation of land surface and global hydrological models against the point of actual evapotranspiration measured by energy flux towers (e.g., Blyth et al. 2010).

Lumped conceptual rainfall–runoff models are widely used to model runoff in catchments and across regions and are almost always calibrated to reproduce the observed streamflow. The models’ parameters are then “regionalized” to estimate runoff in ungauged catchments or across large regions, generally by using calibrated parameter values from the geographically closest gauged catchment, or parameter values related to catchment climate and physical characteristics, or from regional calibration against observed streamflow from multiple catchments within a region (Merz and Bloschl 2004; Parajka et al. 2005; Sivapalan et al. 2003; Wagener and Wheater 2006; Zhang and Chiew 2009). The rainfall–runoff models are more parsimonious than the land surface and global hydrological models. The primary objective of the rainfall–runoff models is to model runoff and its components (e.g., surface flow, subsurface flow). Because they are calibrated against the observed streamflow, rainfall–runoff models can be expected to match the observed streamflow much better than the land surface and global hydrological models.

The purpose of this study is to evaluate the performance of a typical land surface model (CABLE) and global hydrological model (H08) against observations and to compare their estimates against those of two calibrated lumped conceptual rainfall–runoff models [modèle du Génie Rural à 4 paramètres Journalier (GRAJ) and a simplified version of the HYDROLOG model (SIMHYD)] as a “benchmark.” The four models are assessed against runoff data from 644 unregulated catchments and actual ET from 98 eddy covariance towers across the world (Fig. 1). Specifically, the study explores 1) the ability of H08 and CABLE in simulating the temporal characteristics of runoff and ET (compared to the rainfall–runoff models), 2) the relative model performance in different climate regions, and 3) how biases in annual runoff relate to biases in annual ET.

2. Data

All four models are driven by global meteorological time series (precipitation, air temperature, vapor pressure, shortwave downward radiation, longwave downward radiation, and wind speed) at 0.5° × 0.5° spatial resolution from 1980 to 2010 obtained from the Princeton Global Forcing (PGF) dataset (Sheffield et al. 2006).

The monthly streamflow data from 644 catchments (Fig. 1a) used to assess the models come from 1) the Global Runoff Data Centre (http://www.bafg.de/GRDC/EN/Home/homepage_node.html), 2) the Water Information Research and Development Alliance between CSIRO and the Australian Bureau of Meteorology (Y. Q. Zhang et al. 2013), 3) the Model Parameter Estimation Experiment (MOPEX; Duan et al. 2006), and 4) the Chinese Academy of Sciences. The catchments are
larger than 2000 km² and are largely unregulated. Major dam locations from three sources, the International Commission on Large Dams (Vörös márti et al. 2003), Meridian World Data (http://www.meridianworlddata.com/), and National Land and Water Resources Audit of Australia (http://lwa.gov.au/programs/national-land-and-water-resources-audit), were used to exclude the regulated catchments.

The monthly point ET data from 98 flux towers (Fig. 1b) used to assess the models were sourced from several global flux networks, including LaThuile FLUXNET (96 locations; http://www.fluxdata.org/DataInfo/default.aspx), OzFlux (one location; http://www.ozflux.org.au/), and AmeriFlux (one location; http://americflux.ornl.gov/). All the selected sites have more than 2 years of data, are homogeneous to a 1-km radius of the flux tower, and close the energy balance by more than 75%.

3. Descriptions of models and modeling experiments

a. CABLE land surface model

The CABLE model has been developed as a community model. A detailed description of its development, major features, and model structure are provided by Kowalczyk et al. (2006). CABLE calculates the momentum, heat, water, and CO₂ exchange between land and atmosphere. It has three components for simulating biogeochemical, vegetation-dynamics, and disturbance processes. Its integration time step is 3 h and its spatial resolution is 1.0°. CABLE performs well when simulating latent and sensible heat fluxes (Abramowitz and Pitman 2007) and long-term runoff and ET (H. Q. Zhang et al. 2013; Zhou et al. 2012) and compares well to other land surface models (De Kauwe et al. 2015; Wang et al. 2011).

