Cross-scale evaluation of catchment- and global-scale hydrological model simulations of drought characteristics across eight large river catchments

Amit Kumar (amit1792kumar@gmail.com)
University of Nottingham
https://orcid.org/0000-0001-6444-4047

Simon N Gosling
University of Nottingham

Matthew F Johnson
University of Nottingham

Jamal Zaherpour
University of Nottingham

Guoyong Leng
Key Laboratory of Water Cycle & Related Land Surface Processes Chinese Academy of Sciences

Hannes Muller Schmied
Goethe University Frankfurt

Jenny Kupzig
Ruhr-Universitat Bochum

Lutz Breuer
Justus Liebig Universitat Giessen

Naota Hanasaki
National Institute for Environmental Studies, Tsukuba

Qiuhong Tang
Key Laboratory of Water Cycle & Related Land Surface Processes Chinese Academy of Sciences

Rohini Kumar
Helmholtz Centre for Environmental Research - UFZ

Sebastian Ostberg
Potsdam Institute for Climate Impact Research

Tobias Stacke
Helmholtz-Zentrum Geesthacht

Yadu Pokhrel
Michigan State University Department of Civil and Environmental Engineering

Yoshihide Wada
Utrecht University: Universiteit Utrecht

Yoshimitsu Masaki
Ibaraki University

Research Article

Keywords: catchment scale hydrological models, global scale hydrological models, drought
Abstract

Although global- and catchment-scale hydrological models are often shown to accurately simulate long-term runoff time-series, far less is known about their suitability for capturing hydrological extremes, such as droughts. Here we evaluated runoff simulations from nine catchment scale hydrological models (CHMs) and eight global scale hydrological models (GHMs) for eight large catchments: Upper Amazon, Lena, Upper Mississippi, Upper Niger, Rhine, Tagus, Upper Yangtze and Upper Yellow. The simulations were conducted within the framework of phase 2a of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2a). We evaluated the ability of the CHMs, GHMs and their respective ensemble means (Ens-CHM and Ens-GHM) to simulate observed monthly runoff and hydrological droughts over 31 years (1971–2001). Observed and simulated hydrological drought events were identified using the Standardised Runoff Index (SRI) and were classified based on intensity. Our results show that for all eight catchments, CHMs outperformed GHMs in monthly runoff estimation showing a better representation of observed runoff than GHMs. The number of drought events identified under different drought categories (i.e. SRI values of -1 to -1.49, -1.5 to -1.99, and ≤ -2) varied significantly between models. All the models, as well as the two ensemble means present limited ability to accurately simulate severe drought events in all eight catchments, in terms of their timing and intensity. By analysing the monthly runoff time-series for several extreme droughts over the historical period, we identify room for improvement in the models so that extreme droughts may ultimately be better represented by both CHMs and GHMs.

1. Introduction

A drought is an event where water availability is lower than normal, resulting in a failure to fulfil the water demands of different natural systems and socioeconomic sectors (WMO 1986). From 1991 to 2005, 950 million people were affected by droughts worldwide and economic damage of 100 billion US dollars was reported (UN and UNISDR, UNDP 2009). Droughts are usually the consequence of a prolonged period of below normal precipitation that also affects many other environmental, climate and social variables (Lloyd-Hughes 2014; Van Loon 2015). As a result, drought events can be difficult to identify in space and time, which makes it one of the most complex natural hazards (Wilhite 1993; Wilhite, Hayes and Svoboda 2000). Researchers, managers and policy makers quantify drought events using drought indices based on climate data (reviewed in Heim 2002; Keyantash and Dracup 2002; Mishra and Singh 2010). Whilst precipitation is a key input in calculating these indices, other climate and environmental variables that affect water storage and availability, are also significant.

Droughts are complex natural disasters as their onset and magnitude is related to the interaction between many hydrological and climatological processes. Droughts can be classified into different types namely meteorological, hydrological, agricultural and socioeconomic droughts. Meteorological droughts represent below normal precipitation and are mainly presented by precipitation driven indices such as the Standardised Precipitation Index (SPI; McKee, Doesken and Kliest 1993), Regional Drought Area Index (RDAI; Fleig et al. 2011) and Effective drought index (EDI; Byun and Wilhite 1999). In contrast, hydrological droughts define effects on freshwater storage, which are represented by indices that use stream flows, reservoir levels, groundwater levels or other similar variables. Hydrological droughts are often closely related to meteorological droughts and can also be exacerbated by environmental changes, anthropogenic activities, and mismanagement of water resources (Tallaksen et al. 2004).

Studies on hydrological droughts at global or continental scales increasingly use Land Surface Models (LSMs), Global Hydrological Models (GHMs) and Catchment Scale Hydrological Models (CHMs) to quantify and predict drought events (Gosling, Zaherpour, et al. 2017; Hattermann et al. 2017). GHMs, LSMs and CHMs have been widely used to model flood hazards and risk (Amell and Gosling 2016), climate change mitigation (Irvine et al. 2017), forecasting at shorter time scales (Emerton et al. 2016) and food security (Elliott et al. 2014). The use of these models to study droughts is also relatively common (Van Huijgevoort et al. 2013; Prudhomme et al. 2014) but there is relatively less information on the performance of the models for simulating drought. Given the societal significance of drought prediction under climate
change scenarios (Pokhrel et al. 2021) using these tools, and the influence of results on climate change adaptation and mitigation decision-making, it is critical to understand their strengths and limitations when specifically focusing on drought conditions.

