Hydrol. Earth Syst. Sci., 16, 4233–4246, 2012
www.hydrol-earth-syst-sci.net/16/4233/2012/
doi:10.5194/hess-16-4233-2012

Hydrology and Earth System Sciences

Assessment of the potential forecasting skill of a global hydrological model in reproducing the occurrence of monthly flow extremes

N. Candogan Yossef[1,2], **L. P. H. van Beek**[1], **J. C. J. Kwadijk**[2], **and M. F. P. Bierkens**[1,2]

[1]Department of Physical Geography, Utrecht University, Utrecht, The Netherlands
[2]Deltares, Delft, The Netherlands

Correspondence to: N. Candogan Yossef (naze.candoganyossef@deltares.nl)

Received: 14 February 2011 – Published in Hydrol. Earth Syst. Sci. Discuss.: 6 April 2011
Revised: 10 September 2012 – Accepted: 8 October 2012 – Published: 15 November 2012

Abstract. As an initial step in assessing the prospect of using global hydrological models (GHMs) for hydrological forecasting, this study investigates the skill of the GHM PCR-GLOBWB in reproducing the occurrence of past extremes in monthly discharge on a global scale. Global terrestrial hydrology from 1958 until 2001 is simulated by forcing PCR-GLOBWB with daily meteorological data obtained by downscaling the CRU dataset to daily fields using the ERA-40 reanalysis. Simulated discharge values are compared with observed monthly streamflow records for a selection of 20 large river basins that represent all continents and a wide range of climatic zones.

We assess model skill in three ways all of which contribute different information on the potential forecasting skill of a GHM. First, the general skill of the model in reproducing hydrographs is evaluated. Second, model skill in reproducing significantly higher and lower flows than the monthly normals is assessed in terms of skill scores used for forecasts of categorical events. Third, model skill in reproducing flood and drought events is assessed by constructing binary contingency tables for floods and droughts for each basin. The skill is then compared to that of a simple estimation of discharge from the water balance (P-E).

The results show that the model has skill in all three types of assessments. After bias correction the model skill in simulating hydrographs is improved considerably. For most basins it is higher than that of the climatology. The skill is highest in reproducing monthly anomalies. The model also has skill in reproducing floods and droughts, with a markedly higher skill in floods. The model skill far exceeds that of the water balance estimate. We conclude that the prospect for using PCR-GLOBWB for monthly and seasonal forecasting of the occurrence of hydrological extremes is positive. We argue that this conclusion applies equally to other similar GHMs and LSMs, which may show sufficient skill to forecast the occurrence of monthly flow extremes.

1 Introduction

Global hydrological models (GHMs) that simulate land surface dynamics of the hydrological cycle on a global scale have developed rapidly over the past decades. GHMs are comparable to land surface models (LSMs), such as H-TESSEL (Pappenberger et al., 2011; Balsamo et al., 2009), ISBA-SGH (Decharme and Douville, 2006), MOSES (Gedney and Cox, 2003), NOAH (Ek et al., 2003), MATSIRO (Takata et al., 2003) and SWAP (Gusev and Nasonova, 2003), which were introduced in general circulation models (GCMs) to resolve the land component and provide realistic lower boundary conditions on temperature and moisture (Decharme and Douville, 2007). Although largely similar to LSMs, GHMs focus more on modeling runoff and streamflow, as well as a more comprehensive representation of the terrestrial hydrological processes. Examples are VIC (Wood et al., 1992), WaterGap (Döll et al., 2003), LaD (Milly and Schmakin, 2002), WBM (Fekete et al., 2002), and Macro-PDM (Arnell, 1999). GHMs and LSMs have been widely applied to estimate current and future continental runoff (Nijssen et al., 2001a; Fekete et al., 2002; Milly et al., 2005), to investigate the hydrological response to global warming (Arnell, 2004; Nijssen et al., 2001b; Milly et al., 2005), to study future projections of extremes in river discharge (Hirabayashi et al., 2008; Lehner et al., 2006) and to assess freshwater availability (Alcamo et al., 2003; Islam et al., 2007; Oki et al., 2001; Vörösmarty et al., 2000; Van Beek et al., 2011; Wada et al., 2011).

Given the capability of GHMs to quantify streamflow, their relevance for integrated water resources management of large river basins has been recognized (Refsgaard, 2001). Reliable and timely forecasts of extremes in streamflow can help mitigate flood and drought risks and optimize water allocations to different sectors and sub-regions. The application of GHMs could be particularly promising for developing regions of the world where no effective flood and drought early warning systems are in place. However, up to now large-scale hydrological models have rarely been used for river flow forecasting, mainly because appropriate routing of river discharge is not included, and forecasting systems are limited to higher resolution national or regional domains (e.g. the European LISFLOOD system with a grid resolution of 5 × 5 km; De Roo et al., 2000).

In this paper we investigate the skill of the global hydrological model PCR-GLOBWB in reproducing the occurrence of past extremes in the monthly discharges of 20 large rivers of the world that represent all continents and a wide range of climatic zones. The motivation for the paper is twofold. The first objective is to present our evaluation of PCR-GLOBWB as an initial step in assessing the prospect of using a GHM for forecasting hydrological extremes. The second one is to identify a methodology that can serve as a benchmark verification procedure for hydrological forecasting. This procedure uses methods and skill scores that were developed primarily for verification of meteorological forecasts.

Global terrestrial hydrology is simulated for a historical period from 1958 until 2001, by forcing PCR-GLOBWB with a meteorological data set produced by combining ERA-40 reanalysis (Uppala et al., 2005) and Climate Research Unit (CRU) data from the University of East Anglia (New et al., 2000). The use of a historical meteorological dataset implies that the hydrological forecasts are not affected by forecasting uncertainty in the forcing and the propagation thereof with increasing lead times. In this sense, the results presented here are indicative of the maximum skill that can currently be achieved by this and similar GHMs given the associated errors in forcing, discharge observations, model structure and parameterization.

We assess the skill of PCR-GLOBWB in reproducing hydrological extremes in three ways. First, a general verification of simulated hydrographs is carried out. Second, model skill in reproducing significantly higher and lower flows than the monthly normals is assessed by constructing categorical contingency tables and applying skill scores used in meteorology for forecasts of ordinal categorical events. Third, model skill in reproducing flood and drought events is assessed by applying verification measures for forecasts of binary events, where floods and droughts are defined in terms of discharge values being higher or lower than discharges associated with a given return period. The model skill quantified in terms of these three sets of skill scores is then compared with the skill obtained by a simple estimation of discharge from the water balance (P-E) over each basin.

We use discharge observations from the Global Runoff Data Center (GRDC) reference dataset which contains monthly discharges for most basins. Consequently, the forecasting skill that we assess in this study is indicative for the potential skill that could be achieved in monthly and seasonal forecasting, rather than medium-range forecasting.

Among other studies in which the discharge simulations of other GHMs and LSMs have been compared to discharge observations, the novelty of this work is to evaluate the ability of a GHM in reproducing the occurrence of anomalous flows and past flood and drought events with skill measures used in verification of meteorological forecasts, in the prospective context of operational hydrological forecasting.

The rest of this paper is set up as follows: Sect. 2 describes the GHM PCR-GLOBWB, the historical simulation, the meteorological forcing as well as the discharge data used for skill assessment. Section3 describes the assessment of skill in reproducing hydrographs, anomalous flows and floods and droughts. Results are presented and discussed in Sect. 4, followed by conclusions in the last section.

2 Historical simulation

2.1 Hydrological model

PCR-GLOBWB (PCRaster Global Water Balance) is a hydrological model that simulates the terrestrial part of the global water cycle (Van Beek and Bierkens, 2009; Bierkens and Van Beek, 2009). It is coded in the high-level computer language PCRaster for constructing environmental models (Wesseling et al., 1996). PCR-GLOBWB is fully distributed and operates on a regular grid with a cell size of 0.5 × 0.5° (ca. 55 km squared at the Equator). Meteorological forcing is applied on a daily time step and assumed to be constant over the grid cell. Sub-grid variability is taken into account in the representation of short and tall vegetation, open water, different soil types, saturated area, surface runoff, interflow and groundwater discharge.

