Gudmundsson, Lukas; Tallaksen, Lena M.; Stahl, Kerstin; Clark, Douglas B.; Dumont, Egon; Hagemann, Stefan; Bertrand, Nathalie; Gerten, Dieter; Heinke, Jens; Hanasaki, Naota; Voss, Frank; Koirala, Sujan. 2012 Comparing large-scale hydrological model simulations to observed runoff percentiles in Europe. *Journal of Hydrometeorology*, 13 (2). 604-620. 10.1175/JHM-D-11-083.1

Comparing Large-Scale Hydrological Model Simulations to Observed Runoff Percentiles in Europe

LUKAS GUDMUNDSSON,* LENA M. TALLAKSEN,* KERSTIN STAHL,[+] DOUGLAS B. CLARK,[#] EGON DUMONT,[#] STEFAN HAGEMANN,[@] NATHALIE BERTRAND,[&] DIETER GERTEN,[**] JENS HEINKE,[**] NAOTA HANASAKI,[++] FRANK VOSS,[##] AND SUJAN KOIRALA[@@]

* *Department of Geosciences, University of Oslo, Oslo, Norway*
[+] *Institute of Hydrology, University of Freiburg, Freiburg, Germany*
[#] *Centre for Ecology and Hydrology, Wallingford, United Kingdom*
[@] *Max Planck Institute for Meteorology, Hamburg, Germany*
[&] *Laboratoire de Météorologie Dynamique, Paris, France*
[**] *Potsdam Institute for Climate Impact Research, Potsdam, Germany*
[++] *National Institute for Environmental Studies, Tsukuba, Japan*
[##] *Center for Environmental Systems Research, University of Kassel, Kassel, Germany*
[@@] *Department of Mechanical and Environmental Informatics, Tokyo Institute of Technology, Yokohama, Japan*

(Manuscript received 30 June 2011, in final form 12 November 2011)

ABSTRACT

Large-scale hydrological models describing the terrestrial water balance at continental and global scales are increasingly being used in earth system modeling and climate impact assessments. However, because of incomplete process understanding and limits of the forcing data, model simulations remain uncertain. To quantify this uncertainty a multimodel ensemble of nine large-scale hydrological models was compared to observed runoff from 426 small catchments in Europe. The ensemble was built within the framework of the European Union Water and Global Change (WATCH) project. The models were driven with the same atmospheric forcing data. Models were evaluated with respect to their ability to capture the interannual variability of spatially aggregated annual time series of five runoff percentiles—derived from daily time series—including annual low and high flows. Overall, the models capture the interannual variability of low, mean, and high flows well. However, errors in the mean and standard deviation, as well as differences in performance between the models, became increasingly pronounced for low runoff percentiles, reflecting the uncertainty associated with the representation of hydrological processes, such as the depletion of soil moisture stores. The large spread in model performance implies that any single model should be applied with caution as there is a great risk of biased conclusions. However, this large spread is contrasted by the good overall performance of the ensemble mean. It is concluded that the ensemble mean is a pragmatic and reliable estimator of spatially aggregated time series of annual low, mean, and high flows across Europe.

1. Introduction

Large-scale hydrological models have proved to be valuable tools for assessing fluctuations in terrestrial water stores and fluxes on continental and global scales (e.g., Dirmeyer 2011; Dirmeyer et al. 2006; Milly et al. 2005). To date, models describing the terrestrial water balance have been developed by different communities and parallel terminologies, and modeling philosophies have emerged (Haddeland et al. 2011). Among the most commonly used terms are global hydrology models (GHMs), focusing on closing the water balance for the purpose of water resource assessment, and land surface models (LSMs) that were historically developed to provide lower boundary conditions for atmospheric circulation models with a focus on the surface water and energy balances. However, many models (both GHMs and LSMs) share essentially the same conceptualization of the water fluxes (Haddeland et al. 2011). Thus, all models that

Corresponding author address: Lukas Gudmundsson, Department of Geosciences, University of Oslo, P.O. Box 1047, Blindern, 0316 Oslo, Norway.
E-mail: lukas.gudmundsson@geo.uio.no

DOI: 10.1175/JHM-D-11-083.1

resolve the terrestrial part of the water cycle at global and continental scales will in the following be referred to as large-scale hydrological models.

Various efforts have been made to evaluate large-scale hydrological models, including macroscale studies that compare observed and modeled continental river discharge (e.g., Balsamo et al. 2009; Decharme and Douville 2007; Gerten et al. 2004; Hagemann et al. 2009), as well as studies with relatively detailed spatial and temporal resolution on continental and global scales (e.g., Döll et al. 2003; Hunger and Döll 2008; Troy et al. 2008; Widén-Nilsson et al. 2009; Stahl et al. 2011). Generally the focus is on evaluating a single model, possibly with a new representation of certain processes. Another approach is followed by large model intercomparison exercises that focus less on model evaluation by comparison to observations, and rather more on identifying differences in model dynamics. Examples are the Project for Intercomparison of Land Surface Parameterization Schemes (PILPS) (Henderson-Sellers et al. 1995), the Global Soil Wetness Project (GSWP) (Oki et al. 1999; Dirmeyer et al. 2006; Dirmeyer 2011), and the Water Model Intercomparison Project (WaterMIP) (Haddeland et al. 2011). In general, these studies conclude that there are large differences between the models, which may be caused by incomplete process understanding, different parameter estimates, and imperfect atmospheric forcing data.

Several multimodel evaluation studies not only compare individual models to observations, but also investigate the behavior of the mean of all models, commonly referred to as the ensemble mean. Being widely applied in atmospheric science (e.g., Reichler and Kim 2008; Hagedorn et al. 2005; Palmer et al. 2004), so-called ensemble techniques are also increasingly used in the evaluation of large-scale hydrological models. So far most studies that employed ensemble techniques in the context of large-scale hydrological modeling have focused on the mean annual cycle of monthly discharge from large, continental-scale river basins. Generally these studies show that the uncertainty in river discharge introduced by the use of different atmospheric forcing models (Nohara et al. 2006; Hagemann and Jacob 2007) and different land surface schemes (Materia et al. 2010) can be reduced by ensemble techniques. Several studies have compared soil moisture simulations from the GSWP to monthly observations from a global observation network (e.g., Gao and Dirmeyer 2006; Guo and Dirmeyer 2006; Guo et al. 2007). These studies assessed, amongst others, the ability of the ensemble members to capture mean values, the phasing of the annual cycle, and the interannual variability, showing that the ensemble mean was closer to the observations than most participating models (Gao and Dirmeyer 2006; Guo and Dirmeyer 2006; Guo et al. 2007).

Relatively few studies evaluated large-scale hydrological models with respect to their ability to capture hydrological extremes, and consequently no standard procedure has been established. Most available studies have focused on the analysis of daily river discharge, partly because the observational time window is longer, and partly because this increases the number of observations, which renders model validation more reliable. Lehner et al. (2006), for example, evaluated the ability of the Water—Global Analysis and Prognosis (WaterGAP) model to capture the average magnitude and return periods of annual flood and drought statistics in Europe based on daily data. They concluded that the model captured average annual low and high flows reasonably well, but had a tendency to overestimate the return periods of extreme events. Similarly, Hirabayashi et al. (2008) compared the estimated return periods of seven disastrous floods around the globe to the results from a global offline simulation with daily resolution and concluded that the return period of the simulated events compared reasonably well to the observed values. However, Hirabayashi et al. (2008) also pointed out that a statistically reliable evaluation of model performance with respect to extremes on large (global) scales is hampered by the scarcity of long-term observations. Recently, Feyen and Dankers (2009) compared the return periods of selected low-flow statistics derived from observed and simulated daily data from rivers across Europe, highlighting deficiencies of the simulations in the frost season. In an accompanying study, Dankers and Feyen (2009) reported that the simulations captured peak flows from large river basins quite well, whereas the performance was at times poor in small catchments. It shall be noted that all the above studies are based on data from the Global Runoff Data Centre (GRDC; http://grdc.bafg.de/), which provides a collection of observations from relatively large river basins.

The main focus of the studies summarized above was to investigate the impacts of climate change on hydrological variables. Therefore, in these studies model evaluation was only regarded as a prerequisite to further analysis and thus often received little attention. In contrast, Stahl et al. (2011) focused solely on the evaluation of simulated runoff (7-day running mean) from a regional climate model in Europe with respect to 19 different anomaly levels, ranging from low to high flows. Comparing event dynamics and interannual variability, the lowest agreement was found for the dry anomalies and that model performance was best for moderately wet anomalies.

Studies evaluating multimodel ensembles have focused mainly on mean water balance components and

rarely on hydrological extremes. This is partially due to limits from the temporal resolution of the commonly stored summary statistics (e.g., monthly means) and relatively short integrations that preclude a proper analysis of extremes. To overcome such limitations, a major effort was made within the European Framework Project Water and Global Change (WATCH; www.eu-watch.org) to create a multimodel ensemble of large-scale hydrological models with summaries available on a daily resolution. The main objective of this study is to get first insights into the ability of the WATCH multimodel ensemble to capture hydrological extremes, with respect to both their magnitude and interannual variability on a large, continental scale.

The observed data used in this model evaluation exercise comprise time series from a large number of small, nearly natural catchments in Europe that are not nested (see section 2b for details). In contrast to discharge from large river basins, which are often strongly influenced by human activities (Döll et al. 2009), observations from small undisturbed catchments are often more likely to represent the natural system behavior. Further, discharge observations from large rivers are bound to suffer from small sample sizes, as there are a small number of continental-scale drainage basins. A small sample size increases the risk that observation errors lead to biased results in the model evaluation. It is also interesting to note that the mathematical structure underlying individual grid cells in large-scale models is often comparable to the model structure of so-called lumped catchment models, which are commonly used to model streamflow from small catchments (see Clark et al. 2008, 2011b for a comprehensive overview). One example from the current ensemble is the Global Water Availability Assessment (GWAVA) model (Meigh et al. 1999), which uses the commonly applied lumped Probability Distributed Model (Moore 2007, 1985) to parameterize gridcell processes.