Model inter-comparison projects like WaterMIP (Haddeland et al. 2011) and the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP) (Frieler et al. 2017) have made it possible to evaluate and compare simulations for historical and future periods from ensembles of GHMs and CHMs driven by common input datasets. Although recent studies have exploited these two MIPs to evaluate the general performance of ensembles of GHMs (Haddeland et al. 2011; Prudhomme et al. 2014; Zaherpour et al. 2018), CHMs (Dams et al. 2015; Huang et al. 2017), or both GHMs and CHMs together (Hattermann et al. 2017; Krysanova et al. 2020), none have comprehensively evaluated the relative capabilities of two large-scale multi-model ensembles of GHMs and CHMs to represent historical hydrological drought.

This study provides a first insight into the relative abilities of GHMs and CHMs to simulate hydrological drought events (hereafter referred to as ‘droughts’). We used GHM and CHM simulations from ISIMIP2a (Gosling, Müller Schmied, et al. 2017) to identify historical drought events from their respective monthly runoff simulations, and evaluated how these droughts compared to the observed record. The main objectives of this study were, firstly, to systematically evaluate the performance of global scale and catchment scale hydrological models to simulate droughts and, secondly, to analyse the simulated drought by ensemble mean to determine how well it represents droughts simulated by individual models. Finally, we discuss the opportunities for improving the simulation of drought events by both GHMs and CHMs.

2. Data And Methods

2.1. Study Catchments

Eight large (> 65,000 km²) catchments were selected to cover important climate zones and hydrological systems around the globe. These were, the Upper Amazon, Lena, Upper Mississippi, Upper Niger, Rhine, Tagus, Upper Yangtze and Upper Yellow (Fig. 1). They are the same eight catchments used in previous ‘cross-scale’ GHM-CHM comparisons (Gosling, Zaherpour, et al. 2017; Hattermann et al. 2017). For the analysis, only the upper part of the Amazon, Mississippi, Niger, Yangtze and Yellow were modelled due to their complicated geomorphological structure and human alterations further downstream (Krysanova and Hattermann 2017). Catchment boundaries were defined according to Drainage Direction Maps at 30’ (DDM30; Döll and Lehner 2002) for the GHMs and CHMs, and according to the Global Runoff Data Centre (GRDC; http://grdc.bafg.de) for the observed data.

2.2. Models and input data

Simulated runoff data from eight GHMs and nine CHMs along with corresponding observed runoff data from the GRDC, for 1971–2001, were used in this study. Observed runoff data was acquired from the most downstream gauge in the GRDC catalogue for each catchment. All GHMs and CHMs were run with input climate data from WATCH ERA-40 (Weedon et al. 2011) for the period 1971–2001. Output from the models is openly available from the Earth System Grid Federation (ESGF; https://esg.pik-potsdam.de/search/isimip) for the GHMs (Gosling, Müller Schmied, et al. 2017) and the CHMs (Krysanova et al. 2017). All the GHMs provided outputs for all catchments; however, the number of CHMs with simulated runoff varied by catchment (Table 1). Following the method described by Haddeland et al. (2011), monthly observed and simulated runoff data was converted to catchment-mean monthly runoff by using the area upstream of the gauge according to the DDM30 river network. Thus, an area correction factor was applied to the GRDC runoff data to account for the fact that the river network, which is at 0.5° spatial resolution, may not perfectly overlap with the GRDC river catchment boundaries (Table 1).

Both GHMs and CHMs simulate the full hydrological cycle with predominantly daily precipitation and temperature as input data. All the GHMs simulated hydrological processes at a spatial resolution of 0.5° x 0.5° across the global land surface.
In contrast, CHMs operated using various approaches; three CHMs run on a grid (mHM, VIC, WaterGAP3), four by splitting the catchment into sub-catchments and smaller hydrological response units (HBV, HYPE, SWAT, SWIM) and one by considering the whole catchment as a single entity (HYMOD). The GHMs were not calibrated to catchment specific conditions, except WaterGAP2 (which was calibrated against long-term average monthly runoff for a number of gauges worldwide) while the CHMs were calibrated and the performance of the calibration evaluated in a separate validation period using the WATCH reanalysis climate forcing data (Huang et al. 2017).

In addition to the individual model results, we calculated the corresponding ensemble mean for the GHMs and CHMs respectively for every catchment and included them in the analysis, for meeting our second study objective. For the purposes of this study, all the models were treated as independent even though many of the models operate on similar model parameterisation. No hydrological model was excluded or weighted based on their performance on simulating present day runoff.

2.3. Drought Indicator and performance evaluation

Several standardised indices have been developed for identifying hydrological droughts, such as the Palmer Hydrological Drought Index (PDSI; Jacobi et al. 2013), Water Supply Index (WSI; Garen 1993) and Standardised Streamflow/Runoff Index (SSI/SRI; Vicente-Serrano et al. 2012). In comparison with PHDI and WSI, SSI is more commonly accepted due the fact that it is simple to calculate, can be used on various time scales and above all, requires fewer inputs. SSI is extensively used in many studies (Shukla and Wood 2008; Vicente-Serrano et al., 2012; Wu et al. 2018; Liu et al. 2019).