PCR-GLOBWB is a "leaky-bucket" type of model that calculates the water balance for every grid cell by tracking the transfer of water between the atmosphere and the cell, through stores within each cell, and laterally, as discharge, from one cell to the next. The model calculates the storages and fluxes of water, simulates the generation of runoff and its propagation as discharge through the river network. Precipitation falls either as snow or rain depending on atmospheric temperature. It can be intercepted by vegetation and added to the finite canopy storage, which is subject to open water evaporation. Snow is accumulated when the temperature is lower than 0 °C and melts when it is higher. Snow melt is added to rain and throughfall; it is stored in the available pore

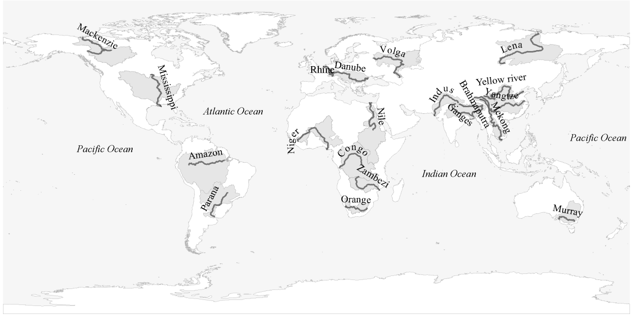

Fig. 1. Selected catchments.

space in the snow cover, or reaches the top soil layer. Part of this water is transformed into surface runoff and the remainder infiltrates into the soil through two vertically stacked soil layers and an underlying groundwater layer. Water is exchanged between these layers following Darcy's law and the resulting soil moisture is subject to evapotranspiration. The remaining water contributes to lateral drainage as interflow from the soil layers or baseflow from the groundwater reservoir. The total drainage which consists of surface runoff, interflow and baseflow is routed through the drainage network of rivers, lakes and wetlands, based on DDM30 (Döll and Lehner, 2002), using the kinematic wave approach. An extensive description of PCR-GLOBWB can be found in Van Beek and Bierkens (2009) and Van Beek et al. (2011).

2.2 Meteorological data set

The meteorological variables required to force PCR-GLOBWB are daily values of precipitation, evapotranspiration and temperature. In the absence of direct estimates of actual evapotranspiration, the model can be forced with values of potential evapotranspiration calculated from temperature, radiation, cloud cover, vapour pressure and wind speed.

In order to force PCR-GLOBWB with daily meteorological data at 0.5° resolution, the monthly fields of the CRU TS 2.1 data set (New et al., 2000) have been downscaled to daily fields using ERA-40 reanalysis (Uppala et al., 2005). Precipitation fields are downscaled multiplicatively while an additive correction is used for temperature. Reference potential evapotranspiration is calculated first on a monthly basis, based on monthly cloud cover and vapour pressure deficit from CRU TS 2.1 as well as radiation and wind speed from CRU CLIM 1.0 (New et al., 2002). Reference evapotranspiration is converted to crop-specific potential evapotranspiration using crop factors derived following FAO guidelines. Finally, potential evapotranspiration is downscaled multiplicatively to daily values using ERA-40 temperature fields. The methodology used to calculate potential evaporation for the different land surfaces in PCR-GLOBWB and the downscaling of the meteorological data is described in detail by Van Beek (2008). The resulting meteorological data set is limited to the period from 1958 to 2001 for which ERA-40 data are available.

Table 1. Basins data.

Basin	Area (km^2)	Q avg (m^3 s^{-1})	Length of records
Amazon	6 915 000	190 000	28 yr
Congo	3 680 000	41 800	26 yr
Mississippi	2 981 076	12 743	40 yr 9 months
Nile	3 400 000	2830	40 yr 7 months
Lena	2 500 000	17 000	24 yr
Parana	2 582 672	18 000	33 yr
Yangtze	1 800 000	31 900	31 yr
Mackenzie	1 805 000	10 700	16 yr 4 months
Volga	1 380 000	8060	24 yr
Niger	2 117 700	6000	21 yr 10 months
Murray	1 061 469	767	16 yr
Orange River	973 000	365	20 yr 3 months
Ganges	907 000	12 015	9 yr
Indus	1 165 000	6600	10 yr 6 months
Danube	817 000	6400	42 yr 10 months
Yellow River	752 000	2571	30 yr
Brahmaputra	930 000	48 160	5 yr 10 months
Rhine	65 638	2200	29 yr
Zambezi	1 390 000	3400	4 yr
Mekong	795 000	16 000	29 yr 5 months

2.3 Simulated and observed discharge time series

The simulated discharge time series represent non-regulated flow. Twenty large river basins are selected for comparison of simulated and observed time series on the basis of three criteria. The first one is to represent all the continents, a wide range of climate zones and latitudes as well as a variety of precipitation regimes. The second criterion is the availability of observed monthly streamflow records for at least part of the period 1958–2001. The third criterion is to focus on developing regions which would benefit most from operational seasonal forecasting. Selected basins can be seen in Fig. 1 (Sperna Weiland et al., 2010). Basin characteristics and record length are presented in Table 1, adapted from Sperna Weiland et al. (2010).

The discharge data for most of the selected basins are obtained from the Global Runoff Data Center (GRDC, 2007). When GRDC data are not available, records from the Global River Discharge Database, RivDis 1.1 (Vörösmarty et al., 1998) are used. The period of record for the discharge values reported in the GRDC and RivDis databases varies widely from basin to basin (Table 1). Simulated daily discharges for the model grid cells corresponding to gauging stations are aggregated into monthly values, since this is the temporal resolution at which observed discharge data are available for validation. The simulated and observed discharge time series are used in the assessment of skill as described in the following section.

3 Skill assessment methodology

3.1 Measuring the skill in reproducing hydrographs

The performance of the model in hydrograph simulation is assessed in terms of verification measures used in forecasting of continuous variables, without applying thresholds. For this assessment, the most commonly applied statistical measure, mean squared error (MSE) is calculated for each river basin. In order to judge the predictive skill, the raw MSE scores are transferred into MSE Skill Scores, (MSESS). The MSESS provide a relative measure of the quality of the simulation compared to the mean climatology as a low skill alternative method of estimation. Here climatology refers to the long term mean of the available monthly discharge records for each of the 12 months of the year. The MSESS is defined as:

$$\mathrm{MSESS} = 1 - \frac{\mathrm{MSE}}{\mathrm{MSE\ climatology}}. \qquad (1)$$

The range of values that MSESS can take is [-∞, 1]; with the maximum value of 1 indicating perfect skill; a value of 0 indicating a model skill equivalent to the climatology; and a negative value implying that the model performs worse than the climatology.

Additionally we use the coefficient of determination (R^2) and Nash and Sutcliffe's coefficient of efficiency (NS), which are often employed in the validation of hydrological models. These coefficients provide a measure of the model skill relative to the long-term mean, and independent of the climatology. NS takes on the values [-∞, 1] and R^2 [0, 1], with higher values indicating higher skill.

Bias due to errors in the meteorological forcing, discharge records, model parameters, or simplifying assumptions, can highly degrade the quality of the output of a hydrological model (Hashino et al., 2007). This is true for our simulations as well. We applied these skill measurement methods on both non bias-corrected and bias-corrected simulation results. Verification with non bias-corrected data presents a better reflection of potential shortcomings in the skill of the GHM and provides the opportunity to compare our simulations with the results of other studies which use non bias-corrected data, such as the Water Model Intercomparison Project (WaterMIP), which quantifies and explains the differences in the results of five GHMs and six LSMs (Haddeland et al., 2011). Verification with bias-corrected data, on the other hand, is relevant for the assessment of forecasting skill, which is the ultimate purpose of this study. It provides an indication of the maximum skill that can be achieved when the systematic bias due to model errors or forcing is eliminated, as is generally the case in operational forecasting.

In this study a simple method of a posteriori bias correction is carried out. It is true that an a priori correction by basin-specific calibration has a stronger physical basis than an a posteriori adjustment of the model output. On the other hand, given the time, data and computational capacity required for model calibration, a simple post-processing has the advantage of being far more straightforward and transparent. The post-processing method we employed is as follows: bias is calculated for each pair of simulation and observation. Calculated biases are grouped into 12 months of the year, and a mean bias is calculated for each of these 12 months. Every discharge value is corrected for the mean bias calculated for the corresponding month of the year. The correction is done by simply subtracting the mean bias for the corresponding month from the simulated monthly discharge value.

3.2 Measuring the skill in reproducing anomalous flows

In order to analyze whether the model is capable of reproducing higher or lower flows than usual for a given month, the discharge time series are transformed into categorical events defined in terms of three categories of high, normal and low flow. The analysis is carried out for two different sets of categories. For the first set, high flow is defined as discharge values above the 75th percentile for the month in question; normal flow between the 75th and the 25th percentile; and low flow below the 25th percentile. For the second set, the 90th and the 10th percentiles are used. Thresholds are identified separately for simulated and observed discharge. This approach eliminates any systematic under- or overestimation in the simulations and thus removes the need for bias correction. The skill in simulating these three classes is assessed by constructing categorical contingency tables and applying skill scores for ordinal categorical events.

Here we use Gerrity Scores (GS) (Gerrity, 1992) which is a subset of the Gandin and Murphy (GM) family of equitable scores for deterministic categorical forecasts (Gandin and Murphy, 1992). The criterion of equitability is based on the principle that random forecasts or constant forecasts of the same single category receive a no-skill score (Murphy and Daan, 1985). GM scores use a scoring matrix which represents the reward or penalty accorded to each pair of simulation and observation on the contingency table. In contrast to other equitable scores such as the Heidke skill score (Heidke, 1926), and Peirce's skill score (PSS) (Haansen and Kuipers, 1965), the GM family considers differences in relative sample probabilities of categories when appropriating a reward or penalty (Livezey, 2003). A correct forecast of a low probability category is rewarded more than that of a high probability category. Likewise failure to forecast a rare event receives a lighter penalty than a common event.