However, the use of streamflow observations from small catchments to evaluate large-scale hydrological models raises several issues. Streamflow observations are prone to measurement errors (e.g., Di Baldassarre and Montanari 2009) that are known to affect the calibration of hydrological models (e.g., Reitan and Petersen-Øverleir 2009; McMillan et al. 2010) and consequently also the performance assessments of large-scale hydrological models. Strategies to incorporate these observational errors into predictive uncertainty, however, are not well established and are subject to ongoing research (e.g., Kavetski et al. 2006; Renard et al. 2010). The model parameters at each grid cell, derived from large-scale maps, are unlikely to perfectly characterize the true catchment properties and this may result in large discrepancies between observed and simulated runoff at the gridcell scale. It is important to note that model parameters such as vegetation and soil properties exhibit high spatial variability (Duan et al. 2006). Maps used to derive model parameters are therefore highly uncertain and parameter estimates based on different map sources may hence result in significant differences in simulated system behavior (Teuling et al. 2009).

One approach to minimize the effect of the large uncertainty in model parameters at the gridcell scale is to focus on spatially aggregated system behavior. For example, in atmospheric sciences it is common to investigate time series of variables that have been averaged over large spatial areas. One example is the assessment of time series of mean global temperature (e.g., Hansen et al. 2006; Macadam et al. 2010). This study adapts this strategy as it agrees with the main objective, which is to evaluate the ability of the WATCH multimodel ensemble to capture key aspects of the interannual variability of runoff in Europe. Importantly, we use data from the level of the grid cell and small catchments, and then aggregate to the larger scale, rather than just using data from continental-scale catchments, for the reasons outlined above.

The remainder of this article is organized as follows: first, the multimodel ensemble of nine large-scale hydrological models and the observed streamflow data are introduced. In the methods section, statistical summaries that represent low, mean, and high flows over large (continental) scales are defined, followed by the introduction of three performance metrics. The results of the analysis are then presented and discussed. The paper concludes with comments on the ability of the multimodel ensemble to simulate European, large-scale hydrology, with special emphasis on low and high river flows.

2. Models and observations

a. Individual models and ensemble mean

Table 1 lists the nine models that were considered in this study and summarizes their evapotranspiration, snow, and runoff schemes. Table 2 briefly summarizes the principles underlying their subsurface parameterization and provides key references. Gridcell runoff is simulated from the water balance

$$\frac{dS}{dt} = P - E - Q_s - Q_{\mathrm{sb}}, \tag{1}$$

where P is precipitation, E evaporation, Q_s surface runoff, Q_{sb} subsurface runoff, and dS/dt denotes changes in storage. Here the total runoff $Q = Q_s + Q_{\mathrm{sb}}$ is derived for each grid cell.

The structure underlying most of the models is illustrated in Fig. 1, indicating the different conceptual storages

TABLE 1. Overview of the participating models and their main characteristics. Models written in italic are classified as LSMs. Surface runoff (Q_s) is in all instances modeled as saturation or infiltration excess or both; the following abbreviations refer to approaches to parameterize subgrid variability: ARNO (Todini 1996), improved ARNO (Dümenil and Todini 1992), and Probability Distributed Model (PDM) (Moore 1985). Subsurface runoff (Q_{sb}) is either modeled as a function of soil moisture $Q_{sb} = Q_d = f(S_{soil})$ or groundwater $Q_{sb} = f(S_{gw})$, where $f(S)$ denotes linear or nonlinear model specific functions (''Richards'': N-layer approximation of Richards equation). Adapted from Haddeland et al. (2011).

Model name	Time step	Evapotranspiration	Snow	Runoff scheme
GWAVA	Daily	Penman–Monteith	Degree day	Q_s: PDM $Q_{sb} = f(S_{gw})$
H08	6 h	Bulk approach	Energy balance	Q_s: Saturation excess $Q_{sb} = Q_d = f(S_{soil})$
HTESSEL	1 h	Penman–Monteith	Energy balance	Q_s: ARNO $Q_{sb} = Q_d = f(S_{soil})$, Richards
JULES	1 h	Penman–Monteith	Energy balance	Q_s: Infiltration excess $Q_{sb} = Q_d = f(S_{soil})$, Richards
LPJmL	Daily	Priestley–Taylor	Degree day	Q_s: Saturation excess $Q_{sb} = Q_d = f(S_{soil})$
MATSIRO	1 h	Bulk approach	Energy balance	Q_s: Infiltration and saturation excess $Q_{sb} = f(S_{gw})$
MPI-HM	Daily	Thornthwaite	Degree day	Q_s: Improved ARNO $Q_{sb} = Q_d = f(S_{soil})$
ORCHIDEE	15 min	Bulk approach	Energy balance	Q_s: Infiltration excess $Q_{sb} = Q_d = f(S_{soil})$
WaterGAP	Daily	Priestley–Taylor	Degree day	Q_s: Saturation excess $Q_{sb} = f(S_{gw})$

and fluxes. Note that not every model considers all elements of this generalized architecture and the models differ in their representation of the processes.

Despite large differences in the description of subsurface processes, all models simulate Q_s (water leaving the grid cell on the surface) and Q_{sb} (water leaving the grid cell below the surface). In Fig. 1, Q_{sb} represents the outflow from groundwater storage (S_{gw}); however, not all models simulate S_{gw}. In such cases, the water draining from the lowest soil layer (Q_d) is used to represent subsurface runoff ($Q_{sb} = Q_d$; Table 1).

The simulation setup is, except for the time window and the temporal resolution of the stored output, identical to that described by Haddeland et al. (2011). Model runs for the time window 1963–2000, with output data available at daily time steps, were considered. The runs were preceded by a spinup period of 5 yr. All model simulations were carried out on the 0.5° grid defined by the Climate Research Unit (CRU) of the University of East Anglia global land mask. No effort was made to harmonize model parameters, but the models were forced by the same meteorological data—the so-called WATCH Forcing Data (WFD; Weedon et al. 2010, 2011). The WFD are based on the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40; Uppala et al. 2005) interpolated to the 0.5° grid defined by the CRU land mask and then adjusted for elevation differences. Air temperature is bias corrected and shortwave radiation adjusted according to cloud cover and aerosol loading using the CRU data (Mitchell and Jones 2005; New et al. 1999, 2000). Precipitation is bias corrected using the Global Precipitation Climatology Centre full product (GPCCv4) data (Rudolf and Schneider 2005; Schneider et al. 2010; Fuchs 2009) and undercatch corrected (Adam and Lettenmaier 2003). The simulations assumed ''naturalized'' conditions, which means that direct anthropogenic effects such as dams and water abstraction were not included. This is consistent with the use of observations from undisturbed catchments.

Besides the runoff simulations of individual models, this study also analyzes the arithmetic mean of the runoff simulations of the multimodel ensemble. This mean will in the following be referred to as the ''ensemble mean'' (or ENSEMBLE) and is treated as a separate model throughout the analysis.

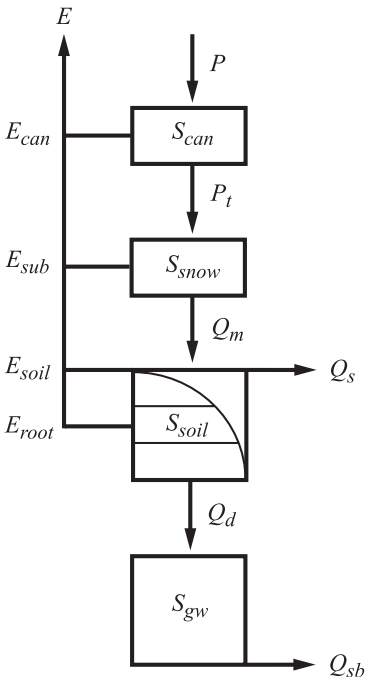
b. Observations

Daily streamflow series from 426 near-natural and spatially independent headwater catchments across Europe were considered. The records cover the time period 1963–2000 and originate from the European Water Archive (EWA)—a database assembled by the European Flow Regimes from International Experimental and Network Data (Euro-FRIEND; http://ne-friend.bafg.de/servlet/is/7413/) project. The EWA is accessible to active members of FRIEND and stored at the GRDC, which also manages data requests. The EWA dataset was recently updated (Stahl et al. 2008)

TABLE 2. Brief descriptions of the nine large-scale hydrological models.

GWAVA	The GWAVA model (Meigh et al. 1999) is based on the PDM rainfall runoff model with an analytic approximation of the subgrid variability of soil moisture (Moore 2007, 1985). The subsurface features several conceptual storages representing the unsaturated and the saturated zone. Two additional storages are used for routing of water via fast pathways to the cell outlet.
H08	H08 is based on a simple bucket model (Manabe 1969) that has been updated to include a nonlinear parameterization of subsurface runoff (Hanasaki et al. 2008).
HTESSEL	The water movement within a grid cell of HTESSEL (Balsamo et al. 2009) is based on the ARNO infiltration excess scheme (Todini 1996), which parameterizes subgrid variability of soil moisture as a function of the standard deviation of orography. HTESSEL features a detailed approximation of the unsaturated zone, which is described by several layers and soil moisture is calculated using an approximation of Richards equation.
JULES	JULES uses four soil layers to calculate subsurface hydrology, with vertical fluxes of water calculated from a solution of Richards equation including root water uptake (Best et al. 2011; D. B. Clark et al. 2011).
LPJmL	LPJmL was developed to model global vegetation dynamics and their coupling to carbon and water fluxes. It features a five-layer soil parameterization where each layer is parameterized as a bucket model that produces saturation excess runoff. Soil moisture responds not only to atmospheric moisture demand, but also to vegetation dynamics (Fader et al. 2010; Bondeau et al. 2007), and new parameterizations as in S. Schaphoff (2011, personal communication).
MATSIRO	The subsurface hydrology of MATSIRO (Takata et al. 2003) is represented by vertical movement of infiltrated moisture through unsaturated soil layers underlain by a groundwater reservoir. The saturated and unsaturated soil zones are in dynamic coupling though an exchange of groundwater recharge, and baseflow is generated from the groundwater reservoir (Koirala et al. 2011a,b, manuscripts submitted to *J. Geophys. Res.*).
MPI-HM	MPI-HM (Hagemann and Dümenil 1998; Roeckner et al. 2003) parameterizes subgrid variability using an updated ARNO scheme (Hagemann and Dümenil Gates 2003) that uses high-resolution soil and orography data to derive the fraction of saturated area of each grid cell. Subsurface runoff is computed as a simple function of storage.
ORCHIDEE	ORCHIDEE has a complex hydrological infiltration scheme (d'Orgeval et al. 2008) that solves the vertical movement of water in the soil using the Fokker–Planck equation with Van Genuchten–Mualem parameters. Subsurface runoff considers orography and surface runoff may reinfiltrate in the same grid cell if the slope is small.
WaterGAP	WaterGAP is based on a series of conceptual storages including surface water bodies, soil moisture, and groundwater (Alcamo et al. 2003). WaterGAP is the only ensemble member that does not solely rely on input maps for parameter estimation, but also undergoes a very limited calibration procedure (see Hunger and Döll 2008 for details).