Calculation of SSI/SRI is similar to that of Standardised Precipitation Index (SPI) proposed by McKee, Doesken and Kliest (1993) but considering runoff instead of precipitation. The SRI values are determined based on long-term runoff records (preferably > 30 years) by aggregating the monthly runoff over an accumulation period (1, 3, 6, 12, or 24 months). The new series formed after accumulation (1 month for this study) is then fitted to a probability distribution that subsequently is transformed to a normal distribution such that the mean SRI is zero. The “SPEI” package in R (Beguería and Vicente-Serrano 2017) was used for all the calculations of SRI. “SPEI” i.e. Standardized Precipitation-Evapotranspiration Index package facilitates computation of SPI and other variants of SPI (SRI for our study) by providing defined functions that can be used directly in the R working environment. Any positive SRI values indicate runoff values greater than the mean monthly runoff and vice versa. Any SRI value less than −1 was considered to indicate drought condition. These drought conditions were categorised based on SRI values into three drought categories namely moderate, severe and extreme droughts. SRI values from −1 to -1.49 signifies moderate drought, from −1.5 to -1.99 severe drought, and all the SRI values less than −2 as extreme drought.

Continuous SRI values ≤-1 signified drought events, which lasted until the SRI rose above −1 again. Two main factors play an important role in the identification a drought event: the intensity and duration. Therefore, all the comparisons made between observed and modelled records were based on intensity, duration and frequency of drought events. We calculated three drought characteristics − drought intensity, drought duration and frequency of drought events to analyse hydrological drought conditions for all eight catchments. Run theory method (Yevjevich and Ica Yevjevich 1967) was used for the extraction of drought characteristics from drought index series i.e. SRI time series. The drought duration (in months) was taken as the period the SRI remained ≤-1, and the minimum SRI value in this period was used as the drought intensity.

We used the coefficient of determination i.e. the square of Pearson product moment correlation coefficient (R²) with a confidence interval of 0.95, and the Nash-Sutcliffe coefficient (NSE; Nash and Sutcliffe 1970) to evaluate the goodness of fit between observed runoff and simulated runoff obtained from different CHMs and GHMs. R² indicated the strength of relationship between the simulated and observed runoff, and NSE indicated the models efficiency (range −1 to 1). NSE values approaching 1 indicate a perfect match of simulated and observed runoff values while negative values show that the observed mean is a better predictor than the model.
3. Results

3.1. Comparison of observed and simulated runoff

Both the $R^2$ and NSE were used for capturing nuances between simulated and observed runoff. Table 2 shows the $R^2$ and NSE values between observed monthly runoff and all the CHMs and GHMs incorporated in the analysis, including the ensemble mean of the GHMs (Ens-GHM) and CHMs (Ens-CHM) respectively. All the direct comparisons drawn between CHMs and GHMs are based upon Ens-CHM and Ens-GHM, and not the individual CHMs or GHMs.

For all eight catchments, CHMs out-performed GHMs in runoff estimation showing a better representation of observed runoff than GHMs. Ens-CHM and Ens-GHM have $R^2$ values greater than 0.7 (Moriasi et al. 2007, 2015) for all catchments except for Lena by Ens-GHM. Comparing both ensemble outputs, $R^2$ values of Ens-CHM were higher than those of Ens-GHM for all catchments. Similarly, NSE values of Ens-CHM for all eight catchments were greater than 0.7 and closer to 1 for Rhine, Tagus and Yangtze catchments, while for Ens-GHM only two catchments i.e. Yangtze and Yellow had NSE values greater than 0.7, with negative values for the Niger and Rhine.

3.2. Comparison of observed and simulated SRI

$R^2$ and NSE values for SRI series obtained from simulated runoff outputs from CHMs and GHMs are shown in Table 3, including the ensembles i.e. Ens-GHM and Ens-CHM. Both $R^2$ and NSE vary widely across all catchments for both GHMs and CHMs. The $R^2$ values of the ensemble series varied between 0.18 (Ens-GHM) and 0.94 (Ens-CHM) while the NSE ranged from −0.14 (Ens-GHM) to 0.95 (Ens-CHM) across all catchments. The catchments where most of the models (both GHMs and CHMs) performed well were the Mississippi and Rhine, both with $R^2$ and NSE values for Ens-CHM and Ens-GHM > 0.85. For both these catchments, individual models always had $R^2 > 0.50$ and NSE values > 0.40. No individual GHM or CHM achieved $R^2$ and NSE values > 0.7 for the other six catchments, with the exception of HBV for the Tagus catchment ($R^2 = 0.75$ and NSE = 0.74). The catchment where all the models performed the worst was the Niger, where no individual GHM or CHM (except SWIM) achieved an $R^2$ value > 0.35 and NSE values were always close to, or below zero. For only three catchments, Ens-CHM had $R^2$ and NSE values > 0.70 and Ens-GHM for only two.