b. H08 global hydrological model

The H08 model is a daily global hydrological model that consists of six main modules: land surface hydrology, river routing, crop growth, reservoir operation, environmental flow requirement estimation, and anthropogenic water withdrawal (Hanasaki et al. 2008). This study only used the land surface hydrology module that calculates the energy and water budgets above and beneath the land surface as forced by meteorological data. Potential ET and snowmelt are calculated from the surface energy balance (Hanasaki et al. 2008), and

FIG. 1. Location of the (a) 644 selected catchments and (b) 98 eddy covariance towers shown by red crosses across globe. Climate zones are classified using the Köppen–Geiger climate classification (Kottek et al. 2006).
actual ET is expressed as a function of potential ET and soil moisture. The runoff scheme in H08 is based on the bucket model concept (Manabe 1969) but modified by using a “leaky bucket” formulation in which subsurface runoff occurs continually as a function of soil moisture. Soil moisture is expressed as a single-layer reservoir with the holding capacity of 15 cm for all the soil and vegetation types. When the reservoir is empty (full), soil moisture is at the wilting point (the field capacity). Its integration time step is 1 day and its spatial resolution is 0.5°.

c. Rainfall–runoff models

The two lumped rainfall–runoff models taken as benchmark are SIMHYD and GR4J (Li et al. 2014; Li et al. 2013). Both are with 1-day integration time step. The model inputs are catchment based with the incorporation of a snowfall–snowmelt module (i.e., spatially aggregated to each catchment), including daily precipitation \( P \), daily potential evapotranspiration (ET\(_p\); Penman 1948), mean temperature, and maximum temperature. Model output for each model is daily runoff.

The SIMHYD model used here has 12 parameters and contains four storages—interception, snowpack, soil moisture, and groundwater—and a routing process (Li et al. 2014). It has been extensively used across Australia (Chiew et al. 2009, 2010; Vaze and Teng 2011; Vaze et al. 2010; Zhang et al. 2008) and modified to incorporate the snowfall–snowmelt module for application in the Tibetan Plateau (Li et al. 2013, 2014). The GR4J used here has seven parameters: three for the snowfall–snowmelt module and four for the runoff module. It has been widely used in Europe, Australia, and the Tibetan Plateau (Coron et al. 2012; Lacombe et al. 2008; Li et al. 2013, 2014; Oudin et al. 2008; Perrin et al. 2003). The details for the model structure and parameters are described by Li et al. (2014).

d. Modeling experiments

Table 1 summarizes the modeling experiments with CABLE, H08, SIMHYD, and GR4J. All four modeling experiments were carried out for 1980–2010 using the 0.5° PGF climate time series. The modeling of CABLE and H08 was carried out globally using the a priori parameter sets that are derived from field experiments, soil properties, vegetation properties, etc. (Kowalczyk et al. 2006). CABLE was run at a 3-h time step and H08 was run at a daily time step. For the daily SIMHYD and GR4J, the models were first calibrated to reproduce the observed monthly runoff by minimizing the objective function described by (Viney et al. 2009)

\[
F_1 = (1 - \text{NSE}) + 5|\ln(1 + \text{bias})|^{2.5},
\]

where the Nash–Sutcliffe efficiency (NSE) is

\[
\text{NSE} = 1 - \frac{\sum_{i=1}^{M} (Q_{\text{obs},i} - Q_{\text{sim},i})^2}{\sum_{i=1}^{M} (Q_{\text{obs},i} - Q_{\text{obs}})^2},
\]

and bias is

Table 1. Summary of the four modeling experiments used in this study.