and further complemented by partners from the WATCH project and is described in detail in Stahl et al. (2010). Observed streamflow ($m^3\ s^{-1}$) was converted into equivalent runoff rates (mm day^{-1}), which we will refer to as observed runoff. Catchment boundaries and mean catchment elevation, based on a high-resolution digital elevation model, were derived from the pan-European river and catchment database Catchment Characterisation and Modeling 2 (CCM2; Vogt et al. 2007). The majority of the catchments have an area that is considerably smaller (median catchment size 258 km^2) than the size of the 0.5° model grid cells (Fig. 2). The size of a grid cell varies, depending on the latitude, between 1065 km^2 (at 70°N) and 2387 km^2 (at 39.5°N). To compare observations and simulations, each gauging station was assigned to the corresponding grid cell and, in cases with more than one station per grid cell, the area-weighted average of the series was used. This procedure resulted in 298 grid cells with observed runoff series. Figure 3 shows the spatial distribution of the grid cells as well as the boundaries of the corresponding catchments. The spatial density and extent of observed runoff were limited



State variables (storages):

S_{can} : Canopy storage
S_{snow} : Snow storage
S_{soil} : Soil storage (unsaturated zone)
S_{gw} : Groundwater storage (saturated zone)

Flux variables:

P : Precipitation
P_t : Throughfall
Q_m : Snowmelt
Q_d : Drainage to groundwater (includes capillary flux)
Q_s : Surface runoff
Q_{sb} : Subsurface runoff
E : Evapotranspiration
E_{can} : Evapotranspiration from canopy
E_{sub} : Sublimation from snow
E_{soil} : Evaporation from soil
E_{root} : Root water uptake

FIG. 1. Simplified conceptualization of state (storage) and flux variables involved in runoff generation. Not all variables are considered in each model. See Table 1 for an overview of the models.

by data availability, with most stations located in central Europe. The median elevation of the catchments is 525 MSL and the average elevation of the selected grid cells is 439 MSL. This systematic lower gridcell elevation may be a result of small headwater catchments being located in higher altitudes, while the grid cells reflect the average elevation of larger areas.

3. Methods

Observed and modeled daily runoff series were aggregated into time series of annual runoff percentiles at five different percentile levels. Low flows are characterized by series of annual 5 percentiles (Q_5), mean flows by series of annual 50 percentiles (Q_{50}; i.e., annual medians), and high flows by series of 95 percentiles (Q_{95}). The notion of percentiles follows the statistical convention commonly used in the United States (representing cumulative or nonexceedance frequencies) and not the hydrological one commonly used in Europe (representing exceedance frequencies). Extreme high and low values are often prone to measurement errors (Laaha and Blöschl 2007) and, therefore, this study excludes annual maximum and minimum values. To provide insights into the entire flow range, two additional percentile series were introduced to characterize moderately low (Q_{25}) and high (Q_{75}) values. It can be argued that this set of five percentile series is sufficient to characterize the overall flow range, as previous results have demonstrated that the information gain by introducing additional percentile levels is limited for continental-scale analysis (Gudmundsson et al. 2011a). This procedure resulted in a set of five time series of annual runoff percentiles for both observed and modeled runoff in each grid cell. The time series from the individual grid cells were then aggregated using the median to obtain one time series for each runoff percentile, resulting in a total of five time series of average percentile values for both simulated and observed values.

Model performance was assessed with respect to three criteria. First, the models' ability to capture the temporal patterns of the interannual variability of the runoff percentiles was quantified using R^2—the squared Pearson-correlation coefficient. Second, the models' ability to capture the average runoff magnitude was characterized using the relative difference in the long-term mean (i.e., the bias)

$$\Delta\mu = \frac{\mu_{\text{mod}} - \mu_{\text{obs}}}{\mu_{\text{obs}}}, \tag{2}$$

where μ denotes the arithmetic mean and the subscripts obs and mod indicate observed and modeled values, respectively. Third, the models' ability to capture the amplitude of the interannual variability was quantified

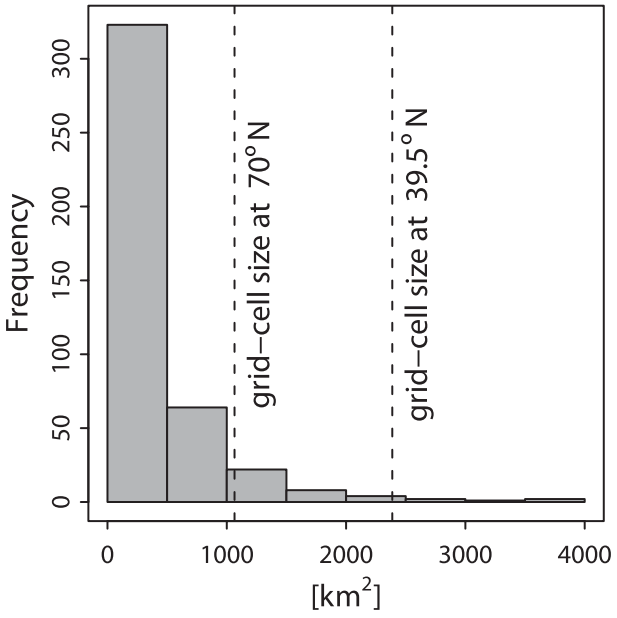


FIG. 2. Histogram of catchment areas. The vertical dashed lines indicate the range of the size of a 0.5° × 0.5° grid cell between the extremes at the lowest and highest latitudes of the spatial domain.

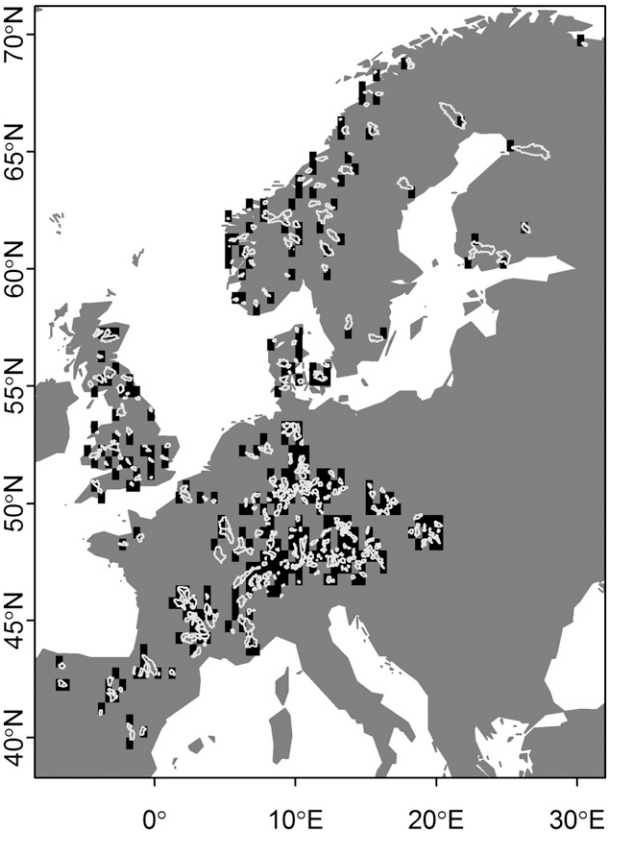


FIG. 3. Map showing grid cells with observations and associated catchment boundaries.

by the relative difference in standard deviation of the annual time series

$$\Delta\sigma = \frac{\sigma_{\text{mod}} - \sigma_{\text{obs}}}{\sigma_{\text{obs}}}, \tag{3}$$

where σ denotes standard deviation.

Finally, the relative merits of the individual models were assessed by ranking their performance (e.g., Gleckler et al. 2008; Macadam et al. 2010). A ranking procedure allows for an easy combination of several performance metrics, even if they have different scales (such as R^2, $\Delta\mu$, and $\Delta\sigma$). However, a ranking will not allow insights into the "absolute performance" of the models; rather it allows the models to be ordered from the one that is on average closest to the observations (rank 1) to the most distant one.

To do an overall ranking, the values of the three performance metrics for each model and runoff percentile were summarized in Table 3, where the columns represent the models and the rows the performance metrics derived for each runoff percentile. First, the values of each row were ranked such that the model being closest to the optimal value (0 for $\Delta\mu$ and $\Delta\sigma$; 1 for R^2) gets rank 1, the next model rank 2, and so on. This procedure results in a new matrix of ranks, which is then summarized to achieve an overall ranking. First, the sum of ranks for each model (columns) is determined and the models are then ordered from the best-performing model (lowest rank sums) to the model with lowest performance (highest rank sums). Finally, the rank sums are replaced by the overall ranks.

Similarly, the percentiles can be ranked by reorganizing the initial matrix in such a way that the columns represent the runoff percentiles and the rows represent the performance of each model. The percentile with the highest rank will then be the percentile value that is overall best reproduced by the models. A similar set of performance metrics was used in a parallel study (Gudmundsson et al. 2011c, manuscript submitted to *Water. Resour. Res.*) to quantify the models' ability to capture the mean annual cycle of runoff with respect to different hydroclimatic regimes as well as the uncertainty of the associated spatial patterns.