3.3. Evaluation of the frequency of drought events

Table 4 presents the number of drought events identified from the SRI series of observed and simulated Ens-GHM and Ens-CHM runoff for all eight catchments classified into different drought categories and sub-categories. Despite having substantial variation in $R^2$ and NSE values for the SRI series (Table 3), the total number of individual drought events were similar (Table 4). For all eight catchments, 136 individual drought events were identified from the observed data and the ensemble models successfully simulated 133 and 123 drought events for Ens-GHM and Ens-CHM, respectively. However, it is possible that the models simulated roughly the same number of drought events as observed, but differ in timing as when and for how long they really occurred. A detailed analysis is presented in Table 5 in the following Sect. 3.4 showing occurrences and timing of observed and simulated extreme drought events (SRI less than −2).

3.4. Evaluation of drought intensity

Table 5 shows observed and simulated (by Ens-GHM or Ens-CHM) extreme drought events, and the performance of individual models in representing these events in terms of drought intensity and drought duration. In the 'Observed' column, the observed drought events are listed from all eight catchments stating the start and end date, and drought duration in months; unshaded cells are observed droughts, while cells shaded grey are simulated drought events that are not present in the observed record. The respective simulated drought duration is shown in the adjacent columns. The results are colour-coded based on the model difference from the observed drought intensity with shades of green.
indicating under-estimation and brown over-estimation. Cells marked blue denote that the model did not simulate that particular event, while cells marked yellow indicate that the model was not run for that particular catchment.

Substantial variation was seen in the ability of individual CHMs and GHMs to simulate drought intensity. Out of the 27 observed extreme drought events across all eight catchments, Ens-GHM and Ens-CHM failed to identify 3 and 1 drought event(s), respectively. In total 14 such extreme drought events were identified by the ensemble models which were not observed (shaded grey under 'Observed' in Table 5), and 5 of these events were identified as extreme drought events by both Ens-GHM and Ens-CHM.

There is a marked difference across catchments in whether the models over- or under-estimated the intensity of observed droughts, and in the ability to simulate the very occurrence of an observed drought itself. We observed cases of false positive, where a model simulates a drought that never occurred in the observed record, and cases of false negative, where a model fails to simulate a drought that occurred in the observed record for almost all catchments under extreme drought category.

For the Amazon, individual CHMs and GHMs performed similarly, with most models under-estimating drought intensity of 6 of the observed extreme drought events in the catchment. In one case here, an observed event went undetected by 10 out of 17 models. All the individual models poorly simulated hydrological conditions in the Yellow catchment, with 4 extreme drought events simulated that were not observed in reality. Model performance is more nuanced in the other catchments, with some observed droughts within a catchment underestimated and other events over-estimated (Niger, Rhine, Yangtze).

### 3.5. Evaluation of drought duration

For all of the observed events which were simulated in each drought category (i.e. moderate, severe and extreme), we calculated the absolute error in months, of the drought duration simulated by each model. The mean of all the absolute errors under each drought category for a model gave the mean absolute error (MAE) for that model under that particular drought category. Figure 2 shows the MAE for the Ens-GHM and Ens-CHM, for every catchment, under the three drought categories along with their respective mean observed drought duration, and Table 6 presents MAE values for all individual models averaged across all catchments.

The largest total MAE (i.e. the MAE summed across the three drought intensity categories) was seen for the Ens-CHM for the Niger, followed by the Mississippi and Tagus (Fig. 2). For both ensembles, the error is, overall, smallest for duration of extreme droughts across all three drought categories Table 6. The ability of the models to simulate extreme drought durations better than for lower intensity drought durations is confirmed in Fig. 2 showing the smallest magnitude of MAE for extreme category events over half the catchments (5), across the three drought categories by Ens-CHM. For the Ens-GHM, this is the case with half the catchments (4). However, overall, it can be concluded that both the CHMs and GHMs struggle to accurately model drought duration (errors are consistently > 1 month).

### 4. Discussion

The aim of this study was to assess the performance of GHMs and CHMs in simulating observed hydrological drought events using the SRI as a hydrological drought indicator, computed using the observed and simulated runoff from GHMs, CHMs and respective ensemble means. No two drought events are the same and, as such, drought events cannot be judged based on a single characteristic. Here we used three characteristics; the intensity, duration and frequency to compare drought events. We used R² and NSE as indicators to judge the ability of these models to replicate observed monthly runoff and SRI patterns, and in turn identify hydrological drought events.

### 4.1. Comparison of simulated hydrologic conditions between GHMs and CHMs
Previous studies reported that model sensitivity of CHMs and GHMs towards climate variability is comparable, and ensemble models exhibit similarity in the effect of global warming on hydrological indicators (Gosling, Zaherpour, et al. 2017; Hattermann et al. 2017). While $R^2$ and NSE values (Table 2) indicated few similarities between observed monthly runoff and simulated runoff by GHMs; CHMs across all catchments show better reproduction of the observed monthly runoff except WaterGAP3 performing satisfactorily in only two catchments (Rhine and Yellow). Interestingly our results indicate that even though the CHMs are better at reproducing better monthly runoff values, SRI series from models of both scales are comparable (Table 3). The individual model performances also reflected well upon both ensemble models. In general practice, evaluation of hydrological models is based on peak flow estimations, and seldom on a model's ability to capture low flow and no flow conditions (Moriasi et al. 2007). However, for drought identification it is of central importance that a model accurately predicts observed low-flow and no-flow conditions because such conditions play critical roles in determining flow deficit, which accumulated over time can trigger drought events. For the purpose of this study, special attention was given to extreme drought category events (events with drought intensity <-2) as they are usually the most devastating.