<table>
<thead>
<tr>
<th>Model name</th>
<th>Type</th>
<th>Forcing data</th>
<th>Parameterization</th>
<th>Spatial resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>CABLE</td>
<td>Land surface model</td>
<td>PGF</td>
<td>A priori values globally</td>
<td>1.0° × 1.0° grid cells across global land surface</td>
</tr>
<tr>
<td>H08</td>
<td>Global hydrological model</td>
<td>PGF</td>
<td>A priori values globally</td>
<td>0.5° × 0.5° grid cells across global land surface</td>
</tr>
<tr>
<td>SIMHYD</td>
<td>Rainfall–runoff model</td>
<td>PGF</td>
<td>Calibrated against monthly streamflow. In model evaluation/validation, parameter values from geographically closest calibrated catchment (nearest neighbor) are used to model runoff.</td>
<td>Lumped catchment (aggregation over catchment)</td>
</tr>
<tr>
<td>GR4J</td>
<td>Rainfall–runoff model</td>
<td>PGF</td>
<td>Calibrated against monthly streamflow. In model evaluation/validation, parameter values from geographically closest calibrated catchment (nearest neighbor) are used to model runoff.</td>
<td>Lumped catchment (aggregation over catchment)</td>
</tr>
</tbody>
</table>
\[
\text{bias} = \frac{\sum_{i=1}^{M} Q_{\text{sim},i} - \sum_{i=1}^{M} Q_{\text{obs},i}}{\sum_{i=1}^{M} Q_{\text{obs},i}},
\]

in which \(Q_{\text{sim}}\) and \(Q_{\text{obs}}\) are the simulated and observed monthly runoff, respectively; \(Q_{\text{obs}}\) is the arithmetic mean of the observed monthly runoff; \(i\) is the \(i\)th month; and \(M\) is the total number of months sampled. NSE measures the agreement between the modeled and observed monthly values, with NSE = 1.0 indicating perfect agreement between all the modeled and observed monthly runoffs, and with negative NSE values indicating that the mean observed runoff better describes the observations than the modeled monthly runoff. The bias measures the water balance error between modeled and observed mean annual runoff, with bias = 0 indicating that the mean annual or total runoff is the same as the observed runoff.

In subsequent evaluations for the two lumped rainfall–runoff models, the regionalization results or application in “ungauged” catchment results are used instead of the calibration results (Table 1). In the rainfall–runoff model regionalization, parameter values from the geographically closest calibrated catchment are used to model runoff in the ungauged catchment. This is a commonly used regionalization approach to predict runoff in ungauged catchments (Bardossy 2007; McIntyre et al. 2005; Merz and Bloschl 2004; Oudin et al. 2008; Parajka et al. 2005; Zhang and Chiew 2009), although parameter values are generally transferred from catchments that are geographically much closer than in the present study. The calibration results therefore reflect practically the best possible runoff simulation with the large-scale dataset used here, while the regionalization results represent application in a more similar context (a priori parameter values or parameter values informed from elsewhere) as the land surface and global hydrological modeling.

e. Model evaluation

Simulations from the four models for 1981–2010 (not including the spinup year 1980) are evaluated against 644 catchment streamflow and 98 point ET measurements using six criteria: 1) NSE of monthly runoff and ET, 2) NSE of annual runoff and ET, 3) Pearson correlation \(R\) of monthly runoff and ET, 4) Pearson correlation \(R\) of annual runoff and ET, 5) volumetric bias of runoff and ET, and 6) standardized bias of runoff and ET. The NSE [Eq. (2)] measures the agreement between the modeled and observed time series, and \(R\) measures how well the seasonal and annual pattern in the observed time series are captured by the models. The volumetric bias [Eq. (3)] measures the precision variation between the 1981–2010 mean annual or total modeled and observed values. Similar to the volumetric bias, the standardized bias biasstd measures runoff and ET bias, but standardized by their standard deviation (std dev). It is defined as

\[
\text{bias}_{\text{std}} = \frac{\sum_{i=1}^{M} Q_{\text{sim},i} - \sum_{i=1}^{M} Q_{\text{obs},i}}{\text{std dev}(Q_{\text{obs},i})}.
\]

The standardized bias may be more appropriate than the volumetric bias for comparing runoff bias relative to the ET bias in the dry regions where mean annual ET is much larger than mean annual runoff.