GS and LEPSCAT scores (Potts et al., 1996) are the two subsets of the GM family that are appropriate for the specific case of ordinal categories, defined as ranges of a continuous variable such as discharge. In this study, GS are preferred since they are recommended by Livezey (2003) for ordinal categorical events, on the practical basis of being more convenient to use compared to LEPSCAT. GS provide higher penalties as the discrepancy between simulated and observed

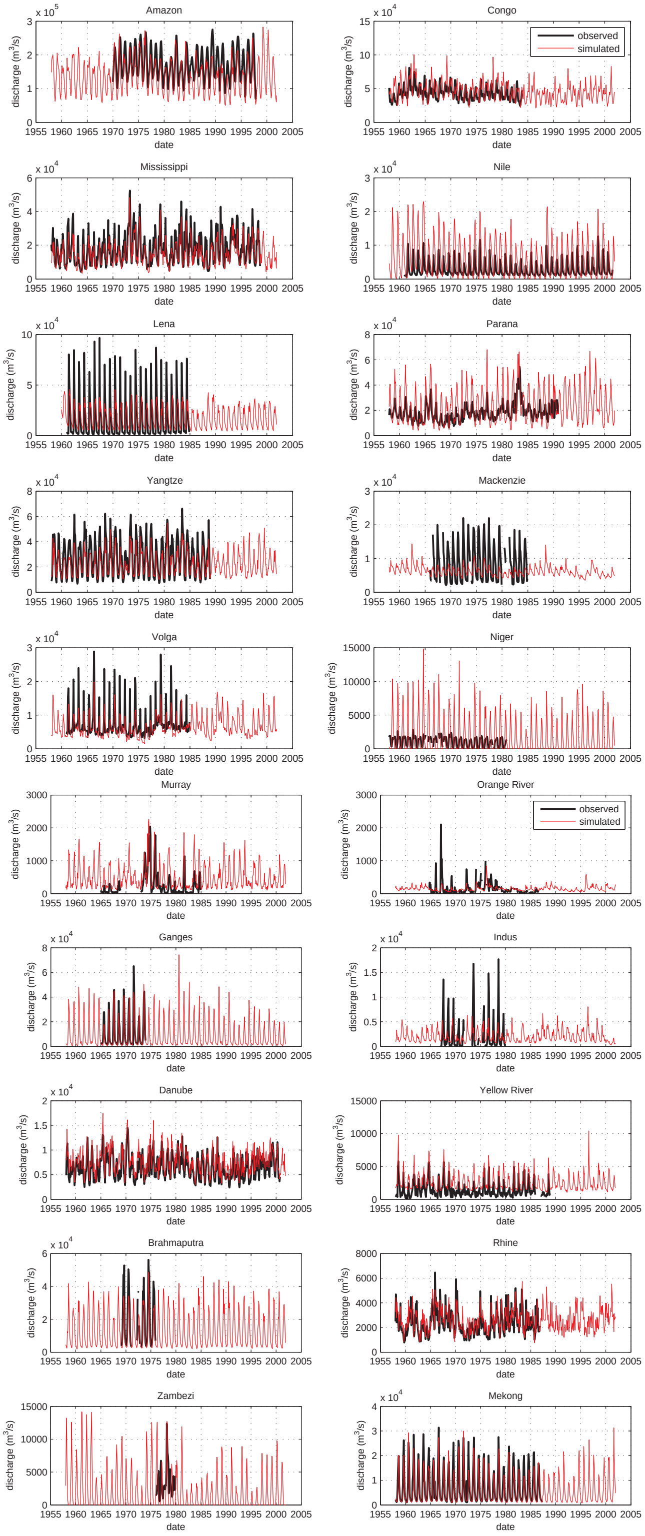

Fig. 2. Discharge time series.

classes increase. For example a forecast of low flow receives a heavier penalty when the observed flow is high, and a lighter one when the observed flow is normal.

Table 2. Skill scores for reproducing hydrographs.

Basin	uncorrected			bias corrected		
	MSESS	R^2	NS	MSESS	R^2	NS
Amazon	−4.92	0.55	−0.13	−0.29	0.79	0.75
Congo	−3.83	0.27	−0.87	−0.35	0.64	0.48
Mississippi	0.40	0.77	0.68	0.72	0.85	0.85
Nile	−31.51	0.59	−4.35	−4.38	0.57	0.11
Lena	−7.81	0.62	0.52	0.40	0.97	0.97
Parana	−2.10	0.48	−1.70	0.48	0.65	0.54
Yangtze	−0.89	0.89	0.64	0.75	0.95	0.95
Mackenzie	−10.51	0.62	0.11	0.33	0.95	0.95
Volga	−0.81	0.58	0.51	0.50	0.86	0.86
Niger	−81.30	0.11	−18.62	−6.75	0.32	−0.85
Murray	−0.70	0.37	−0.45	0.32	0.48	0.42
Orange River	0.11	0.22	0.20	0.17	0.26	0.25
Ganges	0.33	0.90	0.90	0.47	0.92	0.92
Indus	−1.63	0.12	0.12	0.08	0.69	0.69
Danube	−0.04	0.68	0.38	0.50	0.76	0.70
Yellow River	−1.98	0.77	−0.49	0.57	0.79	0.78
Brahmaputra	−1.40	0.88	0.71	0.32	0.92	0.92
Rhine	0.57	0.72	0.65	0.74	0.79	0.79
Zambezi	−1.49	0.16	−1.13	0.24	0.38	0.35
Mekong	−0.61	0.85	0.82	0.13	0.90	0.90

This score takes on the maximum value of 1 for perfect skill, and the value of 0 for no-skill. The value of GS for a categorical forecast with K number of categories is given by Eq. (2):

$$GS = \sum_{i=1}^{K} \sum_{j=1}^{K} p_{ij} s_{ij}, \qquad (2)$$

where the relative sample frequency p_{ij} of each outcome on the K × K contingency table is multiplied by the corresponding scoring factor s_{ij} $(i,\ j = 1, \ldots, K)$ from a scoring matrix **S** with relative levels of rewards and penalties and summing the values. The elements s_{ij} of the scoring matrix **S** is given by Eq. (3):

$$\mathbf{S} = \begin{pmatrix} s_{ii} & s_{ij} & \cdots & s_{iK} \\ s_{ji} & s_{jj} & \cdots & s_{jK} \\ \vdots & \vdots & \ddots & \vdots \\ s_{Ki} & s_{KK} & \cdots & s_{KK} \end{pmatrix}$$
$$s_{ii} = b\left(\sum_{r=1}^{i-1} a_r^{-1} + \sum_{r=i}^{K-1} a_r\right)$$
$$s_{ij} = b\left(\sum_{r=1}^{i-1} a_r^{-1} - (j-i) + \sum_{r=j}^{K-1} a_r\right); (1 \le i \le j \le K) \;.$$
$$s_{ji} = s_{ij}$$
$$a_i = \frac{1-\sum_{r=1}^{i} p_r}{\sum_{r=1}^{i} p_r}$$
$$p_r = \sum_{j=1}^{K} p_{rj}$$
$$b = \frac{1}{K-1} \qquad (3)$$

Table 3. Categorical contingency tables for 75th and 25th percentiles. o: observed, s: simulated, L: low flow, N: normal flow, H: high flow.

Amazon

o \ s	L	N	H
L	53	27	4
N	35	96	37
H	1	32	51

Parana

o \ s	L	N	H
L	73	23	0
N	37	140	27
H	2	34	60

Murray

o \ s	L	N	H
L	30	14	4
N	29	46	21
H	4	18	26

Yellow River

o \ s	L	N	H
L	34	45	4
N	37	116	40
H	2	25	57

Congo

o \ s	L	N	H
L	24	40	8
N	16	101	51
H	1	14	57

Yangtze

o \ s	L	N	H
L	76	20	0
N	21	141	19
H	0	29	66

Orange River

o \ s	L	N	H
L	32	26	1
N	38	76	10
H	5	28	26

Brahmaputra

o \ s	L	N	H
L	6	6	0
N	9	29	7
H	2	7	4

Mississippi

o \ s	L	N	H
L	83	37	0
N	34	181	34
H	2	27	91

Mackenzie

o \ s	L	N	H
L	24	28	0
N	19	73	10
H	3	32	17

Ganges

o \ s	L	N	H
L	18	4	2
N	18	31	11
H	2	8	14

Rhine

o \ s	L	N	H
L	59	24	0
N	25	131	25
H	1	24	59

Nile

o \ s	L	N	H
L	61	49	10
N	57	133	57
H	11	48	61

Volga

o \ s	L	N	H
L	51	19	2
N	38	93	14
H	2	26	43

Indus

o \ s	L	N	H
L	12	11	4
N	25	32	14
H	2	11	15

Zambezi

o \ s	L	N	H
L	0	9	3
N	1	14	9
H	1	5	6

Lena

o \ s	L	N	H
L	26	39	6
N	14	103	28
H	2	29	41

Niger

o \ s	L	N	H
L	11	40	15
N	6	72	52
H	2	25	39

Danube

o \ s	L	N	H
L	92	35	3
N	34	182	38
H	2	38	90

Mekong

o \ s	L	N	H
L	41	36	7
N	24	119	43
H	7	27	49

3.3 Measuring the skill in reproducing floods and droughts

Floods and droughts are regarded as simple binary events defined as exceedences of threshold discharges. For some rivers a monthly time scale may seem to be too coarse to correctly predict flood sizes. However, when we limit ourselves to forecasting monthly flows in terms of binary events, these will certainly be indicative for increased probability of floods for large rivers. It can be seen in Appendix A that at gauging station Lobith on the Rhine, throughout the years with available records during the period from 1815 to 2008, extreme daily discharges almost always coincide with large monthly discharges. When the annual maxima of daily discharge are plotted against the monthly mean discharge of the month in which this daily maximum occurred, resulting points cluster along a straight line (see Fig. A1), with daily maxima higher than monthly mean values as would be expected. Moreover, Fig. A2 shows that for most of the years, the month in which the annual maximum daily discharge occurred is also the month of maximum monthly flow or directly precedes or succeeds this month. Since the Rhine is the smallest of the 20 global rivers in this study, and given the fact that it has a rather complex regime, one can infer that the same assumption holds for other larger basins as well.