Although the $R^2$ and NSE values for the SRI series are less than satisfactory for many individual GHMs and CHMs across almost all catchments, the ensemble means of the two sets of models were better at estimating drought frequency for most of the catchments. There was a marked difference in performance of models when estimating drought intensity across all catchments. One of the key reasons behind this finding is the data used for drought identification, which here is the simulated monthly runoff data. The accuracy of the SRI output is directly proportional to the quality of data used for its calculation (Hayes et al. 1999). Huang et al. (2017) reported that CHMs accurately reproduced monthly runoff, seasonal dynamics, moderate or high flows but simulations of low flows were poor in most catchments. Zaherpour et al. (2018) also found that the majority of GHMs overestimated low flows considerably more than they overestimated high-flows and that GHMs overestimated minimum flow return periods. The majority of the GHMs showed a tendency for overestimating monthly runoff with a wider magnitude range (Veldkamp et al. 2018). Previous studies highlight that this wider spread around ensembles in every catchment is due to the structure of GHMs (Haddeland et al. 2011; Gudmundsson et al. 2012). Physical processes such as transmission losses, having less presence in the GHMs is one main reason for some of the differences between simulated and observed runoff (Gosling and Arnell 2011). In addition, evapotranspiration simulation has been reported to vary widely among the GHMs (Wartenburger et al. 2018).

4.2. Models performance across catchments

Both the ensembles display, overall, a good performance comparative to the individual models and showed comparable outputs (SRI values) despite the GHMs having a wider spread across the ensemble. The ensemble output does not always deliver better performance than individual models, however. For mean monthly runoff estimation, WaterGAP2 showed better results than the Ens-GHM and outperformed other GHMs because the other models overestimated or underestimated the low-flow conditions. The better performance of WaterGAP2 for runoff estimation can be ascribed to its calibration with long-term annual river discharge; however, independently of calibration, the identification of hydrological extremes remains poor among all models. Although the GHMs used the same climate forcing, they used different formulations to compute potential evapotranspiration, which contributes to differences in simulated runoff between the GHMs (Beck et al. 2017).

Our results indicate relatively better performances of GHMs for runoff in the Rhine catchment; however, tendency towards a false positive (Sect. 3.4) can be attributed to dry bias introduced by the choice of potential evapotranspiration formulation for individual models. For example, PCR-GLOBWB consistently appeared near the dry end of Ens-GHM, perhaps because it includes a temperature based evaporation formulation (Hamon) that has been shown to induce a large bias when applied outside its calibration range (Milly and Dunne 2017). In general, for GHMs it is difficult to estimate a drought event at the right time because multiple errors propagate from the inputs (meteorological parameters) and some
GHMs struggle to capture the magnitude and timing of processes like abstraction losses and snowmelt accurately, which is likely to have an impact on drought timing (initiation and duration of drought events).

For CHMs, large biases have been reported in simulating low flow conditions across a majority of studied catchments especially for the Yangtze catchment (Huang et al. 2017). Similarity in performance of both sets of individual models in the Amazon, Lena Tagus, and Yangtze catchments for intensity of extreme drought events (Table 5) can likely be attributed towards the quality of the meteorological data. Some studies have reported WATCH data to be unreliable due to inaccuracies in observed precipitation records, caused by fog/mist (Strauch et al. 2017). Secondly, the inability of individual CHMs to replicate low-flow or no-flow may be due to the choice of objective functions for calibration of the CHMs (Huang et al. 2017). In addition, inaccuracies in low-flow observations may be a factor, or river ice effects in some catchments, which affects estimation of drought conditions. WaterGAP3 among all CHMs comparatively showed the weakest performance, which may be attributed to fewer parameters used for model calibration compared to other CHMs. Performance of the Ens-CHM was better in some catchments than the others, likely due to the fact that different numbers of models were included for each ensemble. Therefore, one of the reasons for comparatively better performance by the Ens-CHM in estimating drought frequency and lower MAE in drought duration for Rhine, Amazon and Lena is an effect of a larger of individual models involved in the ensemble calculations.

5. Conclusion

Our study focused upon the effectiveness of catchment- and global-scale hydrological models to estimate drought conditions at the catchment scale, for 8 large catchments. We found comparably lower performance by most GHMs in simulating observed monthly runoff. R² values for the ensemble mean SRI series varied between 0.18 and 0.95 while the NSE from −0.14 to 0.95 across all catchments. Whilst the Ens-GHM and Ens-CHM produced similar estimates of the total number of drought events across all drought categories, both ensembles struggled to estimate the frequency of drought within the drought categories. Thus, both sets of models have limited ability to simulate the finer, more granular and detailed characteristics of observed droughts. There were marked differences across catchments in estimates of drought intensity by models, and in ability to simulate the occurrence of observed droughts (giving false positives and false negatives in many cases). For both the ensembles, the error is, overall, smallest for duration of extreme droughts across all three drought categories. However, it can also be concluded that both CHMs and GHMs struggled in accurately modelling drought duration for moderate and severe drought categories. We believe that there is still room for improvement in runoff simulations to facilitate drought identification and accurate estimation of drought characteristics. Therefore, we recommend introducing multiple criteria (Krysanova et al., 2018) during model calibration specifically for the evaluation of low- or no-flow simulations.