4. Results

Figure 4 displays the spatially aggregated time series of observed and modeled runoff percentiles and Fig. 5 shows the mean value of each series. Overall, the models capture the temporal evolution of the interannual variability of observed runoff well. However, there are differences in the mean value as well as in the amplitude of

TABLE 3. Model performance for the five runoff percentiles as measured by the correlation coefficient (R^2), the relative difference in mean ($\Delta\mu$), and the relative difference in standard deviation ($\Delta\sigma$). The best models in each row are in boldface. The column labeled percentile median provides median model performance for each runoff percentile. The three rows labeled model median give median performance for each model. The last row ranks the performance of the models. The last column ranks the overall model performance for a given runoff percentile.

	GWAVA	H08	HTESSEL	JULES	LPJmL	MATSIRO	MPI-HM	ORCHIDEE	WaterGAP	ENSEMBLE	Percentile median	Percentile rank
							Q_{95}					
R^2	0.86	0.80	0.79	0.81	0.82	0.69	0.84	0.60	0.86	**0.89**	0.82	
$\Delta\mu$	−0.19	0.25	−0.28	−0.20	0.33	−0.39	0.27	0.24	**−0.02**	−0.09	−0.05	1
$\Delta\sigma$	−0.22	1.05	0.06	−0.20	0.12	−0.08	0.07	**−0.03**	0.05	−0.09	0.01	
							Q_{75}					
R^2	0.79	0.71	0.78	**0.89**	0.66	0.59	0.80	0.78	0.80	0.85	0.79	
$\Delta\mu$	−0.12	−0.56	−0.11	**0.00**	−0.17	−0.13	−0.32	−0.22	−0.36	−0.14	−0.16	3
$\Delta\sigma$	−0.10	−0.95	−0.07	0.06	−0.26	−0.12	−0.21	**0.01**	−0.49	−0.14	−0.13	
							Q_{50}					
R^2	0.83	0.81	0.75	**0.89**	0.71	0.61	0.69	0.77	0.79	0.86	0.78	
$\Delta\mu$	−0.09	−0.36	−0.05	**−0.03**	−0.23	0.20	−0.74	−0.41	−0.28	−0.09	−0.16	2
$\Delta\sigma$	0.22	−0.53	0.28	0.59	0.12	0.40	−0.21	0.15	−0.32	**0.10**	0.13	
							Q_{25}					
R^2	0.87	0.56	0.86	**0.90**	0.82	0.77	0.62	0.77	0.81	0.89	0.81	
$\Delta\mu$	−0.11	−0.43	−0.08	−0.25	−0.77	0.55	−0.96	−0.69	−0.09	−0.10	−0.18	4
$\Delta\sigma$	0.19	−0.04	0.40	0.71	0.19	0.71	−0.98	−0.36	−0.43	0.13	0.16	
							Q_5					
R^2	0.72	0.59	0.87	0.81	0.57	0.74	0.65	0.68	0.79	**0.85**	0.73	
$\Delta\mu$	−0.10	−0.59	**0.01**	−0.41	−0.97	1.04	−0.95	−0.80	0.16	−0.08	−0.26	5
$\Delta\sigma$	0.13	0.19	0.54	0.61	−0.94	1.23	−0.99	−0.48	**0.12**	0.20	0.16	
							Model median					
R^2	0.83	0.71	0.79	**0.89**	0.71	0.69	0.69	0.77	0.80	0.86		
$\Delta\mu$	−0.11	−0.43	**−0.08**	−0.20	−0.23	0.20	−0.74	−0.41	−0.09	−0.09		
$\Delta\sigma$	0.13	−0.04	0.28	0.59	0.12	0.40	−0.21	**−0.03**	−0.32	0.10		
							Model rank					
Rank	2	8	4	3	7	10	9	6	5	**1**		

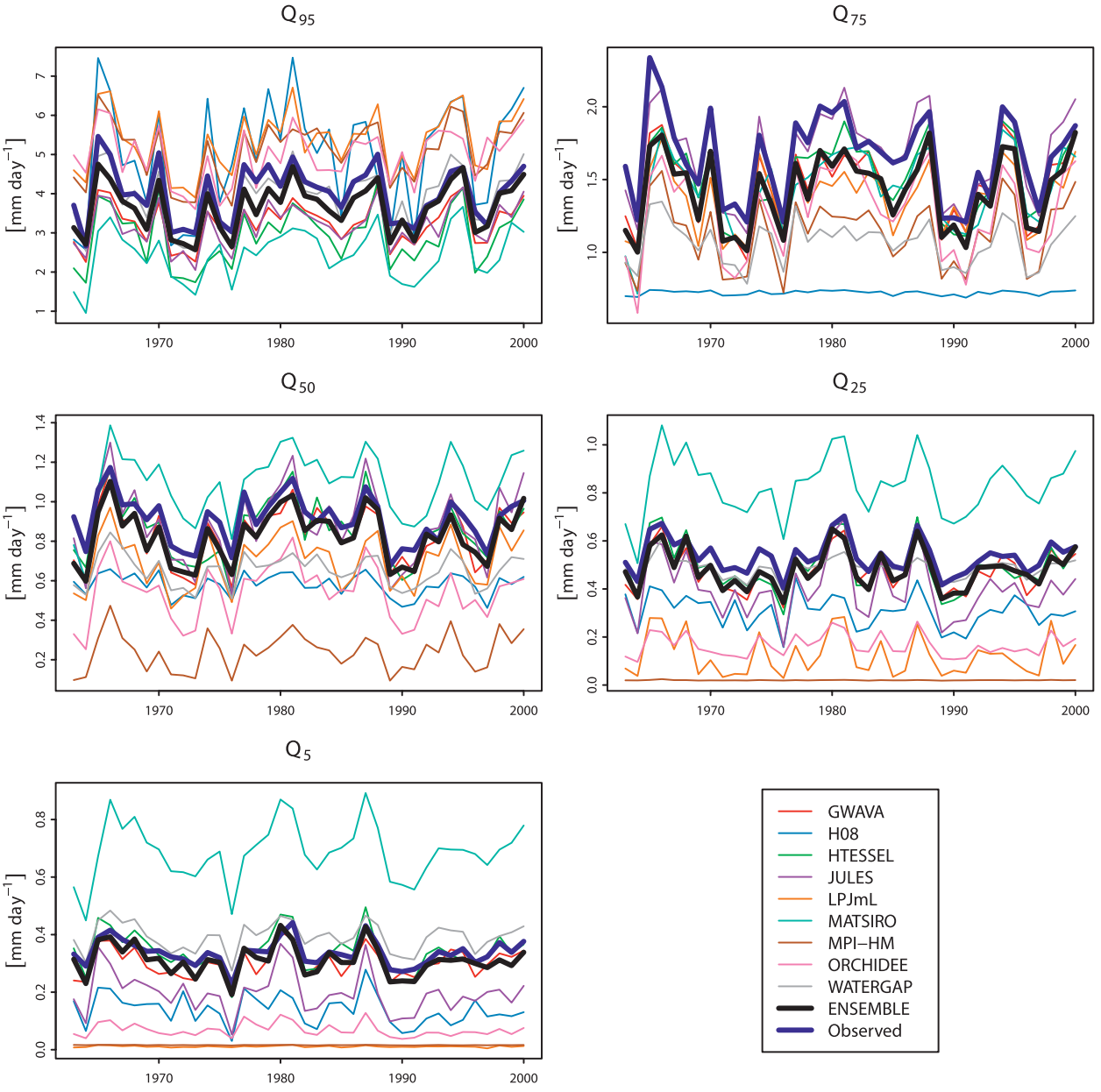


FIG. 4. Annual time series of observed and modeled runoff percentiles across Europe. Note the different scales of the *y* axes.

the annual percentile series. For the highest runoff percentile (Q_{95}), the models scatter evenly around the observed values. For all other runoff percentiles, most of the models underestimate the observations and there are, in some instances, also pronounced differences in the amplitude of the series. For example, H08 has a lower amplitude in the Q_{75} series than any other model, and some models [the hydrological model of the Max Planck Institute for Meteorology (MPI-HM) and Lund–Potsdam–Jena managed Land (LPJmL)] have almost constant values throughout the years for the two lowest runoff percentiles (Q_5 and Q_{25}). The LSM Minimal Advanced Treatments of Surface Interaction and Runoff (MATSIRO) is the only model that consistently overestimates the three lowest percentile levels.

Table 3 quantifies the differences between the observed and modeled runoff percentiles based on the three performance metrics R^2, $\Delta\mu$, and $\Delta\sigma$, and Fig. 6 summarizes the range of the performance metrics for each of the five runoff percentiles. The column ''percentile median'' in Table 3 provides the median of each performance metric for the different runoff percentiles

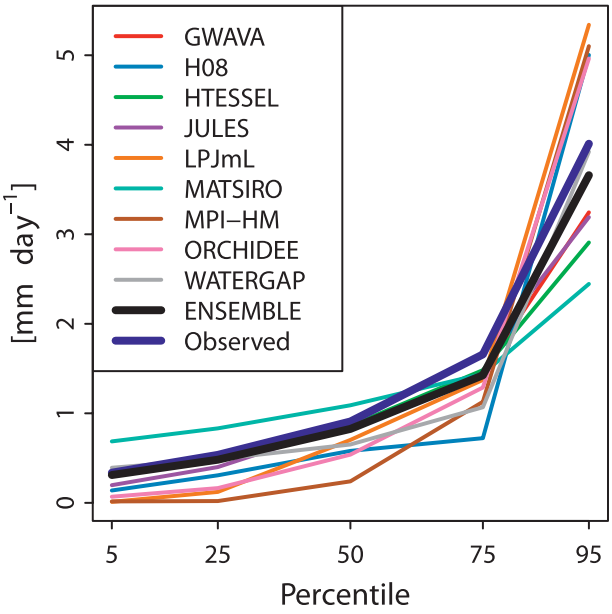


FIG. 5. Mean value of the runoff percentiles series (see Fig. 4).