The Pearson correlation is expressed as

\[
R = \frac{\sum_{i=1}^{M} (Q_{\text{sim},i} - \overline{Q}_{\text{sim}})(Q_{\text{obs},i} - \overline{Q}_{\text{obs}})}{\sqrt{\sum_{i=1}^{M} (Q_{\text{sim},i} - \overline{Q}_{\text{sim}})^2 \sum_{i=1}^{M} (Q_{\text{obs},i} - \overline{Q}_{\text{obs}})^2}},
\]

where \(\overline{Q}_{\text{sim}}\) is the mean simulated monthly runoff and \(\overline{Q}_{\text{obs}}\) is the mean observed monthly runoff.

4. Results

Figure 2 summarizes the models’ performance in simulating monthly runoff. All four models cannot reproduce the monthly observed runoff time series. The NSE from CABLE and H08 modeling is negative in more than half the catchments (Figs. 2a,b,e). The median NSEs from GR4J and SIMHYD, using parameter values from the nearest gauged catchment, are 0.15 and 0.25, respectively (Figs. 2c–e). It is noted that in subsequent results/discussion, unless otherwise mentioned, results are always benchmarked against regionalization (parameter values from nearest gauged catchment) results of the rainfall–runoff models. However, there is generally reasonable correlation between the modeled and observed monthly streamflows (Figs. 2f–j). The median correlation from CABLE and H08 is about 0.5 (i.e., the monthly correlation in more than half the 644 catchments is greater than 0.5), with GR4J and SIMHYD performing better with median correlations of 0.6 and 0.65, respectively.

Figure 3 summarizes the models’ performance in simulating annual runoff. All four models cannot reproduce the annual observed runoff time series. The NSE from CABLE and H08 is negative in practically all the catchments (Figs. 3a,b,e) and the NSE from GR4J and SIMHYD is negative in more than half the catchments (Figs. 3c–e). The correlation between the modeled and observed annual runoff time series is reasonable and surprisingly similar in all four models, with a median correlation of about 0.6 (Figs. 3f–j).
The model performances are shown separately for the different Köppen–Geiger climate zones (Kottek et al. 2006) in Fig. 4 (NSE for monthly and annual runoff) and Fig. 5 (correlation for monthly and annual runoff). The snow and polar climates are grouped together so that there are a sufficient number of catchments to allow a meaningful interpretation. There are 93, 79, 151, and 321 catchments in the equatorial climates, arid climates, warm temperate climates, and snow and polar climates, respectively. All four models generally show reasonable correlations between modeled and observed runoff time series, with similar correlations in the equatorial, arid, and warm temperate climates and slightly poorer in the snow/polar climate (Figs. 5e–h). The monthly runoff correlation in CABLE is highest in the equatorial and warm temperate climates (median correlation greater than 0.6; Fig. 5a) and the monthly correlation in H08 is highest in the equatorial, arid, and warm temperate climates (median correlation greater than 0.6; Fig. 5b). The monthly runoff correlation is highest in SIMHYD.
(Fig. 5d), while the monthly runoff correlation in GR4J is generally similar or slightly higher than in CABLE and H08 (Fig. 5c). CABLE and H08 appear to perform best in the warm temperate climate, being the only climate with positive median NSE comparing agreement between the modeled and observed monthly runoff time series (Figs. 2a,b). Likewise, the two rainfall–runoff models also perform best in the warm temperate climate, although the median NSE in all climates is generally greater than 0.2.

The results in Figs. 3k–o also show that H08 overestimates the runoff (positive bias) and CABLE tends to underestimate the runoff. The two rainfall–runoff models have smaller biases and similar tendency in overestimating or underestimating across the 644 catchments. Figure 6 further compares the volumetric biases in the 1981–2010 mean annual runoff and ET estimated by the four models for the different climate zones (results are not shown for GR4J as it does not have a separate ET module). The large runoff overestimation by H08 largely occurs in equatorial and arid climates (Fig. 6b). The CABLE model also largely overestimates runoff in the equatorial climate, but generally slightly underestimates runoff in the other climate zones (Fig. 6a). There is equal tendency in the rainfall–runoff models to overestimate or underestimate runoff, with the largest biases occurring in the arid climate. Figure 7 compares the standardized bias in runoff to that in ET, and the results are similar to the volumetric bias results in Fig. 6 and described above.