References


**Tables**

**Table 1** Catchments, gauging station, upstream area of gauging station, and the GHMs and CHMs that comprise ensemble (shaded)
<table>
<thead>
<tr>
<th>Catchment</th>
<th>Upper Amazon</th>
<th>Lena</th>
<th>Upper Mississippi</th>
<th>Upper Niger</th>
<th>Rhine</th>
<th>Tagus</th>
<th>Upper Yangtze</th>
<th>Upper Yellow</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRDC number</td>
<td>3623100</td>
<td>2903430</td>
<td>4119800</td>
<td>1134100</td>
<td>6435060</td>
<td>6113050</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Gauging station</td>
<td>Sao Paulo de Olivenca</td>
<td>Stolb</td>
<td>Alton</td>
<td>Koulikoro</td>
<td>Lobith</td>
<td>Almourol</td>
<td>Cuntan</td>
<td>Tangnaihai</td>
</tr>
<tr>
<td>Upstream drainage area (km²) - GRDC</td>
<td>990,781</td>
<td>2,460,000</td>
<td>444,185</td>
<td>120,000</td>
<td>160,800</td>
<td>67,490</td>
<td>804,859</td>
<td>121,000</td>
</tr>
<tr>
<td>Upstream drainage area (km²) - DDM30</td>
<td>994,469</td>
<td>2,456,513</td>
<td>448,575</td>
<td>121,058</td>
<td>162,092</td>
<td>71,007</td>
<td>851,303</td>
<td>117,543</td>
</tr>
<tr>
<td>Difference between GRDC and DDM30 areas (%)</td>
<td>-0.4</td>
<td>0.1</td>
<td>-1.0</td>
<td>-0.9</td>
<td>-0.8</td>
<td>-5.2</td>
<td>-5.8</td>
<td>2.9</td>
</tr>
</tbody>
</table>

**GHM ensemble (Ens-GHM)**
- CLM (Oleson et al. 2010)
- DBH (Tang et al. 2007)
- H08 (Hanasaki et al. 2008)
- MATSIRO (Pokhrel et al. 2015)
- MPI-HM (Hagemann and Dümenil 1998)
- PCR-GLOBWB (Wada, Wisser and Bierkens 2014)
- WaterGAP2 (Muller Schmied et al. 2016)
- LPJmL (Bondeau et al. 2007)

**Number of GHM simulations** | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 |

**CHM ensemble (Ens-CHM)**
- ECOMAG (Motovilov et al. 1999)
- HBV (Bergstrom and Forsman 1973)
- HYMOD (Boyle 2001)
- HYPE (Lindström et al. 2010)
- mHM (Samaniego, Kumar and Attinger 2010)
- SWAT (Arnold, Allen and Bernhardt 1993)
- SWIM (Krysanova, Müller-Wolflell and Becker 1998)
- VIC (Liang et al. 1994)
- WaterGAP3 (Verzano 2009)