and corresponds to the horizontal bars in Fig. 6. Numerical values reported in the following paragraph refer to these values if not specified differently. The correlation coefficients (R^2), quantifying the similarity of the temporal evolution of observed and modeled runoff percentiles, are on average highest for Q_{95} ($R^2_{95} = 0.82$; median value—subscripts indicate the runoff percentile) and lowest for Q_5 ($R^2_5 = 0.73$). The differences in correlation between the runoff percentiles are in most cases small, reflecting that the models capture the interannual dynamics of all flow levels relatively well. The relative difference in mean ($\Delta\mu$) is on average negative for all runoff percentiles, indicating that the models tend to underestimate runoff. The $\Delta\mu$ is smallest for Q_{95} ($\Delta\mu_{95} = -0.05$) and largest for Q_{25} ($\Delta\mu_{25} = -0.26$). The spread in $\Delta\mu$ is smallest for Q_{75} and largest for Q_5. In the latter case, differences between observed and simulated values range from $\Delta\mu = -0.97$ (LPJmL) to $\Delta\mu = 1.04$ (MATSIRO). The relative difference in standard deviation ($\Delta\sigma$) shows a rather complex picture. On average it is underestimated only for Q_{75} ($\Delta\sigma_{75} = -0.13$). It is closest to zero for Q_{95} ($\Delta\sigma_{95} = 0.01$), which means that the amplitude of the interannual variability of observed and simulated high flows are almost equal. On average, $\Delta\sigma$ is overestimated for the three lower runoff percentiles (Q_{50}, Q_{25}, and Q_5) and has its largest absolute value for the lowest flows ($\Delta\sigma_{25} = \Delta\sigma_5 = 0.16$). The relative difference in standard deviation also exhibits a large spread that increases toward the lower runoff percentiles. For Q_5 the spread is most pronounced and the relative error in standard deviation ranges from a strong underestimation $\Delta\sigma = -0.99$ (MPI-HM) to a strong overestimation $\Delta\sigma = 1.23$ (MATSIRO).

Figure 7 summarizes the performance of the individual models. The rows "model median" in Table 3 provide the median performance for each model averaged over all runoff percentiles and correspond to the bars in Fig. 7. The numbers reported in this paragraph refer to these median values if not stated differently. On average the Joint U.K. Land Environment Simulator (JULES) captures the interannual variability of the observed Q_{95} series best ($R^2_{\text{JULES}} = 0.89$; median values—subscripts indicate model name), closely followed by the ENSEMBLE ($R^2_{\text{ENSEMBLE}} = 0.86$) and GWAVA ($R^2_{\text{GWAVA}} = 0.83$). These models, as well as WaterGAP and Hydrology Tiled ECMWF Scheme for Surface Exchanges over Land (HTESSEL), also have a small spread in R^2 that is contrasted by the larger differences in R^2 found for the other models. On average, HTESSEL has the smallest bias ($\Delta\mu_{\text{HTESSEL}} = -0.08$), closely followed by WaterGAP ($\Delta\mu_{\text{WaterGAP}} = -0.09$), the ENSEMBLE ($\Delta\mu_{\text{ENSEMBLE}} = -0.09$), and GWAVA ($\Delta\mu_{\text{GWAVA}} = -0.11$). These models have almost equal biases for all runoff percentiles, which contrasts the large spread in $\Delta\mu$ found for LPJmL, MATSIRO, MPI-HM, and Organizing Carbon and Hydrology in Dynamic Ecosystems (ORCHIDEE). For most of these models, the large spread is associated with a overestimation of Q_{95} followed by pronounced underestimations of the lowest runoff percentiles. MATSIRO, the only model that consistently overestimates runoff, has an opposite pattern with underestimated high flows and overestimated low flows. On average ORCHIDEE captures the variance of annual runoff percentiles almost perfectly ($\Delta\mu_{\text{ORCHIDEE}} = -0.03$), followed by H08 ($\Delta\mu_{\text{H08}} = -0.03$) and the ENSEMBLE ($\Delta\mu_{\text{ENSEMBLE}} = 0.10$). However, H08 also has the largest spread in $\Delta\sigma$, with a large overestimation of the standard deviation of Q_{95} followed by a pronounced underestimation of the standard deviation of Q_{75}. All other models capture the standard deviation of the high flows reasonably well. However, the absolute values of $\Delta\sigma$ tend to increase for the low runoff percentiles, causing a large spreads in $\Delta\sigma$ for most models.

The last column in Table 3 ranks the ability of the models (including the ENSEMBLE) to reproduce the interannual dynamics of European runoff percentiles. The overall model performance decreases systematically from high (Q_{95}; rank 1) to low (Q_5; rank 5) percentiles, implying that the models capture annual high flows better than annual low flows. Note, however, that this ranking is not strictly monotonic (anomaly in the ordering of Q_{50} and Q_{75}). Interestingly, the tendency for poorer model performance for the low runoff percentiles is not only manifested in a drop in average model performance, but also by an increasing spread in $\Delta\mu$ and $\Delta\sigma$ (Fig. 6).

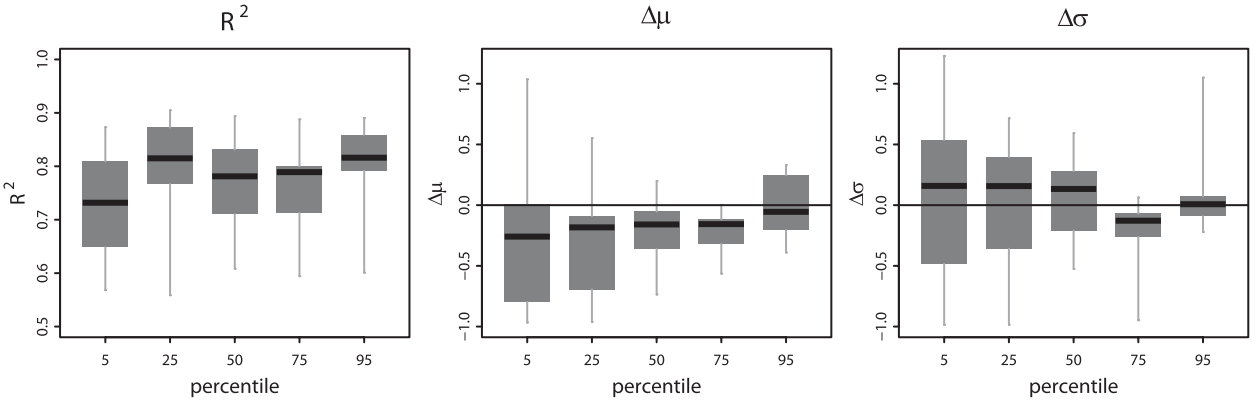


FIG. 6. Comparison of model performance for the different runoff percentiles. Performance is measured by (left to right) correlation (R^2), relative bias ($\Delta\mu$), and the relative difference in standard deviation ($\Delta\sigma$) for the five runoff percentiles. (Bar: median, box: interquartile range, and whiskers: range).

Table 3 also shows the ranking of the models themselves. The ENSEMBLE ranks number one, followed by GWAVA, JULES, and HTESSEL. A careful inspection of Table 3 confirms that the three highest-ranking models are closest to the observations with respect to the correlation coefficient (R^2) and the relative difference in mean ($\Delta\mu$). For the relative difference in standard deviation ($\Delta\sigma$), however, this is not strictly the case, and more midranking models exhibit a closer similarity to the observations. In general, no single performance metric could be identified that clearly explains why some models perform better than others. There is rather a tendency for a uniform decrease in all three criteria from the highest- to the lowest-ranked model.

5. Discussion

The comparison of the five aggregated time series of observed and simulated annual runoff percentiles not only provided insights into the ability of individual models to capture the magnitude and dynamics of annual runoff percentiles, but also allowed for an assessment of the overall performance of the multimodel ensemble. A good model performance with respect to interannual variability of all runoff percentiles (as reflected by relatively high R^2) is most likely related to the fact that the dynamics of annual runoff closely follow those of the atmospheric drivers. Shorthouse and Arnell (1997, 1999), for example, have demonstrated the coupling between atmospheric oscillation indices and river flow in Europe, and recently Gudmundsson et al. (2011b) showed that the dominant space–time patterns of European low-frequency runoff variability (variability on time scales longer than 1 yr) were closely related to the corresponding patterns of precipitation and temperature. This dependence of runoff on atmospheric variability suggests that simulated runoff on interannual time scales may be more sensitive to the data product used to force the models

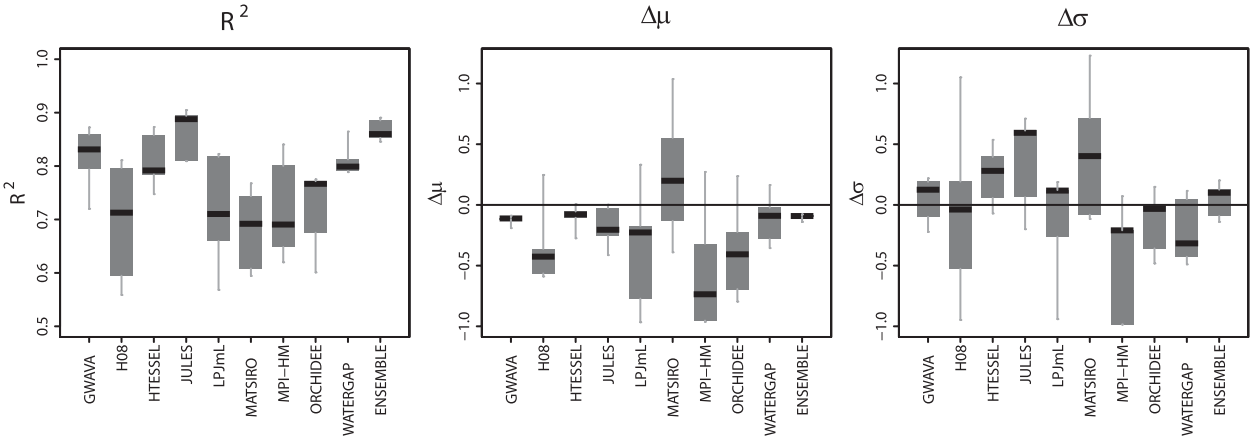


FIG. 7. As in Fig. 6, but for the participating models.

than to the parameterization of terrestrial hydrological processes. In fact, it has been previously demonstrated that simulated river discharge from continental-scale basins is highly sensitive to the choice of forcing data (e.g., Nasonova et al. 2011; Materia et al. 2010; Gerten et al. 2008; Hagemann and Jacob 2007).