To close the energy and water balance, the runoff bias can be expected to be complementary to the ET bias. This is clear for CABLE, which shows runoff underestimation and ET overestimation in the arid, warm temperate, and snow/polar climates and runoff overestimation and ET underestimation in the equatorial climates (Figs. 6a, 6c, 7a, 7c). In H08, the runoff overestimation in snow/polar regions is complemented by
the ET underestimation, but less obvious in the other climate zones (Figs. 6b, 6f, 7b, 7f). In SIMHYD, there is a slight tendency to overestimate ET, while there is equal tendency for runoff bias in both directions (Figs. 6d, 6h, 7d, 7h; note that the large bias in arid climate simply reflects the very small runoff relative to the larger precipitation and ET). Nevertheless, note that this is not a like-for-like interpretation because the runoff and ET simulations are compared against observations from different catchments and locations.

Figure 8 summarizes the models’ performance in simulating monthly ET. The results here should be interpreted cautiously because there is generally only several years of ET data, and for the same reason comparison of annual ET is not shown here. The ET simulations from the three models (there is no ET comparison for GR4J) are better than the runoff simulations. The relative differences between the model performances in simulating ET are also smaller than their relative performances in simulating runoff. This is largely because ET is generally a larger number compared to runoff, which in some regions is a small difference between the relatively larger precipitation and ET. The correlations between the modeled and observed monthly ET time series are generally reasonable in all the climate zones (except H08 in equatorial and warm temperate climates; Figs. 8d–f). The NSE comparing the agreement between modeled and observed monthly ET is larger than 0.2 in more than half the locations in the warm temperate and snow/polar climates, compared to being negative for monthly runoff in practically all the catchments.

5. Discussion

a. Influence of forcing data on model performance

The quality of the forcing data partly contributes to the poor performance of the models. This is best seen in the rainfall–runoff models, where the NSEs from the model calibration (median NSE of 0.4–0.5) are generally 0.2–0.3 lower than the application of GR4J and SIMHYD in catchments forced with local meteorological time
series (Merz and Bloschl 2004; Oudin et al. 2008; Parajka et al. 2005; Viney et al. 2008; Wagener and Wheater 2006; Zhang and Chiew 2009). The 3-h PGF data used in this study are based on the NCEP–NCAR reanalyses (Sheffield et al. 2012), with the precipitation and temperature bias corrected using the Climatic Research Unit (CRU) Time Series, version 3.0 (TS3.0), dataset (Mitchell and Jones 2005) on a monthly time scale. This reanalysis’s coarse-resolution data provide a convenient gridded long meteorological forcing time series (from 1948 to present) across the world, but are obviously less accurate at the catchment and point scale from which observations are obtained.

b. CABLE and H08 runoff simulations relative to the rainfall–runoff models

The modeled runoff from CABLE and H08 are considerably poorer than the runoff modeled by the rainfall–runoff models. CABLE and H08 cannot reproduce the observed runoff time series, with NSE values comparing monthly and annual modeled runoff versus the observed runoff being negative in the large majority of the 644 catchments. The biases in the mean annual runoff modeled by CABLE and H08 are also relatively large (more than 50% in more than one-third of the catchments) compared to the biases from the rainfall–runoff models (less than 20% in more than half the catchments). We further analyze and present in Fig. 9 the performance of 14 land surface and global hydrological models from the Global Soil Wetness Project, version 2 (GSWP-2), an ongoing environmental modeling research activity of the Global Land Atmosphere System Study and the International Satellite Land Surface Climatology Project. The 14 models (including CABLE) were driven by the 1° × 1° GSWP-2 atmospheric forcing and default parameter sets. Details for the model setup and application are described in Zhou et al. (2012). Figure 9 summarizes the 14 model estimates against observed runoff in the 644 catchments over 1986–95 using the same criteria presented earlier. Like CABLE and H08 driven with the PGF dataset, the 14 models driven by the GSWP-2
forcing also cannot reproduce the observed runoff time series.