**Number of CHM simulations** | 7 | 5 | 7 | 7 | 7 | 4 | 4 | 6 |

Table 2 $R^2$ and NSE (in parenthesis) values for simulated (individual GHMs and CHMs including both ensembles) versus observed monthly runoff across all eight catchments. Shaded cells denotes where both $R^2$ and NSE values are above 0.7, and marked with x denote that the particular model was not run for the specific catchment.
<table>
<thead>
<tr>
<th>GHMs</th>
<th>Upper Amazon</th>
<th>Lena</th>
<th>Upper Mississippi</th>
<th>Upper Niger</th>
<th>Rhine</th>
<th>Tagus</th>
<th>Upper Yangtze</th>
<th>Upper Yellow</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLM</td>
<td>0.54 (0.21)</td>
<td>0.52 (0.39)</td>
<td>0.60 (-0.40)</td>
<td>0.86 (-2.80)</td>
<td>0.51</td>
<td>0.54</td>
<td>0.29 (0.14)</td>
<td>0.60 (0.37)</td>
</tr>
<tr>
<td>DBH</td>
<td>0.60 (0.48)</td>
<td>0.05 (-0.70)</td>
<td>0.54 (-4.40)</td>
<td>0.75 (-4.50)</td>
<td>0.64</td>
<td>0.78</td>
<td>0.89 (-0.80)</td>
<td>0.78 (-0.20)</td>
</tr>
<tr>
<td>H08</td>
<td>0.78 (-0.40)</td>
<td>0.0 (0.10)</td>
<td>0.62 (-1.00)</td>
<td>0.73 (-1.70)</td>
<td>0.62</td>
<td>0.84</td>
<td>0.59 (-0.46)</td>
<td>0.58 (-0.36)</td>
</tr>
<tr>
<td>MATSIRO</td>
<td>0.49 (0.50)</td>
<td>0.67 (0.57)</td>
<td>0.75 (-1.10)</td>
<td>0.88 (-0.60)</td>
<td>0.56</td>
<td>0.82</td>
<td>0.77 (0.57)</td>
<td>0.76 (0.03)</td>
</tr>
<tr>
<td>MPI-HM</td>
<td>0.57 (0.19)</td>
<td>0.45 (0.37)</td>
<td>0.75 (-0.30)</td>
<td>0.88 (-0.90)</td>
<td>0.71</td>
<td>0.70</td>
<td>0.57 (0.46)</td>
<td>0.49 (0.15)</td>
</tr>
<tr>
<td>PCR-GLOBWB</td>
<td>0.65 (0.34)</td>
<td>0.44 (0.0)</td>
<td>0.72 (0.55)</td>
<td>0.55 (-0.26)</td>
<td>0.79</td>
<td>0.84</td>
<td>0.91 (0.84)</td>
<td>0.70 (-0.50)</td>
</tr>
<tr>
<td>WaterGAP2</td>
<td>0.76 (0.73)</td>
<td>0.62 (0.58)</td>
<td>0.76 (0.74)</td>
<td>0.77 (0.87)</td>
<td>0.88</td>
<td>0.83</td>
<td>0.90 (0.80)</td>
<td>0.64 (0.53)</td>
</tr>
<tr>
<td>LPJML</td>
<td>0.70 (-0.20)</td>
<td>0.50 (0.28)</td>
<td>0.47 (0.40)</td>
<td>0.71 (-1.10)</td>
<td>0.80</td>
<td>0.85</td>
<td>0.93 (0.87)</td>
<td>0.66 (0.62)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHMs</th>
<th>ECOMAG</th>
<th>x</th>
<th>0.93 (0.90)</th>
<th>x</th>
<th>x</th>
<th>x</th>
<th>x</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>HBV</td>
<td>0.83 (0.82)</td>
<td>x</td>
<td>0.72 (0.70)</td>
<td>0.89 (0.77)</td>
<td>0.87 (0.87)</td>
<td>0.92 (0.83)</td>
<td>0.96 (0.94)</td>
<td>0.84 (0.82)</td>
</tr>
<tr>
<td>HYMOD</td>
<td>0.66 (0.66)</td>
<td>x</td>
<td>0.65 (0.64)</td>
<td>0.87 (0.86)</td>
<td>0.82 (0.87)</td>
<td>x</td>
<td>0.79 (0.78)</td>
<td>0.79</td>
</tr>
<tr>
<td>HYPE</td>
<td>x</td>
<td>0.95 (0.94)</td>
<td>x</td>
<td>x</td>
<td>0.90 (0.85)</td>
<td>0.92 (0.91)</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>mHM</td>
<td>0.78 (0.50)</td>
<td>x</td>
<td>0.89</td>
<td>0.87</td>
<td>0.90 (0.85)</td>
<td>x</td>
<td>x</td>
<td>0.86 (0.85)</td>
</tr>
<tr>
<td>SWAT</td>
<td>0.78 (0.62)</td>
<td>x</td>
<td>0.79</td>
<td>0.85</td>
<td>0.94 (0.86)</td>
<td>x</td>
<td>0.94 (0.93)</td>
<td>0.83 (0.93)</td>
</tr>
<tr>
<td>SWIM</td>
<td>0.85 (0.85)</td>
<td>0.92</td>
<td>0.78</td>
<td>0.90</td>
<td>0.91 (0.86)</td>
<td>x</td>
<td>0.95 (0.93)</td>
<td>0.83 (0.93)</td>
</tr>
<tr>
<td>VIC</td>
<td>0.81 (0.8)</td>
<td>0.85</td>
<td>0.75</td>
<td>0.87</td>
<td>0.88</td>
<td>0.88</td>
<td>0.96 (0.85)</td>
<td>0.72 (0.81)</td>
</tr>
<tr>
<td>WaterGAP3</td>
<td>0.35 (0.21)</td>
<td>0.44</td>
<td>0.62</td>
<td>0.59</td>
<td>0.84</td>
<td>0.79</td>
<td>x</td>
<td>0.79</td>
</tr>
<tr>
<td>Ens-GHM</td>
<td>0.81 (0.24)</td>
<td>0.58</td>
<td>0.87</td>
<td>0.86</td>
<td>0.84</td>
<td>0.87</td>
<td>0.83</td>
<td>0.80 (0.80)</td>
</tr>
<tr>
<td>Ens-CHM</td>
<td>0.78 (0.75)</td>
<td>0.92</td>
<td>0.89</td>
<td>0.90</td>
<td>0.96</td>
<td>0.94</td>
<td>0.97</td>
<td>0.84 (0.84)</td>
</tr>
</tbody>
</table>