The models' ability to capture the interannual variability was contrasted by a systematic underestimation of observed runoff in Europe. In a global analysis of discharge from continental-scale river basins (e.g., Amazon, Congo, and Lena) using a multimodel ensemble comparable to the ensemble used in this study, Haddeland et al. (2011) did not find similar consistent patterns of underestimation. They rather found large regional differences, with a tendency to underestimate observed discharge from river basins at high latitudes. In principal, a bias in the mean can either be attributed to biased atmospheric input variables (e.g., Nasonova et al. 2011; Teutschbein and Seibert 2010) or to a too-rapid depletion of stores through modeled evapotranspiration. The consistency of the underestimation in the present study, however, points toward biased forcing data, for example, because of the fact that local orographic effects on precipitation cannot be resolved within large grid cells of atmospheric reanalysis or interpolated data products. It is, for example, well documented that the ERA-40 data underlying the WFD underestimate precipitation in regions with complex topography (e.g., Adam et al. 2006; Barstad et al. 2009) and the bias correction procedure underlying the WFD does not account for orographic effects on precipitation (Weedon et al. 2010, 2011). Thus, this likely explains some of the biases in simulated runoff. Additional observations would be needed to investigate this further, which is beyond the scope of this study.

One of the most striking results of the model evaluation is the systematic decrease in model performance from wet to dry runoff percentiles (Table 3, Fig. 6). Both $\Delta\mu$ and $\Delta\sigma$ are relative measures and the impact of small absolute errors is larger for small observed values. Therefore, both $\Delta\mu$ and $\Delta\sigma$ can increase in magnitude for the lower runoff percentiles even if the absolute value of the error is constant. The existence of such effects is to some extent supported by Fig. 5, where the differences in observed and simulated mean values are almost constant throughout the runoff percentiles. This shows that there are only minor differences in the absolute model error between high and low flows. Despite such artifacts there are good reasons for normalizing the model error. The difference between low and high flows is larger than one order of magnitude. Therefore, model errors that are not normalized simply would follow this pattern, rendering interpretations difficult. Further, an error of a particular magnitude will be less relevant for large than for small values. This is especially the case if the error has the same magnitude as the observed quantity itself. In this context, it shall also be emphasized that the ranking of model performance has to be interpreted with caution and is only thought of as guidance for the careful inspection of the performance metrics themselves. Because of the nature of the procedure, small, possibly insignificant, differences may alter the ranking. Therefore, it is likely that neighboring ranks in fact represent broadly comparable performances. An alternative approach to make an average ranking (such as in this study) more reliable is to introduce weights for the different performance metrics such that metrics with a larger spread will have a larger influence on the overall ranking (e.g., Gulden et al. 2008; Gleckler et al. 2008). However, the choice of weights is nontrivial and results may depend on the method selected. Therefore, we opted to present only an unweighted ranking.

The large differences in model performance, especially for the lowest runoff percentiles, demonstrate the uncertainty associated with the appropriate mathematical representation of hydrological systems. Resolving this structural uncertainty is a subject of ongoing research (e.g., Gupta et al. 2008; Rosero et al. 2009; Martinez and Gupta 2010; Clark et al. 2011b,a) and would go beyond of the scope of the current study. Other sources of uncertainty are related to the estimation of model parameters. The models use a wide range of data products to determine soil properties and vegetation characteristics and different models may even have different interpretations of the same data source. For example, Teuling et al. (2009) demonstrated that soil properties derived from three different data products used in the European Land Data Assimilation System (ELDAS) project led to significant differences in the system behavior of a stochastic soil moisture model. The data products used to retrieve model parameters were not harmonized for the present ensemble and, even if some of the models rely on the same input maps, the processing and interpretation of the mapped values to derive the parameters may differ substantially. For example, H08 assumes a soil layer with an uniform soil with a depth of 1 m and a field capacity of 15 cm throughout all grid cells (Hanasaki et al. 2008), while the soil parameters of HTESSEL are taken from the Food and Agriculture Organization (FAO) dataset (FAO 2003), and ORCHIDEE determines the parameters of the Van Genuchten equations based on the suggestions of Carsel and Parrish (1988) for U.S. Department of Agriculture (USDA) soil types. A similar diversity of data products and approaches is also the case for other parameters such as vegetation characteristics.

It is regularly observed that hydrological models with mathematical structures that are comparable to the

models in the current ensemble often have deficiencies in simulating the lowest flows correctly (Smakhtin 2001; Stahl et al. 2011). To date, the reason for high flows being better (and more consistently) simulated than low flows is not fully understood. The fact that four of the five lowest-ranking models overestimate Q_{95}, followed by an increasingly pronounced underestimation in all other runoff percentiles (see Fig. 5), suggests that some models release too much of the incoming precipitation too quickly. Consequently, too little water is stored in soils and aquifers, which in turn may lead to pronounced underestimation of the lowest flows. The only model to exhibit an opposite behavior is MATSIRO, which reacts too slowly to precipitation as it underestimates the magnitude of high flows and overestimates the low flows.

Most models capture the standard deviation of Q_{95} relatively well, but large discrepancies are found in the standard deviations of the annual low flows. This may be a result of high flows (and floods) being more directly coupled to atmospheric variability than low flows. Thus, the variance of high flows, as well as the temporal evolution, is likely to be directly related to precipitation variability, whereas low flows are to a much larger extent influenced by terrestrial hydrological processes. Various empirical studies support this. For example, Gudmundsson et al. (2011a) demonstrated, using the same observed dataset that is the basis for this study, that annual high flows have a high degree of synchronization across Europe, reflecting their link to atmospheric variability. Low flows, on the other hand, were found to have a more complex spatial pattern and a lower degree of synchronization, suggesting an increasing influence of catchment processes under dry conditions. Similarly, Bouwer et al. (2008) found that annual maximum river discharges in Europe were more sensitive to variations in the atmospheric forcing than annual mean discharges. It is also noteworthy that statistical moments of mean annual floods have been reported to be significantly correlated to the hydroclimatic conditions, but not to static catchment properties such as geology and soil types (Merz and Blöschl 2009). In summary, these results suggest that continental-scale patterns of runoff response are closely linked to the atmospheric forcing under wet conditions, irrespective of the properties of the catchments. Under dry conditions on the other hand, runoff depends primarily on depleting storages, the extent and properties of which vary strongly with topography and hydrogeology (Smakhtin 2001; Whitehouse et al. 1983) as well as on the antecedent moisture conditions.

The large differences in performance between models are contrasted by the good performance of the ensemble mean (ENSEMBLE). The present study showed that the ENSEMBLE is actually closer to the observed series of annual high flows (Q_{95}) and low flows (Q_5) than any other model with respect to R^2, and has a performance comparable to the best models with respect to $\Delta\mu$ and $\Delta\sigma$ (Table 3). The ENSEMBLE is also superior for the simulation of low and high flows, which can likely be related to the fact that the percentile series provide robust estimates of annual high and low flows, but do not take the actual timing of flow events into account. Accordingly, ensemble techniques appear to increase the reliability of simulations of the terrestrial water cycle with respect to extremes on large spatial and temporal scales. The reason for the superiority of the ENSEMBLE compared to any individual model is not clear, but a possible explanation is that the model solutions scatter more or less evenly around the true value (unless the errors are systematic), and thus, the errors behave like random noise that can be efficiently removed by averaging. Note, however, that in the present study this is only the case for the highest flows (Fig. 4). For climate simulations, such noise arises from the simulated internal climate variability and from uncertainties in the model parameterizations (Reichler and Kim 2008). Similar arguments also hold for hydrological systems where the uncertainty on the ''true'' physical representation may lead to an even scatter of model errors around the observations, and thus increases the reliability of the predictions.

6. Summary and conclusions

This study assessed the ability of an ensemble of nine large-scale, hydrological models to capture the magnitude and the interannual variability of runoff percentiles representing dry, mean, and wet conditions in Europe. In contrast to other studies that evaluate the performance of large-scale hydrological models using only a few continental-scale river basins, this study uses observation-based runoff estimates in 298 grid cells. The gridded runoff was derived from gauged river flow series from 426 small, near-natural catchments, reducing the risk of biased conclusions due to observation error. To minimize the effect of local parameter uncertainty and to focus on the dominant patterns of interannual variability, spatially aggregated time series were analyzed.

Overall, the ensemble members were able to capture the temporal evolution of the interannual variability, measured by the correlation coefficient R^2, reasonably well. However, an overall tendency toward underestimation of runoff was found, and both structural issues common to all models and biases in the forcing data are plausible explanations.

Model performance decreases from wet to dry conditions. This change in average model performance is

accompanied by an increasing spread in the relative error in the mean ($\Delta\mu$) as well as in the standard deviation ($\Delta\sigma$) for the low runoff percentiles. One possible explanation is that hydrological systems are more closely coupled to the meteorological forcing under wet conditions, whereas runoff under dry conditions depends more on storage processes whose parameterization are highly uncertain.

The large differences in performance among the models are contrasted by the fact that the ENSEMBLE, the mean over all models, provides the most reliable estimation of spatially aggregated time series of all annual runoff percentiles. The ensemble mean not only provides a good overall estimator, but is also closer to the series of annual high flows (Q_{95}) and low flows (Q_5) than most models. This leads us to caution against the use of a single model in climate impact assessment, which is associated with a high risk of biased conclusions, and rather recommend the use of multimodel ensembles.

A principle limitation of this study is the loss of information due to the spatial aggregation in data preprocessing. Possible approaches to gain insights to the spatial patterns of model performance could include the analysis of smaller regions or more ''intelligent'' data preprocessing to define and extract signals (e.g., the mean annual cycle and leading empirical orthogonal functions) that are expected to be reproduced by the models. These issues are subject to ongoing research and addressed in a parallel study (Gudmundsson et al. 2011c, manuscript submitted to *Water. Resour. Res.*).