Nevertheless, although CABLE and H08 cannot reproduce the observed runoff values, they (and the models in Fig. 9) can generally mimic the seasonal runoff variability and interannual runoff variability. The correlation between monthly modeled and observed runoff in CABLE and H08 is generally only slightly poorer than the correlation from the rainfall–runoff models, and the correlation between annual modeled and observed runoff in CABLE and H08 is generally similar to the annual correlation from the rainfall–runoff models. The median value of monthly correlation from CABLE and H08 across the 644 catchments is about 0.5 and the median annual correlation is about 0.65, and these lie within the range of results from the 14 land surface and global hydrological models in Fig. 9. These results highlight that climatic variability (from the meteorological forcing data) is probably more important than the model parameterization in modeling the variability in monthly and annual runoff. These results also suggest that while the land surface and global hydrological models, applied with a priori parameters globally, cannot adequately simulate the actual runoff time series and long-term average volumes, they can reasonably simulate the monthly and interannual runoff variability and trends, and probably also spatial differences in runoff values.

c. Regional model performance

The runoff simulations from CABLE and H08 are poorest in the cold/polar climates, indicated by the lower correlations between modeled and observed monthly and annual runoffs compared to the simulations in the
other climate regions. The relative difference between CABLE and H08 versus the rainfall–runoff models is also largest in the cold/polar climates. Two possible reasons for the poorer CABLE and H08 simulations in cold/polar climates are deficiency in the model representation of snowfall, snowmelt, and ice processes and the resolution of the forcing meteorological data. The coarse 0.5° spatial resolution in the forcing data cannot adequately drive the snowfall–snowmelt processes that are influenced by local-scale precipitation, temperature, and topography (Gudmundsson et al. 2012). In contrast, CABLE and H08 performed remarkably well in simulating the monthly ET in the snow and polar catchments. This is largely because ET is evaluated mainly against data from the vegetation growing seasons while runoff is evaluated using all available data.

It is also interesting to note that although CABLE and H08 simulations of ET can be expected to be better in very dry regions where precipitation is a dominant factor in controlling ET, this was not observed in the results. This is largely because only two of the flux towers are located in very arid catchments, while most of the flux towers that are classified in the arid climate are largely located in well-vegetated semiarid catchments.

The model bias in runoff is generally reflected as a complementary opposite bias in ET. This is particularly obvious in CABLE but less so in H08. This highlights that, although the models are attempting to reproduce the hydroclimate and ecohydrological processes, the different conceptualizations can lead to quite different results and interpretations, even at this aggregated large-scale energy and water balance. Nevertheless, it should be noted that there are generally only several years of observed ET data and the flux towers and streamflow gauges are not necessarily located in the same catchments.

d. Improving parameterization of land surface and global hydrological models

A practical way to improve parameterization of land surface and global hydrological models is to calibrate some of the more sensitive and less physically defined parameters in these models. To ensure consistent model application regionally and globally, these parameters

---

**Fig. 7.** As in Fig. 6, but for the biasstd.
can be calibrated to reproduce jointly reasonable simulations of the observations over very large regions, rather than best simulation at individual observation locations. For example, Huang and Liang (2006) and Zhao et al. (2012) showed that calibration of six sensitive parameters in the VIC, with the same values used for the entire southeastern Australia, considerably improved the runoff simulations compared to using the default a priori parameters (median daily NSE of 0.5 compared to 0.1 across 191 catchments). Huang et al. (2013) found that the most sensitive parameters in the Community Land Model are those related to the subsurface runoff parameterization followed by parameters describing soil textures and surface runoff generation. Nasonova et al. (2009) found that calibrating 15 parameters (seven for soil, one for runoff transformation, three for vegetation, and four for forcing data) in the Soil–Water–Atmosphere–Plant (SWAP) model significantly improved the daily streamflow simulation for 12 MOPEX catchments compared to modeling with a priori default parameter values.