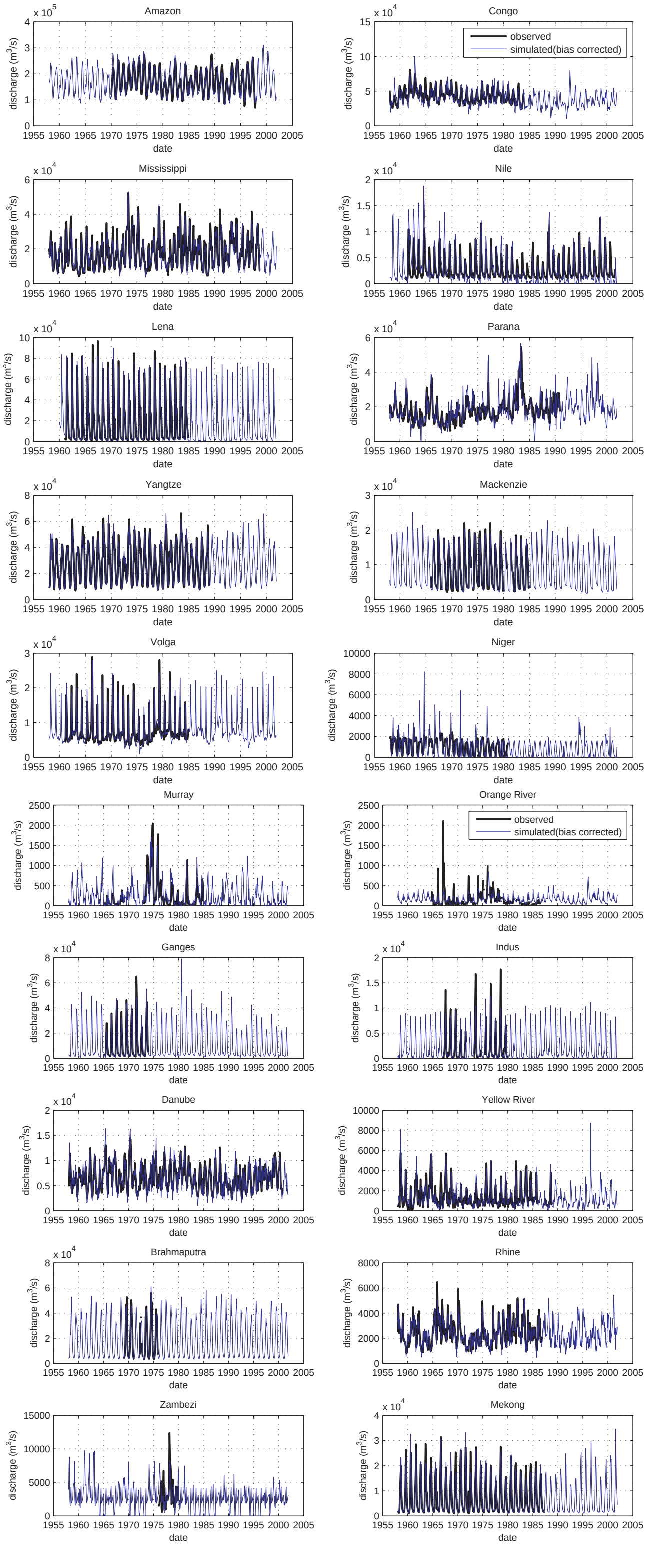

Fig. 3. Bias-corrected discharge time series.

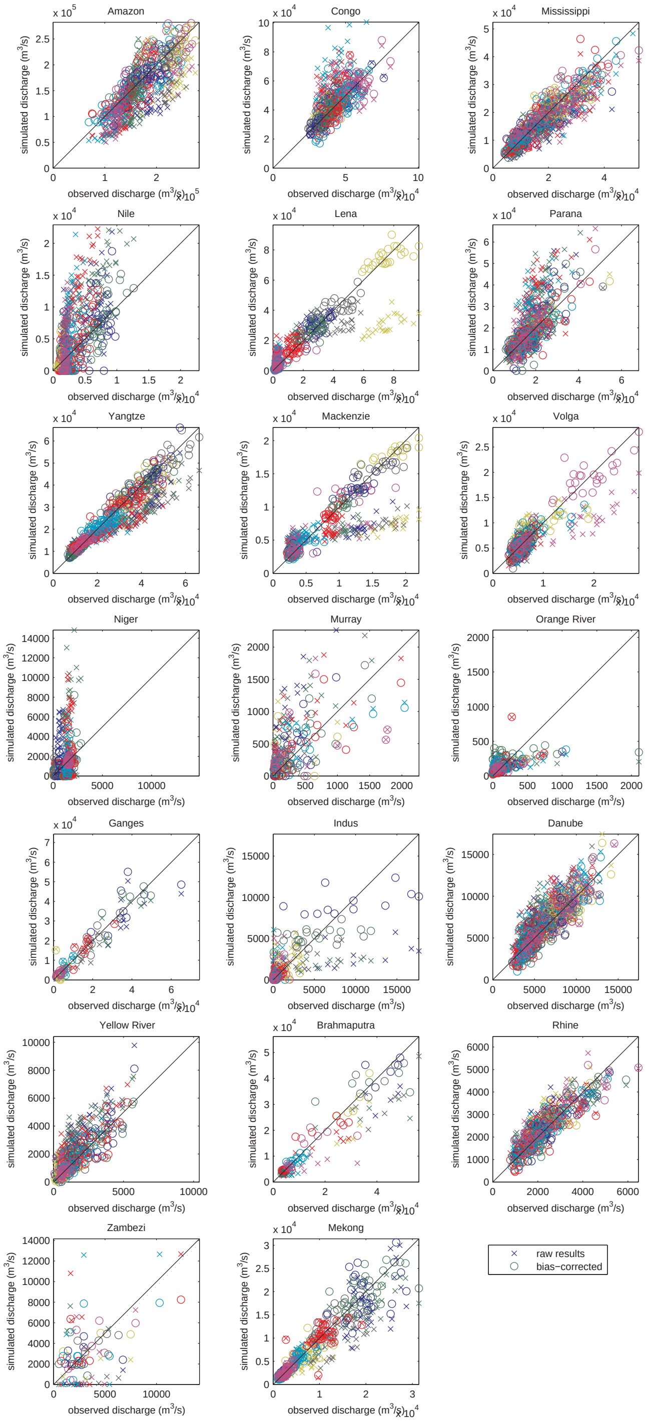

Fig. 4. Reliability diagrams (different colors indicate different months of the year).

Table 4. Categorical contingency tables for 90th and 10th percentiles. o: observed, s: simulated, L: low flow, N: normal flow, H: high flow.

Amazon

o \ s	L	N	H
L	18	18	0
N	16	228	20
H	0	19	17

Parana

o \ s	L	N	H
L	21	15	0
N	25	291	8
H	0	12	24

Murray

o \ s	L	N	H
L	6	17	0
N	17	115	13
H	1	12	11

Yellow River

o \ s	L	N	H
L	9	27	0
N	17	253	18
H	0	15	21

Congo

o \ s	L	N	H
L	4	31	1
N	7	214	19
H	0	12	24

Yangtze

o \ s	L	N	H
L	20	16	0
N	17	277	7
H	0	13	22

Orange River

o \ s	L	N	H
L	13	11	0
N	15	174	5
H	1	15	8

Brahmaputra

o \ s	L	N	H
L	3	6	0
N	3	45	4
H	0	8	1

Mississippi

o \ s	L	N	H
L	25	23	0
N	18	360	15
H	0	15	33

Mackenzie

o \ s	L	N	H
L	4	19	0
N	10	149	2
H	0	12	10

Ganges

o \ s	L	N	H
L	6	5	1
N	10	70	4
H	0	6	6

Rhine

o \ s	L	N	H
L	18	16	0
N	17	252	9
H	0	13	23

Nile

o \ s	L	N	H
L	16	29	3
N	31	332	28
H	1	31	16

Volga

o \ s	L	N	H
L	12	10	0
N	24	213	5
H	0	13	11

Indus

o \ s	L	N	H
L	1	11	0
N	9	84	9
H	0	11	1

Zambezi

o \ s	L	N	H
L	0	0	0
N	0	35	13
H	0	0	0

Lena

o \ s	L	N	H
L	3	21	0
N	8	211	21
H	1	17	6

Niger

o \ s	L	N	H
L	0	22	2
N	4	181	29
H	1	14	9

Danube

o \ s	L	N	H
L	25	23	0
N	23	373	22
H	0	22	26

Mekong

o \ s	L	N	H
L	15	20	1
N	8	250	23
H	0	25	11

Decision thresholds for a basin may be defined using various hydrological and economical criteria. A comprehensive approach with verification over the full range of possible thresholds for each basin is beyond the scope of this study. Therefore, a single set of decision thresholds for floods and droughts common for all river basins is selected that can reasonably distinguish between the usual and extreme states of each basin. The flood and drought thresholds used in this study correspond to 5-yr return periods for each river. The discharges corresponding to the 5-yr flood and drought events have been derived using the Annual Maximum Series method.