Acknowledgments. This research contributes to the European Union (FP6) funded Integrated Project WATCH (Contract 036946). The provision of streamflow data by all agencies that contributed data to the EWA-FRIEND or to the WATCH project is gratefully acknowledged. We further acknowledge the contribution of Pedro Viterbo and Sandra Gomes from the University of Lisbon and Jan Polcher Laboratoire de Meteorologie Dynamique (Paris) for providing model results and helpful comments.

REFERENCES

Adam, J. C., and D. P. Lettenmaier, 2003: Adjustment of global gridded precipitation for systematic bias. *J. Geophys. Res.,* **108,** 4257, doi:10.1029/2002JD002499.

——, E. A. Clark, D. P. Lettenmaier, and E. F. Wood, 2006: Correction of global precipitation products for orographic effects. *J. Climate,* **19,** 15–38.

Alcamo, J., P. Petra Döll, T. Henrichs, F. Kaspar, B. Lehner, T. Roumlsch, and S. Siebert, 2003: Development and testing of the WaterGAP 2 global model of water use and availability. *Hydrol. Sci. J.,* **48,** 317–337, doi:10.1623/hysj.48.3.317.45290.

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

Barstad, I., A. Sorteberg, F. Flatø, and M. Déqué, 2009: Precipitation, temperature and wind in Norway: Dynamical downscaling of ERA40. *Climate Dyn.,* **33,** 769–776, doi:10.1007/s00382-008-0476-5.

Best, M. J., and Coauthors, 2011: The Joint UK Land Environment Simulator (JULES), model description—Part 1: Energy and water fluxes. *Geosci. Model Dev.,* **4,** 677–699, doi:10.5194/gmd-4-677-2011.

Bondeau, A., and Coauthors, 2007: Modelling the role of agriculture for the 20th century global terrestrial carbon balance. *Global Change Biol.,* **13,** 679–706, doi:10.1111/j.1365-2486.2006.01305.x.

Bouwer, L. M., J. E. Vermaat, and J. C. J. H. Aerts, 2008: Regional sensitivities of mean and peak river discharge to climate variability in Europe. *J. Geophys. Res.,* **113,** D19103, doi:10.1029/2008JD010301.

Carsel, R. F., and R. S. Parrish, 1988: Developing joint probability distributions of soil water retention characteristics. *Water Resour. Res.,* **24,** 755–769.

Clark, D. B., and Coauthors, 2011: The Joint UK Land Environment Simulator (JULES), model description—Part 2: Carbon fluxes and vegetation dynamics. *Geosci. Model Dev.,* **4,** 701–722, doi:10.5194/gmd-4-701-2011.

Clark, M. P., A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener, and L. E. Hay, 2008: Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. *Water Resour. Res.,* **44,** W00B02, doi:10.1029/2007WR006735.

——, D. Kavetski, and F. Fenicia, 2011a: Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resour. Res.,* **47,** W09301, doi:10.1029/2010WR009827.

——, H. K. McMillan, D. B. G. Collins, D. Kavetski, and R. A. Woods, 2011b: Hydrological field data from a modeller's perspective: Part 2: Process-based evaluation of model hypotheses. *Hydrol. Processes,* **25,** 523–543, doi:10.1002/hyp.7902.

Dankers, R., and L. Feyen, 2009: Flood hazard in Europe in an ensemble of regional climate scenarios. *J. Geophys. Res.,* **114,** D16108, doi:10.1029/2008JD011523.

Decharme, B., and H. Douville, 2007: Global validation of the ISBA sub-grid hydrology. *Climate Dyn.,* **29,** 21–37, doi:10.1007/s00382-006-0216-7.

Di Baldassarre, G., and A. Montanari, 2009: Uncertainty in river discharge observations: A quantitative analysis. *Hydrol. Earth Syst. Sci.,* **13,** 913–921, doi:10.5194/hess-13-913-2009.

Dirmeyer, P. A., 2011: A history and review of the Global Soil Wetness Project (GSWP). *J. Hydrometeor.,* **12,** 729–749.

——, X. Gao, M. Zhao, Z. Guo, T. Oki, and N. Hanasaki, 2006: GSWP-2: Multimodel analysis and implications for our perception of the land surface. *Bull. Amer. Meteor. Soc.,* **87,** 1381–1397.

Döll, P., F. Kaspar, and B. Lehner, 2003: A global hydrological model for deriving water availability indicators: Model tuning and validation. *J. Hydrol.,* **270** (1–2), 105–134, doi:10.1016/S0022-1694(02)00283-4.

——, K. Fiedler, and J. Zhang, 2009: Global-scale analysis of river flow alterations due to water withdrawals and reservoirs. *Hydrol. Earth Syst. Sci.,* **13,** 2413–2432.

d'Orgeval, T., J. Polcher, and P. de Rosnay, 2008: Sensitivity of the West African hydrological cycle in ORCHIDEE to infiltration

processes. *Hydrol. Earth Syst. Sci.,* **12,** 1387–1401, doi:10.5194/hess-12-1387-2008.

Duan, Q., and Coauthors, 2006: Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. *J. Hydrol.,* **320** (1–2), 3–17, doi:10.1016/j.jhydrol.2005.07.031.

Dümenil, L., and E. Todini, 1992: A rainfall-runoff scheme for use in the Hamburg climate model. *Advances in Theoretical Hydrology: A Tribute to James Dooge,* J. P. O'Kane, Ed., Elsevier Science, 129–157.

Fader, M., S. Rost, C. Müller, A. Bondeau, and D. Gerten, 2010: Virtual water content of temperate cereals and maize: Present and potential future patterns. *J. Hydrol.,* **384** (3–4), 218–231, doi:10.1016/j.jhydrol.2009.12.011.

FAO, 2003: Digital soil map of the world and derived soil properties. Food and Agriculture Organization of the United Nations, CD-ROM.

Feyen, L., and R. Dankers, 2009: Impact of global warming on streamflow drought in Europe. *J. Geophys. Res.,* **114,** D17116, doi:10.1029/2008JD011438.

Fuchs, T., 2009: GPCC Annual report for year 2008. Global Precipitation Climatology Centre Tech. Rep., DWD, 13 pp. [Available online at http://gpcc.dwd.de.]

Gao, X., and P. A. Dirmeyer, 2006: A multimodel analysis, validation, and transferability study of global soil wetness products. *J. Hydrometeor.,* **7,** 1218–1236.

Gerten, D., S. Schaphoff, U. Haberlandt, W. Lucht, and S. Sitch, 2004: Terrestrial vegetation and water balance—Hydrological evaluation of a dynamic global vegetation model. *J. Hydrol.,* **286** (1–4), 249–270, doi:10.1016/j.jhydrol.2003.09.029.

——, S. Rost, W. von Bloh, and W. Lucht, 2008: Causes of change in 20th century global river discharge. *Geophys. Res. Lett.,* **35,** L20405, doi:10.1029/2008GL035258.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.,* **113,** D06104, doi:10.1029/2007JD008972.

Gudmundsson, L., L. M. Tallaksen, and K. Stahl, 2011a: Spatial cross-correlation patterns of European low, mean and high flows. *Hydrol. Processes,* **25,** 1034–1045, doi:10.1002/hyp.7807.

——, ——, ——, and A. K. Fleig, 2011b: Low-frequency variability of European runoff. *Hydrol. Earth Syst. Sci.,* **15,** 2853–2869, doi:10.5194/hess-15-2853-2011.

Gulden, L. E., E. Rosero, Z.-L. Yang, T. Wagener, and G.-Y. Niu, 2008: Model performance, model robustness, and model fitness scores: A new method for identifying good land-surface models. *Geophys. Res. Lett.,* **35,** L11404, doi:10.1029/2008GL033721.

Guo, Z., and P. A. Dirmeyer, 2006: Evaluation of the Second Global Soil Wetness Project soil moisture simulations: 1. Intermodel comparison. *J. Geophys. Res.,* **111,** D22S02, doi:10.1029/2006JD007233.

——, ——, X. Gao, and M. Zhao, 2007: Improving the quality of simulated soil moisture with a multi-model ensemble approach. *Quart. J. Roy. Meteor. Soc.,* **133,** 731–747, doi:10.1002/qj.48.

Gupta, H. V., T. Wagener, and Y. Liu, 2008: Reconciling theory with observations: Elements of a diagnostic approach to model evaluation. *Hydrol. Processes,* **22,** 3802–3813, doi:10.1002/hyp.6989.

Haddeland, I., and Coauthors, 2011: Multimodel estimate of the global terrestrial water balance: Setup and first results. *J. Hydrometeor.,* **12,** 869–884.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. *Tellus,* **57A,** 219–233, doi:10.1111/j.1600-0870.2005.00103.x.

Hagemann, S., and L. Dümenil, 1998: A parametrization of the lateral waterflow for the global scale. *Climate Dyn.,* **14,** 17–31, doi:10.1007/s003820050205.

——, and L. Dümenil Gates, 2003: Improving a subgrid runoff parameterization scheme for climate models by the use of high resolution data derived from satellite observations. *Climate Dyn.,* **21,** 349–359, doi:10.1007/s00382-003-0349-x.

——, and D. Jacob, 2007: Gradient in the climate change signal of European discharge predicted by a multi-model ensemble. *Climatic Change,* **81** (Suppl.), 309–327, doi:10.1007/s10584-006-9225-0.

——, H. Göttel, D. Jacob, P. Lorenz, and E. Roeckner, 2009: Improved regional scale processes reflected in projected hydrological changes over large European catchments. *Climate Dyn.,* **32,** 767–781, doi:10.1007/s00382-008-0403-9.

Hanasaki, N., S. Kanae, T. Oki, K. Masuda, K. Motoya, N. Shirakawa, Y. Shen, and K. Tanaka, 2008: An integrated model for the assessment of global water resources—Part 1: Model description and input meteorological forcing. *Hydrol. Earth Syst. Sci.,* **12,** 1007–1025.

Hansen, J., M. Sato, R. Ruedy, K. Lo, D. W. Lea, and M. Medina-Elizade, 2006: Global temperature change. *Proc. Natl. Acad. Sci. USA,* **103,** 14 288–14 293, doi:10.1073/pnas.0606291103.