Table 3 $R^2$ and NSE (in parenthesis) values for simulated (individual GHMs and CHMs including both ensembles) versus observed SRI across all eight catchments. Shaded cells denotes where both $R^2$ and NSE values are >0.70 and marked with x denote that the particular model was not run for the specific catchment.
Table 4 Number of drought events identified from SRI series of observed and simulated Ens-GHM and Ens-CHM runoff for all eight catchments, classified into drought categories (moderate, severe and extreme) based on drought intensity.
<table>
<thead>
<tr>
<th>Catchments (1)</th>
<th>Drought category (→)</th>
<th>Moderate (-1 to -1.49)</th>
<th>Severe (-1.5 to -1.99)</th>
<th>Extreme (-2 to -2.49)</th>
<th>(-2.5 to -2.99)</th>
<th>(-3 or Below)</th>
<th>Total individual drought events (-1 or Below)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper Amazon</td>
<td>Ens-GHM</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Ens-CHM</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Obs</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>Lena</td>
<td>Ens-GHM</td>
<td>7</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Ens-CHM</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Obs</td>
<td>11</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>Upper Mississippi</td>
<td>Ens-GHM</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td></td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Ens-CHM</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Obs</td>
<td>9</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
<td>14</td>
</tr>
<tr>
<td>Upper Niger</td>
<td>Ens-GHM</td>
<td>10</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td></td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>Ens-CHM</td>
<td>9</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Obs</td>
<td>11</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td></td>
<td>22</td>
</tr>
<tr>
<td>Rhine</td>
<td>Ens-GHM</td>
<td>7</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td></td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Ens-CHM</td>
<td>8</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td></td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Obs</td>
<td>10</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td></td>
<td>19</td>
</tr>
<tr>
<td>Tagus</td>
<td>Ens-GHM</td>
<td>11</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Ens-CHM</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Obs</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>Upper Yangtze</td>
<td>Ens-GHM</td>
<td>12</td>
<td>5</td>
<td>5</td>
<td></td>
<td></td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>Ens-CHM</td>
<td>12</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td></td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>Obs</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td></td>
<td>21</td>
</tr>
<tr>
<td>Upper Yellow</td>
<td>Ens-GHM</td>
<td>6</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td></td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Ens-CHM</td>
<td>9</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td></td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Obs</td>
<td>7</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td>16</td>
</tr>
<tr>
<td><strong>Total individual drought events across all catchments</strong></td>
<td>Ens-GHM</td>
<td>64</td>
<td>45</td>
<td>18</td>
<td>5</td>
<td>1</td>
<td>133</td>
</tr>
<tr>
<td></td>
<td>Ens-CHM</td>
<td>59</td>
<td>38</td>
<td>17</td>
<td>9</td>
<td>2</td>
<td>123</td>
</tr>
<tr>
<td></td>
<td>Obs</td>
<td>69</td>
<td>40</td>
<td>21</td>
<td>4</td>
<td>2</td>
<td>13</td>
</tr>
</tbody>
</table>
Table 6 MAE (across all catchments) in simulating drought duration by individual models, separately shaded for each drought category considering GHMs and CHMs simultaneously with shades of grey, darker denoting higher values.
### Drought category (SRI level) →

<table>
<thead>
<tr>
<th></th>
<th>Moderate (-1 to -1.49)</th>
<th>Severe (-1.5 to -1.99)</th>
<th>Extreme (-2 or Below)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Models (1)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GHMs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLM</td>
<td>3.69</td>
<td>4.53</td>
<td>2.83</td>
</tr>
<tr>
<td>DBH</td>
<td>2.3</td>
<td>3.5</td>
<td>5.29</td>
</tr>
<tr>
<td>H08</td>
<td>2.53</td>
<td>3.21</td>
<td>6.13</td>
</tr>
<tr>
<td>MATSIRO</td>
<td>5.62</td>
<td>5.78</td>
<td>7.47</td>
</tr>
<tr>
<td>MPI-HM</td>
<td>3.78</td>
<td>5.15</td>
<td>3.58</td>
</tr>
<tr>
<td>PCR-GLOBWB</td>
<td>3.01</td>
<td>4.27</td>
<td>3.64</td>
</tr>
<tr>
<td>WaterGAP2</td>
<td>2.69</td>
<td>3.91</td>
<td>2.77</td>
</tr>
<tr>
<td>LPJML</td>
<td>2.77</td>
<td>2.65</td>
<td>6.09</td>
</tr>
<tr>
<td>CHMs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECOMAG</td>
<td>3.63</td>
<td>2</td>
<td>5.33</td>
</tr>
<tr>
<td>HBV</td>
<td>3.89</td>
<td>4.78</td>
<td>3.16</td>
</tr>
<tr>
<td>HYMOD</td>
<td>3</td>
<td>4.35</td>
<td>5.85</td>
</tr>
<tr>
<td>HYPE</td>
<td>3.29</td>
<td>2.47</td>
<td>2.11</td>
</tr>
<tr>
<td>mHM</td>
<td>3.5</td>
<td>6.15</td>
<td>2.36</td>
</tr>
<tr>
<td>SWAT</td>
<td>3.11</td>
<td>3.51</td>
<td>3.78</td>
</tr>
<tr>
<td>SWIM</td>
<td>2.8</td>
<td>4.28</td>
<td>2.38</td>
</tr>
<tr>
<td>VIC</td>
<td>2.92</td>
<td>4.26</td>
<td>2.66</td>
</tr>
<tr>
<td>WaterGAP3</td>
<td>2.92</td>
<td>3.05</td>
<td>4.61</td>
</tr>
<tr>
<td>Ens-GHM</td>
<td>3.24</td>
<td>3.9</td>
<td>3</td>
</tr>
<tr>
<td>Ens-CHM</td>
<td>2.94</td>
<td>4.18</td>
<td>2.45</td>
</tr>
</tbody>
</table>

### Figures

![Map showing eight study catchments](map.png)

**Figure 1**

Location of the eight study catchments labelled as (1) Upper Amazon, (2) Lena, (3) Upper Mississippi, (4) Upper Niger, (5) Rhine, (6) Tagus, (7) Upper Yangtze, and (8) Upper Yellow. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning
the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

Figure 2

MAE in simulating drought under each drought category by ensemble model (Ens-GHM and Ens-CHM) for all eight catchments along with respective mean observed drought duration