The choice of 5-yr return periods for floods as well as droughts is made on the basis of two considerations. On the one hand, events with return periods of a few years do not reflect the long-term variability, and do not represent unusually extreme states of a river. On the other hand, the limited availability of discharge observations does not allow the estimation of rare events beyond a fraction of the record length. Five years in this case appears to be a practical return period for the assessment of model skill in reproducing both types of hydrological extremes observed in all basins, the record lengths for which are given in Table 1. For the two basins with the longest records, i.e. the Danube and the Mississippi, we repeat the analysis for return periods of ten years.

Similar to the approach used in the construction of categorical tables described in Sect. 3.2, for the construction of binary tables, the thresholds for observations and simulations are identified separately in order to decrease the effect of any systematic under- or overestimation. The skill in simulating

Table 5. Gerrity skill scores in reproducing anomalous flows for 75th and 25th percentiles, and 90th and 10th percentiles.

Basin	GS-75/25	GS-90/10	Basin	GS-75/25	GS-90/10
Amazon	0.47	0.43	Murray	0.33	0.27
Congo	0.40	0.34	Orange River	0.34	0.39
Mississippi	0.63	0.57	Ganges	0.47	0.42
Nile	0.32	0.26	Indus	0.21	0.01
Lena	0.35	0.13	Danube	0.60	0.48
Parana	0.58	0.58	Yellow River	0.39	0.36
Yangtze	0.67	0.56	Brahmaputra	0.25	0.16
Mackenzie	0.29	0.28	Rhine	0.61	0.54
Volga	0.53	0.45	Zambezi	0.07	n.a.
Niger	0.15	0.12	Mekong	0.39	0.31

flood and drought events is assessed by constructing 2 × 2 contingency tables and applying binary skill scores. Binary contingency tables present the 2 × 2 possible combinations of simulated and observed event outcomes: hit, false alarm, miss and correct rejection.

Equitable skill scores used in the verification of binary forecasts are Heidke skill score (HSS) (Heidke, 1926), Peirce's skill score (PSS) (Haansen and Kuipers, 1965), Gilbert's skill score (GSS) (Schaefer, 1990) and odds ratio skill score (ORSS) (Stephenson, 2000). As stated in Sect. 3.2, the criterion of equitability is based on the principle that random forecasts or constant forecasts of the same single category receive a no-skill score (Murphy and Daan, 1985). Two of these four equitable scores, namely HSS and GSS, are markedly dependent on sample climate. Sample climate, defined as the sample estimate of the unconditional probability of occurrence of an event, is purely a characteristic of the observations with no direct relevance to skill assessment (Mason, 2003). Since dependence on sample climate makes a skill score unjustifiably sensitive to variations in observed climate and therefore unreliable, HSS and GSS are excluded in this study. The remaining two equitable scores PSS and ORSS are independent of the sample climate and recommended in several studies (McBride and Ebert, 2000; Stephenson, 2000; Göber et al., 2004). However, ORSS is also excluded because the presence of zero in any cell of the contingency table renders this skill score inappropriate (Livezey, 2003). PSS is preferred to other scores in this study on the basis of these considerations.

The possible values of PSS are within the range [−1, 1] and its true zero-skill value is 0. Negative values imply less skill than a random prediction. The PSS for floods and droughts for each basin are calculated in terms of cell counts of the relevant contingency tables according to the formula:

$$\mathrm{PSS} = \frac{a}{a+c} - \frac{b}{b+d}, \tag{4}$$

where a, b, c and d represent the cell counts for each of the possible outcomes of hit, false alarm, miss and correct rejection respectively.

3.4 Measuring added skill over a simple water balance estimate

In order to demonstrate the added value of running a complex hydrological model over a simple estimation of the water balance, the MSESS (non-bias corrected), GS and PSS are applied on an alternative set of monthly discharge values at the outlet of each basin. These discharge values are computed as follows: monthly actual evapotranspiration (E) is subtracted from the precipitation (P) on a monthly basis, then aggregated over the drainage network including downstream losses due to open water evaporation to obtain the instantaneous monthly discharge. This estimate of P-E incorporates the same information from the climatic forcing, but ignores hydrological information on stores and fluxes that lead to temporal and spatial redistribution. Skill comparison of model results with this estimate shows the added value of the routing and hydrology, while both suffer from the same poor climatological forcing.

4 Results and discussion

4.1 Skill in reproducing hydrographs

The results of the historical simulation and observed discharge time series for the selected rivers are presented in Fig. 2 for visual inspection. The difference between the simulations and observations can be attributed to several errors such as those in the meteorological forcing, discharge records, model parameters, or simplifying assumptions. The possible model errors are discussed in depth in Van Beek and Bierkens (2009) and Van Beek et al. (2011).

Three groups of rivers present a large discrepancy between the simulations and observations. The first group is the Arctic rivers, such as the Lena and Mackenzie, and snow and glacier dominated rivers such as the Indus. Undercatch in the CRU snowfall amounts reported by Fiedler and Döll (2007) results in a large underestimation of the spring discharge after the start of snowmelt. The second group consists of those basins with heavy regulation and large amounts of withdrawal for

Table 6. Binary contingency tables for floods and droughts. o: observed, s: simulated.

Amazon

	Flood			Drought	
o \ s	yes	no	o \ s	yes	no
yes	4	5	yes	4	6
no	5	322	no	3	323

Parana

	Flood			Drought	
o \ s	yes	no	o \ s	yes	no
yes	11	6	yes	0	17
no	7	372	no	13	366

Congo

	Flood			Drought	
o \ s	yes	no	o \ s	yes	no
yes	3	6	yes	0	5
no	3	300	no	10	297

Yangtze

	Flood			Drought	
o \ s	yes	no	o \ s	yes	no
yes	4	0	yes	5	5
no	2	366	no	2	360

Mississippi

	Flood			Drought	
o \ s	yes	no	o \ s	yes	no
yes	7	3	yes	7	11
no	3	476	no	11	460

Mackenzie

	Flood			Drought	
o \ s	yes	no	o \ s	yes	no
yes	1	0	yes	0	4
no	3	202	no	7	195

Nile

	Flood			Drought	
o \ s	yes	no	o \ s	yes	no
yes	5	6	yes	1	49
no	8	468	no	11	426

Volga

	Flood			Drought	
o \ s	yes	no	o \ s	yes	no
yes	2	1	yes	2	6
no	3	282	no	4	276

Lena

	Flood			Drought	
o \ s	yes	no	o \ s	yes	no
yes	2	4	yes	0	1
no	3	279	no	5	282

Niger

	Flood			Drought	
o \ s	yes	no	o \ s	yes	no
yes	1	6	yes	0	31
no	3	252	no	6	225

Murray

	Flood			Drought	
o \ s	yes	no	o \ s	yes	no
yes	5	9	yes	2	25
no	4	174	no	2	163

Yellow River

	Flood			Drought	
o \ s	yes	no	o \ s	yes	no
yes	6	6	yes	2	6
no	3	345	no	7	345

Orange River

	Flood			Drought	
o \ s	yes	no	o \ s	yes	no
yes	1	1	yes	4	13
no	4	236	no	12	213

Brahmaputra

	Flood			Drought	
o \ s	yes	no	o \ s	yes	no
yes	1	0	yes	1	2
no	0	69	no	0	67

Ganges

	Flood			Drought	
o \ s	yes	no	o \ s	yes	no
yes	1	2	yes	0	3
no	2	103	no	2	103

Rhine

	Flood			Drought	
o \ s	yes	no	o \ s	yes	no
yes	3	2	yes	5	9
no	3	340	no	6	328

Indus

	Flood			Drought	
o \ s	yes	no	o \ s	yes	no
yes	1	2	yes	0	15
no	1	122	no	2	109

Zambezi

	Flood			Drought	
o \ s	yes	no	o \ s	yes	no
yes	0	0	yes	0	0
no	1	47	no	1	47

Danube

	Flood			Drought	
o \ s	yes	no	o \ s	yes	no
yes	7	7	yes	4	7
no	6	494	no	6	497

Mekong

	Flood			Drought	
o \ s	yes	no	o \ s	yes	no
yes	2	3	yes	2	6
no	4	344	no	9	336

Table 7. Peirce's skill scores for floods and droughts.