Henderson-Sellers, A., A. J. Pitman, P. K. Love, P. Irannejad, and T. H. Chen, 1995: The Project for Intercomparison of Land Surface Parameterization Schemes (PILPS): Phases 2 and 3. *Bull. Amer. Meteor. Soc.,* **76,** 489–503.

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

Hunger, M., and P. Döll, 2008: Value of river discharge data for global-scale hydrological modeling. *Hydrol. Earth Syst. Sci.,* **12,** 841–861, doi:10.5194/hess-12-841-2008.

Kavetski, D., G. Kuczera, and S. W. Franks, 2006: Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resour. Res.,* **42,** W03407, doi:10.1029/2005WR004368.

Laaha, G., and G. Blöschl, 2007: A national low flow estimation procedure for Austria. *Hydrol. Sci. J.,* **52,** 625–644, doi:10.1623/hysj.52.4.625.

Lehner, B., P. Döll, J. Alcamo, T. Henrichs, and F. Kaspar, 2006: Estimating the impact of global change on flood and drought risks in Europe: A continental, integrated analysis. *Climatic Change,* **75,** 273–299, doi:10.1007/s10584-006-6338-4.

Macadam, I., A. J. Pitman, P. H. Whetton, and G. Abramowitz, 2010: Ranking climate models by performance using actual values and anomalies: Implications for climate change impact assessments. *Geophys. Res. Lett.,* **37,** L16704, doi:10.1029/2010GL043877.

Manabe, S., 1969: Climate and the ocean circulation. I. The atmospheric circulation and the hydrology of the earth's surface. *Mon. Wea. Rev.,* **97,** 739–774.

Martinez, G. F., and H. V. Gupta, 2010: Toward improved identification of hydrological models: A diagnostic evaluation of the ''abcd'' monthly water balance model for the conterminous United States. *Water Resour. Res.,* **46,** W08507, doi:10.1029/2009WR008294.

Materia, S., P. A. Dirmeyer, Z. Guo, A. Alessandri, and A. Navarra, 2010: The sensitivity of simulated river discharge to land surface representation and meteorological forcings. *J. Hydrometeor.,* **11,** 334–351.

McMillan, H., J. Freer, F. Pappenberger, T. Krueger, and M. Clark, 2010: Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions. *Hydrol. Processes,* **24,** 1270–1284, doi:10.1002/hyp.7587.

Meigh, J. R., A. A. McKenzie, and K. J. Sene, 1999: A grid-based approach to water scarcity estimates for eastern and southern Africa. *Water Resour. Manage.,* **13,** 85–115, doi:10.1023/A:1008025703712.

Merz, R., and G. Blöschl, 2009: Process controls on the statistical flood moments—A data based analysis. *Hydrol. Processes,* **23,** 675–696, doi:10.1002/hyp.7168.

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

Mitchell, T. D., and P. D. Jones, 2005: An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *Int. J. Climatol.,* **25,** 693–712, doi:10.1002/joc.1181.

Moore, R. J., 1985: The probability-distributed principle and runoff production at point and basin scales. *Hydrol. Sci. J.,* **30,** 273–297.

——, 2007: The PDM rainfall-runoff model. *Hydrol. Earth Syst. Sci.,* **11,** 483–499, doi:10.5194/hess-11-483-2007.

Nasonova, O. N., Ye. M. Gusev, and Ye. E. Kovalev, 2011: Impact of uncertainties in meteorological forcing data and land surface parameters on global estimates of terrestrial water balance components. *Hydrol. Processes,* **25,** 1074–1090, doi:10.1002/hyp.7651.

New, M., M. Hulme, and P. Jones, 1999: Representing twentieth-century space–time climate variability. Part I: Development of a 1961–90 mean monthly terrestrial climatology. *J. Climate,* **12,** 829–856.

——, ——, and ——, 2000: Representing twentieth-century space–time climate variability. Part II: Development of 1901–96 monthly grids of terrestrial surface climate. *J. Climate,* **13,** 2217–2238.

Nohara, D., A. Kitoh, M. Hosaka, and T. Oki, 2006: Impact of climate change on river discharge projected by multimodel ensemble. *J. Hydrometeor.,* **7,** 1076–1089.

Oki, T., T. Nishimura, and P. Dirmeyer, 1999: Assessment of annual runoff from land surface models using Total Runoff Integrating Pathways (TRIP). *J. Meteor. Soc. Japan,* **77,** 235–255.

Palmer, T. N., and Coauthors, 2004: Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER). *Bull. Amer. Meteor. Soc.,* **85,** 853–872.

Reichler, T., and J. Kim, 2008: How well do coupled models simulate today's climate? *Bull. Amer. Meteor. Soc.,* **89,** 303–311.

Reitan, T., and A. Petersen-Øverleir, 2009: Bayesian methods for estimating multi-segment discharge rating curves. *Stochastic Environ. Res. Risk Assess.,* **23,** 627–642, doi:10.1007/s00477-008-0248-0.

Renard, B., D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks, 2010: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resour. Res.,* **46,** W05521, doi:10.1029/2009WR008328.

Roeckner, E., and Coauthors, 2003: The atmospheric general circulation model ECHAM 5. Part I: Model description. Max Planck Institute for Meteorology Tech. Rep. 349, 127 pp.

Rosero, E., Z.-L. Yang, L. E. Gulden, G.-Y. Niu, and D. J. Gochis, 2009: Evaluating enhanced hydrological representations in Noah LSM over transition zones: Implications for model development. *J. Hydrometeor.,* **10,** 600–622.

Rudolf, B., and U. Schneider, 2005: Calculation of gridded precipitation data for the global land-surface using in-situ gauge observations. *Proc. Second Workshop of the International Precipitation Working Group,* Monterey, CA, IPWG, 231–247. [Available online at http://gpcc.dwd.de.]

Schneider, U., A. Becker, A. Meyer-Christoffer, M. Ziese, and B. Rudolf, 2010: Global precipitation analysis products of the GPCC. Global Precipitation Climatology Centre Tech. Rep., DWD, 12 pp. [Available online at ftp://ftp.dwd.de/pub/data/gpcc/PDF/GPCC_intro_products_2008.pdf.]

Shorthouse, C., and N. Arnell, 1997: Spatial and temporal variability in European river flows and the North Atlantic oscillation. *FRIEND'97—Regional Hydrology: Concepts and Models for Sustainable Water Resource Management,* A. Gustard et al., Eds., IAHS, 77–85.

——, and ——, 1999: The effects of climatic variability on spatial characteristics of European river flows. *Phys. Chem. Earth,* **24B** (1–2), 7–13, doi:10.1016/S1464-1909(98)00003-3.

Smakhtin, V. U., 2001: Low flow hydrology: A review. *J. Hydrol.,* **240** (3–4), 147–186, doi:10.1016/S0022-1694(00)00340-1.

Stahl, K., H. Hisdal, L. Tallaksen, H. van Lanen, J. Hannaford, and E. Sauquet, 2008: Trends in low flows and streamflow droughts across Europe. UNESCO Tech. Rep., 39 pp.

——, and Coauthors, 2010: Streamflow trends in Europe: Evidence from a dataset of near-natural catchments. *Hydrol. Earth Syst. Sci. Discuss.,* **7,** 5769–5804, doi:10.5194/hessd-7-5769-2010.

——, L. M. Tallaksen, L. Gudmundsson, and J. H. Christensen, 2011: Streamflow data from small basins: A challenging test to high-resolution regional climate modeling. *J. Hydrometeor.,* **12,** 900–912.

Takata, K., S. Emori, and T. Watanabe, 2003: Development of the minimal advanced treatments of surface interaction and runoff. *Global Planet. Change,* **38** (1–2), 209–222, doi:10.1016/S0921-8181(03)00030-4.

Teuling, A. J., R. Uijlenhoet, B. van den Hurk, and S. I. Seneviratne, 2009: Parameter sensitivity in LSMs: An analysis using stochastic soil moisture models and ELDAS soil parameters. *J. Hydrometeor.,* **10,** 751–765.

Teutschbein, C., and J. Seibert, 2010: Regional climate models for hydrological impact studies at the catchment scale: A review of recent modeling strategies. *Geography Compass,* **4,** 834–860, doi:10.1111/j.1749-8198.2010.00357.x.

Todini, E., 1996: The ARNO rainfall–runoff model. *J. Hydrol.,* **175** (1–4), 339–382, doi:10.1016/S0022-1694(96)80016-3.

Troy, T. J., E. F. Wood, and J. Sheffield, 2008: An efficient calibration method for continental-scale land surface modeling. *Water Resour. Res.,* **44,** W09411, doi:10.1029/2007WR006513.

Uppala, S. M., and Coauthors, 2005: The ERA-40 Re-Analysis. *Quart. J. Roy. Meteor. Soc.,* **131,** 2961–3012, doi:10.1256/qj.04.176.

Vogt, J., and Coauthors, 2007: A pan-European river and catchment database. JRC Reference Rep. EUR 22920 EN, 119 pp.

Weedon, G. P., S. Gomes, P. Viterbo, H. Österle, J. C. Adam, N. Bellouin, O. Boucher, and M. Best, 2010: The WATCH

forcing data 1958–2001: A meteorological forcing dataset for land surface- and hydrological-models. WATCH Tech. Rep. 22, 41 pp. [Available online at http://www.eu-watch.org.]

——, and Coauthors, 2011: Creation of the WATCH Forcing Data and its use to assess global and regional reference crop evaporation over land during the twentieth century. *J. Hydrometeor.,* **12,** 823–848.

Whitehouse, I., M. McSaveney, and G. Horrell, 1983: Spatial variability of low flows across a portion of the central Southern Alps, New Zealand. *J. Hydrol.,* **20,** 123–137.

Widén-Nilsson, E., L. Gong, S. Halldin, and C.-Y. Xu, 2009: Model performance and parameter behavior for varying time aggregations and evaluation criteria in the WASMOD-M global water balance model. *Water Resour. Res.,* **45,** W05418, doi:10.1029/2007WR006695.