It is likely that validation against observed data, like the gauged streamflow data and flux ET measurements used here, followed by better specification of a priori parameters, can improve the performance of land surface and global hydrological models in simulating the energy and water balance, fluxes, and stores. This is particularly so given that more and more measurements are becoming available, with increasingly longer length of data as well as measurements from more locations. Likewise, running the global models at higher resolution will allow better utilization of and validation against multiple data sources.

Another practical way for parameterization is multi-objective model calibration and assessment against satellite data like temperature, satellite-model-derived ET, leaf area index, and snow cover area (SCA). These data are not as accurate as the on-ground measurements, but

---

**FIG. 8.** The stratified model performance using the (a)–(c) NSE and (d)–(f) correlation of monthly ET. The 98 eddy covariance towers are grouped into four Köppen–Geiger climate zones as in Fig. 6.
they provide a complete spatial coverage. The model simulations can therefore be assessed against sparse on-ground measurements as well as the temporal and spatial characteristics of the satellite-derived data. Several studies have shown that model calibration against multiple data sources can improve the model performance. For instance, Crow et al. (2003) showed that multi-objective calibration of the VIC against radiometric skin temperature and streamflow improved the monthly ET estimates by up to 20% compared to calibration against streamflow alone. Zhang et al. (2009) and Zhang et al. (2011) showed that rainfall–runoff model calibration against remotely sensed monthly ET, soil moisture, and LAI, compared to calibration against streamflow alone, considerably improved the estimation of ET, soil moisture, and LAI and marginally improved the streamflow simulation for a validation period.

Various studies have also showed that the use of satellite-derived snow cover data can improve model performance. Udnaes et al. (2007) showed that calibration of a hydrological model against both streamflow and SCA data considerably improved the SCA estimate but made little difference to the streamflow estimate. Parajka and Blöschl (2008) also showed that use of the
MODIS snow cover data improved the simulation of snow cover. Franz and Karsten (2013) demonstrated that the use of remotely sensed SCA data together with recorded streamflow data to calibrate the U.S. National Weather Service Snow Accumulation and Ablation Model (SNOW-17) improve both the spatial SCA and streamflow estimates.

6. Conclusions

This study evaluates a typical land surface model (CABLE) and a typical global hydrological model (H08) against streamflow and ET observations across the world and benchmarks their performance against two lumped conceptual rainfall–runoff models (GRAJ and SIMHYD). The results show that all four models perform poorly in simulating the monthly and annual runoff values, with CABLE and H08 performing much worse than the rainfall–runoff models. The model biases in runoff are generally reflected as a complementary opposite bias in the ET. The models are generally more capable of reproducing the observed ET compared to runoff, most likely because runoff is generally a small number (often a small difference between two larger quantities in precipitation and ET, particularly in semi-arid and arid regions) that is more difficult to accurately model.

The models can generally mimic the seasonal and interannual runoff variability. The correlations between the modeled and observed runoff time series are reasonable, with CABLE and H08 performing slightly worse than the rainfall–runoff models at the monthly time scale, and all four models performing similarly at the annual time scale. This highlights that climatic variability is more important than the model parameterization in modeling the variability in monthly and annual runoff. The results also suggest that while the land surface and global hydrological models, applied with a priori parameters globally, cannot adequately simulate the actual runoff time series and long-term average volumes, they can reasonably simulate the monthly and interannual runoff variability and trends and can therefore be reliably used for broadscale or comparative regional and global water and energy balance assessments and simulations of future trajectories. It is likely that the performance of land surface and global hydrological models can be improved through validating the models, or calibrating some of the more sensitive and less physically based parameters, against streamflow and ET data as well as satellite-derived data.

Acknowledgments. This study is supported by the CSIRO Land and Water regional and global water modeling project. The authors would also like to thank Dr. Jaifu Mao, the two anonymous reviewers, and the editor for their useful comments and suggestions.

REFERENCES

Crow, W. T., E. F. Wood, and M. Pan, 2003: Multiobjective calibration of land surface model evapotranspiration predictions using streamflow observations and spaceborne surface