Basin	PSS-f	PSS-d	Basin	PSS-f	PSS-d
Amazon	0.44	0.40	Murray	0.36	0.07
Congo	0.33	0.00	Orange River	0.50	0.24
Mississippi	0.70	0.39	Ganges	0.33	0.00
Nile	0.45	0.02	Indus	0.33	0.00
Lena	0.33	0.00	Danube	0.50	0.36
Parana	0.65	0.00	Yellow River	0.50	0.25
Yangtze	1.00	0.50	Brahmaputra	1.00	0.33
Mackenzie	1.00	0.00	Rhine	0.60	0.36
Volga	0.67	0.25	Zambezi	n.a.	n.a.
Niger	0.14	0.00	Mekong	0.40	0.25

irrigation and consumption, such as the Murray, Zambezi and Parana. The routing scheme in the current version of PCR-GLOBWB simulates natural discharge and does not include reservoir operations and withdrawals. Therefore the simulated natural flow on these heavily regulated rivers is in disagreement with the measured discharge. Although it is one of the most heavily regulated rivers, the Nile does not show this discrepancy since measurements of natural flow upstream of the High Dam is available for comparison. The last group consists of rivers in the tropics, which show either overestimation as in Africa, or underestimation as in the Amazon. This is mostly attributable to the low station coverage over the tropics in the CRU dataset and to a lesser extent poor precipitation forecasts in ERA-40 (Troccoli and Kalberg, 2004).

The improvement in predictive skill due to the correction of bias can be seen on the discharge time series before and after the bias correction (Figs. 2 and 3), as well as the reliability diagrams (Fig. 4). It can be observed from these figures that bias correction highly improves the results. This improvement is documented quantitatively in Table 2, which shows the MSE skill scores for the selected basins, both before and after the bias correction. Table 2 shows that without a bias correction, the MSESS for the majority of basins are negative. The improvement in the MSESS due to the correction varies widely, but is quite high in general, yielding a skill higher than the climatology for most basins. The three basins where the highest skill is observed are the Yangtze, the Rhine and the Mississippi, with MSESS above 0.70. The model performs worse than the climatology in four basins. It is interesting to note that the three basins with the worst performance, namely the Niger, the Nile, and the Congo are all African rivers. The fourth basin with negative skill is the Amazon. The relatively low skill in the Amazon and other monsoon-dominated basins such as the Indus and the Mekong can be explained to a certain degree by the fact that for such basins the climatology is already a good estimate of the expected discharge, so that it is difficult to perform better than that. The relatively high values of R^2 and NS for these basins, which are also presented in Table 2, indicate that the model performance is not poor in monsoon-dominated basins, provided that it is evaluated using measures independent of the climatology.

4.2 Skill in reproducing anomalous flows

A complete summary of the joint distribution of categorical simulations and observations for the selected basins is presented in 3 × 3 contingency tables (Tables 3 and 4). These tables provide the basis for the calculation of the Gerrity Scores for each basin. As can be seen in Table 5, all the resulting values of GS are positive, indicating that the model has skill in reproducing categorical events. In general, GS values are higher for reproducing the 75th and the 25th percentile flows than for the 90th and the 10th, as the skill is expected to decrease for more extreme flow.

The same three rivers with the highest skill in simulating exact discharges, namely the Yangtze, the Rhine and the Mississippi, have again the highest scores for categorical events. The model performance in categorical simulations for the African rivers the Niger, the Nile, and the Congo is much better than in reproducing hydrographs. The lowest skill among all the basins is observed for another African river, the Zambezi, though still above the climatology. For the Amazon, where the skill in reproducing hydrographs is less than that of the climatology, we observe that the skill in reproducing anomalous flows is rather high compared to other basins. This shows that even in cases where the model simulations are biased and do not outperform the climatology in reproducing hydrographs, the skill in reproducing anomalous flows can be relatively high.

4.3 Skill in reproducing floods and droughts

The 2 × 2 contingency tables for flood and drought events for the selected basins can be seen in Table 6. The PSS calculated on the basis of these tables are presented in Table 7. The resulting PSS show that the skill obtained by binary forecasts of 5-yr floods and droughts is also higher than an unskilled forecasting system. The system has a markedly higher skill in forecasting floods compared to droughts.

Model structure and process descriptions explain the difference in skill in reproducing floods and droughts. Floods are largely controlled by the rapid response of basins and thus react almost directly to the above-average rainfall of the forcing depending on the antecedent conditions. In contrast, droughts or low flows represent the response of the hydrological system to prolonged periods of below-average rainfall. As such, they are more sensitive to the uncertainty in model parameterization affecting processes such as the build-up of soil moisture deficit, the depletion of the groundwater system by baseflow and the regulation of discharge by reservoirs or changed withdrawal. With respect to baseflow, PCR-GLOBWB contains a conceptual model to describe the influence of lithology and drainage density. This model is parameterized using global datasets but not calibrated. As a consequence it can resolve the general trend but not all local variations. Moreover, the simulated discharge in this study is the natural one and regulation and consumption are not considered. All in all, this makes droughts more sensitive to model uncertainty, all the more so as the rank order of these events can be less accurately assessed due to the relatively larger variability of this phenomenon.

There are no basins where the model has a negative skill score in reproducing either floods or droughts; but for seven basins, the PSS indicates no skill in reproducing droughts. This is because the PSS takes on the value of zero when the contingency table shows no hits. For some basins, the model demonstrates perfect skill in reproducing floods. This is a shortcoming of the skill score used. The score takes on the value of one in cases where there are either no misses or no false alarms. Yet, to be able to assign perfect skill, one would expect the number of both misses and false alarms to be zero.

The skill assessment in reproducing 5-yr events is not applicable to the Zambezi for which the available discharge record only covers four years (see Table 1). For this basin, PSS is undefined due to the absence of any observed event. The short length of the observed discharge records affects the assessment of skill negatively for the Brahmaputra (five years and ten months) and the Ganges (nine years; Table 1).

For the two basins with the longest records, i.e. the Danube and the Mississippi, we have repeated the analysis for return periods of ten years. The results, which are presented in Appendix C, show that for both basins, PSS in floods decrease when the return period increases, as expected. For the Mississippi, the PSS in reproducing 10-yr droughts is surprisingly slightly higher than in 5-yr droughts. For the Danube, the PSS in 10-yr droughts is zero since there are no hits on the contingency tables.

Notwithstanding the problems related to limited observation lengths, skill in reproducing flood and drought events is demonstrated.

4.4 Added skill over a simple water balance estimate

The added value of running a complex hydrological model over a simple estimation of the water balance is demonstrated by comparison of the skill scores MSESS (non-bias corrected), GS and PSS for model simulated discharges and for the P-E estimate. Skill scores for both model results and for the P-E estimate are presented in Appendix B.

The results show that model skill by far exceeds that of the P-E estimate in all cases. Skill comparison of model results with this estimate shows the added value of the routing and hydrology, while both suffer from the same poor climatological forcing. In contrast, the monthly climatology of observed discharge performs better than the P-E estimate as it is more attuned to the actual climate, save for its anomalies, as well as the regulation.

5 Conclusions and recommendations

As an initial step in assessing the prospect of global hydrological forecasting, we tested the ability of a global hydrological model PCR-GLOBWB in reproducing the occurrence of past extremes in the monthly discharge of 20 large rivers of the world. We assessed the model skill in three ways: first in simulating hydrographs, second in reproducing monthly anomalies and third in reproducing flood and drought events. The advantage of such a procedure is that it provides a more detailed assessment of forecasting skill and an insight into which types of forecasting are more promising.

Verification of non bias-corrected hydrographs reflects model and forcing errors, thus providing the opportunity for improvement. In addition it allows comparison with the results of other studies which use non bias-corrected data. Eliminating the systematic bias due to model errors or forcing, on the other hand, provides an indication of the maximum skill that can be achieved in operational forecasting. Simulations with PCR-GLOBWB are biased for most basins, and the skill in reproducing hydrographs is lower than the observed climatology. The model skill improves significantly after a post-processing bias correction and surpasses the observed climatology in most basins.

Results of the analysis indicate that the skill obtained in reproducing monthly anomalies using non bias-corrected data is higher than the climatology for all basins. The model also has skill in reproducing floods and droughts, with a markedly better performance in the case of floods. The model skill surpasses that of a simple water balance estimate in all cases.

Although simulated hydrographs may be biased and do not always outperform the observed climatology even after bias correction, higher skills can be attained in forecasting the occurrence of monthly anomalies as well as floods. The prospects for operational forecasting of monthly hydrological extremes are thus positive. PCR-GLOBWB is similar to other GHMs in model structure and parameterization; and the forcing data is similar to those used in simulations with other GHMs and LSMs. The performance of PCR-GLOBWB in reproducing runoff is comparable to those of other GHMs (Sperna Weiland et al., 2010; Wada et al., 2008) and to LSMs (Sperna Weiland et al., 2011). Given these similarities we argue that our conclusion is valid for other comparable GHMs and LSMs as well.

This assessment in retrospect is a preliminary one and it shows a potential skill given the current GHM, with a meteorological forcing based on observations. The true skill should be assessed in forecasting mode using meteorological forecasts subject to uncertainty from numerical weather prediction (NWP) models.

Appendix A

Correlation between annual maxima of daily and monthly discharges at gauging station Lobith on the Rhine

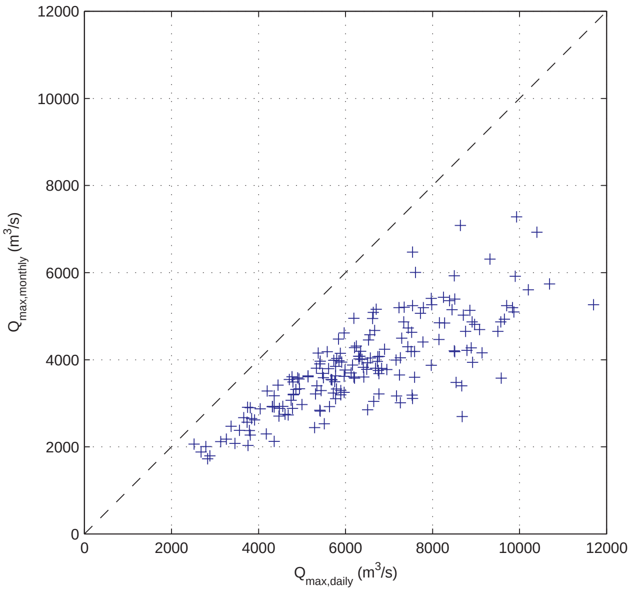

Fig. A1. Annual maxima of daily discharge vs. corresponding monthly mean flows.

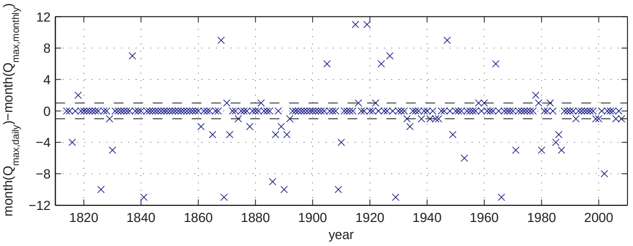

Fig. A2. The difference between the month in which the annual maximum daily discharge occurred and the month of maximum monthly flow.

Appendix B

Table B1. Skill comparison of model results of routed streamflow and streamflow estimates based on P-E fields from the water balance.

Basin	MSESS		GS		PSSf		PSSd	
	model	estimate	model	estimate	model	estimate	model	estimate
Amazon	−4.92	−21.03	0.47	0.18	0.44	0.66	0.40	0.00
Congo	−3.83	−50.09	0.40	0.19	0.33	0.00	0.00	0.00
Mississippi	0.40	−6.69	0.63	0.11	0.70	0.00	0.39	0.14
Nile	−31.51	−75 474.70	0.32	0.02	0.45	n.a.	0.02	0.00
Lena	−7.81	−13.21	0.35	0.02	0.33	n.a.	0.00	0.03
Parana	−2.10	−19.80	0.58	0.15	0.65	0.22	0.00	0.00
Yangtze	−0.89	−4.35	0.67	0.23	1.00	0.33	0.50	0.00
Mackenzie	−10.51	−12 285.40	0.29	0.04	1.00	0.00	0.00	0.00
Volga	−0.81	−30.34	0.53	−0.01	0.67	0.00	0.25	0.03
Niger	−81.30	−696.49	0.15	0.05	0.14	0.00	0.00	0.03
Murray	−0.70	−13.63	0.33	0.04	0.36	0.00	0.07	0.00
Orange River	0.11	−2.58	0.34	0.08	0.50	0.00	0.24	0.01
Ganges	0.33	−14.04	0.47	0.06	0.33	n.a.	0.00	0.00
Indus	−1.63	−3.26	0.21	−0.03	0.33	0.00	0.00	0.00
Danube	−0.04	−15.17	0.60	0.13	0.50	0.00	0.36	0.02
Yellow River	−1.98	−32.76	0.39	0.11	0.50	0.33	0.25	0.01
Brahmaputra	−1.40	−2.25	0.25	0.12	1.00	n.a.	0.33	n.a.
Rhine	0.57	−2.40	0.61	0.35	0.60	1.00	0.36	0.00
Zambezi	−1.49	−17.34	0.07	0.04	n.a.	n.a.	n.a.	n.a.
Mekong	−0.61	−8.85	0.39	0.19	0.40	0.00	0.25	0.07

Appendix C

Comparison of skill in reproducing 5-yr and 10-yr floods and droughts for the Mississipi and the Danube

Table C1. Binary contingency tables and PSS for the Mississippi.

5-yr floods				10-yr floods				5-yr droughts				10-yr droughts		
o \ s	yes	no		o \ s	yes	no		o \ s	yes	no		o \ s	yes	no
yes	7	3		yes	3	2		yes	7	11		yes	4	5
no	3	476		no	2	482		no	11	460		no	6	474
PSS= 0.70				PSS= 0.60				PSS= 0.39				PSS= 0.44		

Table C2. Binary contingency tables and PSS for the Danube.

5-yr floods				10-yr floods				5-yr droughts				10-yr droughts		
o \ s	yes	no		o \ s	yes	no		o \ s	yes	no		o \ s	yes	no
yes	7	7		yes	3	4		yes	4	7		yes	0	5
no	6	494		no	4	503		no	6	497		no	5	504
PSS= 0.50				PSS= 0.43				PSS= 0.36				PSS= 0.00		

Acknowledgements. Would you like to write an acknowledgement?

Edited by: R. Merz

References

Alcamo, J., Döll, P., Henrichs, T., Kaspar, F., Lehner, B., Rösch, T., and Siebert, S.: Development and testing of the WaterGAP 2 global model of water use and availability, Hydrol. Sci. J., 48, 317–337, 2003.

Arnell, N.: A simple water balance model for the simulation of streamflow over a large geographic domain, J. Hydrol., 27, 314–335, 1999.

Balsamo, G., Viterbo, P., Beljaars, A., Van den Hurk, B., Hirschi, M., Betts, A. K., and Scipal, K.: A revised hydrology for the ECMWF model: Verification from field site to terrestrial water storage and impact in the integrated forecast system, J. Hydrometeorol., 10, 623–643, 2009.

Bierkens, M. F. P. and Van Beek, L. P. H.: Seasonal predictability of European discharge: NAO and hydrological response time, J. Hydrometeorol., 10, 953–968, 2009.

De Roo, A. P. J., Wesseling, C. G., and Van Deursen, W. P. A.: Physically based river basin modeling within a GIS: The LISFLOOD model, Hydrolog. Process., 14, 1981–1992, 2000.

Decharme, B. and Douville, H.: Introduction of a sub-grid hydrology in the ISBA land surface model, Clim. Dynam., 26, 65–78, 2006.

Decharme, B. and Douville, H.: Global validation of the ISBA sub-grid hydrology, Clim. Dynam., 29, 21–37, 2007.

Döll, P. and Lehner, B.: Validation of a new global 30-minute drainage direction map, J. Hydrol., 258, 214–231, 2002.

Döll, P., Kaspar, F., and Lehner, B.: A global hydrological model for deriving water availability indicators: model tuning and validation, J. Hydrol., 270, 105–134, 2003.

Ek, M. B., Mitchell, K. E., Lin, Y., Grunmann, P., Rogers, E., Gayno, G., Koren, V., and Tarpley, J. D.: Implementation of the upgraded Noah land-surface model in the NCEP operational mesoscale Eta model, J. Geophys. Res., 108, 8851, doi:10.1029/2002JD003296, 2003.

Fekete, B. M., Vörösmarty, C. J., and Grabs, W.: High-resolution fields of global runoff combining observed river discharge and simulated water balances, Global Biogeochem. Cy., 16, 1042, doi:10.1029/1999GB001254, 2002.

Fiedler, K. and P. Döll: Global modeling of continental water storage changes – sensitivity to different climate data sets, Adv. Geosci., 11, 63–68, 2007, http://www.adv-geosci.net/11/63/2007/.

Gandin, L. S. and Murphy, A. H.: Equitable scores for categorical forecasts, Mon. Weather Rev., 120, 361–370, 1992.

Gedney, N. and Cox, P. M.: The sensitivity of global climate model simulations to the representation of soil moisture heterogeneity, J. Hydrometeorol., 4, 1265–1275, 2003.

Gerrity Jr., J. P.: A note on Gandin and Murphy's equitable score, Mon. Weather Rev., 120, 2707–2712, 1992.

Göber, M., Wilson, C. A., Milton, S. F., and Stephenson, D. B.: Fairplay in the verification of operational quantitative precipitation forecasts, J. Hydrol., 288, 225–236, 2004.

GRDC.: Major River Basins of the World, Global Runoff Data Centre, Federal Institute of Hydrology, D 56002, Koblenz, Germany, 2007.

Gusev, Y. M. and Nasonova, O. N.: The simulation of heat and water exchange in the boreal spruce forest by the landsurface model SWAP, J. Hydrol., 280, 162–191, 2003.

Haddeland, I., Clark, D. B., Franssen, W., Ludwig, F., Voß, F., Arnell, N. W., Bertrand, N., Best, M., Folwell, S., Gerten, D., Gomes, S., Gosling, S. N., Hagemann, S., Hanasaki, N., Harding, R., Heinke, J., Kabat, P., Koirala, S., Oki, T., Polcher, J., Stacke, T., Viterbo, P., Weedon, G. P., and Yeh, P.: Multimodel Estimate of the Global Terrestrial Water Balance: Setup and First Results, J. Hydrometeor., 12, 869–884, doi:10.1175/2011JHM1324.1, 2011.

Hanssen, A. W. and Kuipers, W. J. A.: On the relationship between the frequency of rain and various meteorological parameters, Koninklijk Nederlands Meteorologisch Institut, Mededelingen en Verhandelingen, 81, 2–15, 1965.

Hashino, T., Bradley, A. A., and Schwartz, S. S.: Evaluation of bias-correction methods for ensemble streamflow volume forecasts, Hydrol. Earth Syst. Sci., 11, 939–950, doi:10.5194/hess-11-939-2007, 2007.

Heidke, P.: Berechnung des Erfolges und der Güte der Windstärkevorhersagen im Sturmwarnungsdienst, Geografiska Annaler Stockholm, 8, 301–349, 1926.

Hirabayashi, Y., Kanae, S., Emori, S., Oki, T., and Kimoto, M.: Global projections of changing risks of floods and droughts in a changing climate, Hydrol. Sci. J., 53, 754–772, 2008

Islam, M. S., Oki, T., Kanae, S., Hanasaki, N., Agata, Y., and Yoshimura, K.: A grid-based assessment of global water scarcity including virtual water trading, Water Resour. Manage., 21, 19–33, 2007.

Lehner, B., Döll, P., Alcamo, J., Henrichs, T., and Kaspar, F.: Estimating the impact of global change on flood and drought risks in Europe: a continental integrated analysis, Climatic Change 75, 273–299, 2006.

Livezey, R. E.: Categorical events, in: Forecast Verification: A Practitioner's Guide in Atmospheric Science, edited by: Jollife, I. T. and Stephenson, D. B., Wiley, West Sussex, United Kingdom, 77–96, 2003.

Mason, I. B.: Binary events, in: Forecast Verification: A Practitioner's Guide in Atmospheric Science, edited by: Jollife, I. T. and Stephenson, D. B., Wiley, West Sussex, United Kingdom, 37–73, 2003.

McBride, J. L. and Ebert, E. E.: Verification of quantitative precipitation forecasts from operational numerical weather prediction models over Australia, Weather Forecast., 15, 103–121, 2000.

Milly, P. C. D. and Schmakin, A. B.: Global modelling of land water and energy balances, Part I: the land dynamics (LaD) model, J. Hydrometeorol., 3, 283–299, 2002.

Milly, P. C. D., Dunne, K. A., and Vecchia, A. V.: Global pattern of trends in streamflow and water availability in a changing climate, Nature, 438, 347–350, doi:10.1038/nature04312, 2005.

Murphy, A. H. and Daan, H.: Forecast evaluation, in: Probability, Statistics and Decision Making in the Atmospheric Sciences, edited by: Murphy, A. H. and Katz, R. W., Westview Press, Boulder, Colorado, USA, 379–437, 1985.

New, M., Hulme, M., and Jones, P.: Representing twentieth-century space-time climate variability, Part 1: Development of a 1961–90 mean monthly terrestrial climatology, J. Climate, 12, 829–856, 2000.

New, M., Lister, D., Hulme, M., and Makin, I.: A high-resolution data set of surface climate over global land areas, Climate Res., 21, 1–25, 2002.

Nijssen, B., O'Donnell, G. M., Lettenmaier, D. P., Lohmann, D., and Wood, E. F.: Predicting the discharge of global rivers, J. Clim., 14, 3307–3323, 2001a.

Nijssen, B., O'Donnell, G. M., Hamlet, A. F., and Lettenmaier, D. P.: Hydrologic sensitivity of global rivers to climate change, Clim. Change, 50, 143–175, 2001b.

Oki, T., Agata, Y., Kanae, S., Saruhashi, T., Yang, D., and Musiake, K.: Global assessment of current water resources using total runoff integrating pathways, Hydrolog. Sci. J., 46, 983–995, 2001.

Pappenberger, F., Cloke, H. L., Balsamo, G., Oki, P., and Ngo-Duc, T.: Global routing of surface and subsurface runoff produced by the hydrological component of the ECMWF NWP system, Int. J. Climatol., 30, 2155–2174, 2011.

Potts, J. M., Folland, C. K., Jolliffe, I. T., and Sexton, D.: Revised LEPS scores for assessing climate model simulations and long-range forecasts, J. Climate, 9, 34–53, 1996.

Refsgaard, J. C.: Discussion of model validation in relation to the regional and global scale, in: Model Validation: Perspectives in Hydrological Science, edited by: Anderson, M. G. and Bates P. D., Wiley, West Sussex, United Kingdom, 461–483, 2001.

Schaefer, J. T.: The critical success index as an indicator of forecasting skill, Weather Forecast., 5, 570–575, 1990.

Sperna Weiland, F. C., van Beek, L. P. H., Kwadijk, J. C. J., and Bierkens, M. F. P.: The ability of a GCM-forced hydrological model to reproduce global discharge variability, Hydrol. Earth Syst. Sci., 14, 1595–1621, doi:10.5194/hess-14-1595-2010, 2010.

Sperna Weiland, F. C., Van Beek, L. P. H., Kwadijk, J. C. J., and Bierkens M. F. P.: On the suitability of GCM runoff fields for river discharge modeling; a case study using model output from HadGEM2 and ECHAM5, J. Hydrometeorol.,13, 140–154, doi:10.1175/JHM-D-10-05011.1, 2011.

Stephenson, D. B.: Use of the “Odds Ratio” for diagnosing forecast skill, Weather Forecast., 15, 221–232, 2000.

Troccoli, A. and Kallberg, P.: Precipitation correction in the ERA-40 reanalysis, ERA-40 Project Rep. Series 13, 6. ECMWF, Reading, United Kingdom, 2004.

Uppala, S. M., Kållberg, P. W., Simmons, A. J., Andrae U., da Costa Bechtold, V., Fiorino, M., Gibson, J. K., Haseler, J., Hernandez, A., Kelly, G. A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R. P., Andersson, E., Arpe, K., Balmaseda, M. A., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Caires, S., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B. J., Isaksen, L., Janssen, P. A. E. M., McNally, A. P., Mahfouf, J. F., Jenne, R., Morcrette, J. J., Rayner, N. A., Saunders, R. W., Simon, P., Sterl, A., Trenberth, K. E., Untch, A., Vasiljevic, D., Viterbo, P., and Woollen, J.: The ERA-40 re-analysis, Q. J. Roy. Meteor. Soc., 131, 2961–3012, 2005.

Van Beek, L. P. H.: Forcing PCR-GLOBWB with CRU meteorological data, available at: http://vanbeek.geo.uu.nl/suppinfo/vanbeek2008.pdf (last access: November 2012), 2008.

Van Beek, L. P. H. and Bierkens, M. F. P.: The global hydrological model PCR-GLOBWB: Conceptualization, parametrization and verification, available at: http://vanbeek.geo.uu.nl/suppinfo/vanbeekbierkens2009.pdf (last access: November 2012), 2009.

Van Beek, L. P. H., Wada, Y., and Bierkens, M. F. P.: Global Monthly Water Stress: 1.Water balance and water availability, Water Resour. Res., 47, W07517, doi:10.1029/2010WR009791, 2011.

Vörösmarty, C. J., Fekete, B., and Tucker, B. A.: River Discharge Database, Version 1.1. Institute for the Study of Earth, Oceans, and Space, University of New Hampshire, Durham NH, USA, 1998.

Vörösmarty, C. J., Green, P., Salisbury, J., and Lammers, R. B.: Global Water Resources: Vulnerability from Climate Change and Population Growth, Science, 289, 284–288, 2000.

Wada, Y., Van Beek, L. P. H., Viviroli, D., Dürr, H. H., Weingartner, R., and Bierkens, M. F. P.: Water Stress over the Year: Quantitative Analysis of Seasonality and Severity on a Global Scale, MSc Thesis, University Utrecht, Utrecht, Netherlands, available at: http://igitur-archive.library.uu.nl/student-theses/2010-0308-200229/UUindex.html (last access: November 2012), 2008.

Wesseling, C. G., Karssenberg, D., Van Deursen, W. P. A., and Burrough, P. A.: Integrating dynamic environmental models in GIS: the development of a Dynamic Modeling language, Transactions in GIS. 1, 40–48, 1996.

Wood, E. F., Lettenmaier, D. P., and Zartarian, V. G.: A land-surface hydrology parameterization with subgrid variability for general circulation models, J. Geophys. Res., 97, 2717–2728, 1992.