How Well Do Large-Scale Models Reproduce Regional Hydrological Extremes in Europe?

CHRISTEL PRUDHOMME, SIMON PARRY, JAMIE HANNAFORD, AND DOUGLAS B. CLARK
Centre for Ecology and Hydrology, Wallingford, United Kingdom

STEFAN HAGEMANN
Max Planck Institute for Meteorology, Hamburg, Germany

FRANK VOSs
Center for Environmental Systems Research, Universität-Kassel, Kassel, Germany

(Manuscript received 5 October 2010, in final form 31 March 2011)

ABSTRACT

This paper presents a new methodology for assessing the ability of gridded hydrological models to reproduce large-scale hydrological high and low flow events (as a proxy for hydrological extremes) as described by catalogues of historical droughts [using the regional deficiency index (RDI)] and high flows [regional flood index (RFI)] previously derived from river flow measurements across Europe. Using the same methods, total runoff simulated by three global hydrological models from the Water Model Intercomparison Project (WaterMIP) [Joint U.K. Land Environment Simulator (JULES), Water Global Assessment and Prognosis (WaterGAP), and Max Planck Institute Hydrological Model (MPI-HM)] run with the same meteorological input (watch forcing data) at the same spatial 0.5° grid was used to calculate simulated RDI and RFI for the period 1963–2001 in the same European regions, directly comparable with the observed catalogues. Observed and simulated RDI and RFI time series were compared using three performance measures: the relative mean error, the ratio between the standard deviation of simulated over observed series, and the Spearman correlation coefficient. Results show that all models can broadly reproduce the spatiotemporal evolution of hydrological extremes in Europe to varying degrees. JULES tends to produce prolonged, highly spatially coherent events for both high and low flows, with events developing more slowly and reaching and sustaining greater spatial coherence than observed—this could be due to runoff being dominated by slow-responding subsurface flow. In contrast, MPI-HM shows very high variability in the simulated RDI and RFI time series and a more rapid onset of extreme events than observed, in particular for regions with significant water storage capacity—this could be due to possible underrepresentation of infiltration and groundwater storage, with soil saturation reached too quickly. WaterGAP shares some of the issues of variability with MPI-HM—also attributed to insufficient soil storage capacity and surplus effective precipitation being generated as surface runoff—and some strong spatial coherence of simulated events with JULES, but neither of these are dominant. Of the three global models considered here, WaterGAP is arguably best suited to reproduce most regional characteristics of large-scale high and low flow events in Europe. Some systematic weaknesses emerge in all models, in particular for high flows, which could be a product of poor spatial resolution of the input climate data (e.g., where extreme precipitation is driven by local convective storms) or topography. Overall, this study has demonstrated that RDI and RFI are powerful tools that can be used to assess how well large-scale hydrological models reproduce large-scale hydrological extremes—an exercise rarely undertaken in model intercomparisons.

1. Introduction and background

There is growing evidence that the hydrological cycle is intensifying (e.g., Huntington 2006; Stott et al. 2010) as a result of anthropogenically forced climatic change.
Generally speaking, at regional to continental scales, two contrasting approaches are used to examine the influence of climate change on the hydrological cycle:

(i) through analysis of historical data, to detect emerging trends (e.g., in Europe by Stahl et al. 2010, in North America by Douglas et al. 2000 and Khaliq et al. 2008, and globally by Dai et al. 2009), and

(ii) through modeling approaches [e.g., European analyses of future flood risk (Dankers and Feyen 2008, 2009) and global analysis of droughts by Burke et al. 2006].

The latter attempts to reproduce observed variability in hydrological characteristics using process-driven models, which can then be used in combination with climate prediction to provide assessments of potential hydrological changes. Wilby et al. (2008) observe that the two approaches often yield conflicting results, but that this "conceptual controversy" is in fact partly due to the different philosophies and specific methodologies underlying them. Achieving greater reconciliation between models and observations, in order to better understand the influence of climate change on the hydrological cycle, is therefore one of the most pressing challenges of contemporary hydrology.

While there is inevitably much attention on aspects of the hydrological cycle that are directly influenced by anthropogenic forcing (e.g., increased evaporation through higher temperatures and increased rainfall as a result of higher moisture holding capacity in a warmer atmosphere), one of the most important potential impacts of climate change is on hydrological extremes (i.e., drought and flooding). Extremes are likely to be sensitive to climate change, raising the possibility that changes in the extremes of hydrological parameters may be more detectable than changes in mean conditions (Wilby et al. 2008; Sheffield and Wood 2008b). Furthermore, changes to extremes are of significant importance to society; in reality, researchers, policy makers, and planners are more interested in hydrological extremes than mean conditions. Droughts and floods have had significant impacts in Europe in the recent past: the cost of the 2003 drought in Europe to the agricultural and forestry sectors was estimated at 13.1 billion Euros (EUR), and the drought and heat wave caused an estimated 15 000 excess deaths (Fink et al. 2004); the 2007 floods that affected United Kingdom caused 15 fatalities (Marsh and Hannaford 2008) and cost 3.2 billion British pounds (GBP) (Chatterton et al. 2010) while the 2002 floods in central Europe resulted in over 100 fatalities and cost over 14 billion Euros (Ulbrich et al. 2003). Clearly, any future increase in the severity and frequency of hydrological extremes could have potentially far-reaching socioeconomic and environmental consequences.

Climate change influences on the hydrological cycle are often studied using large-scale gridded models, which enable elements of the hydrological cycle (e.g., rainfall, evapotranspiration, and runoff) and potential perturbations to the cycle as a result of climate change to be studied on regional (e.g., across Europe) to global scales. Such studies have predominantly analyzed climatic variables such as temperature and precipitation (e.g., Burke et al. 2006) or land surface variables such as soil moisture (e.g., Sheffield and Wood 2007, 2008a; Sheffield et al. 2009). However, while some of the most pressing direct impacts of droughts and floods are primarily related to runoff and river flows, comparatively little effort has focused on examining runoff extremes in large-scale models. A number of studies have examined runoff from global models at continental scales, but these studies have tended to focus on mean runoff (e.g., Milly et al. 2005; Gedney et al. 2006; Hamlet et al. 2007). Two notable exceptions are research by Milly et al. (2002), who compared the frequency of large floods, globally, in the twentieth century to future scenarios, and Lehner et al. (2006), who examined future flood and drought occurrence for Europe using a global hydrological model. At a finer spatial scale, Feyen and Dankers (2009) and Dankers and Feyen (2008) investigated future changes in high flows and drought in Europe using a hydrological model driven by a regional climate model, followed by a multimodel ensemble (Dankers and Feyen 2009). Generally, however, high or low flow extremes have been underexamined.

Results from all modeling studies are sensitive to the models used; model outputs must be corroborated against observed data if they are to be used with confidence. Most of the above studies perform some validation of model outputs against observed data, often based on comparing modeled flow at a limited number of catchment outlets against equivalent observed data from relatively large basins. Haddeland et al. (2011) detail a model intercomparison study comparing the performance of 11 models that simulate hydrological processes on a global scale. Their study, in common with previous model intercomparison studies (see references therein), focused on comparing mean and median runoff. A focus on the means also underpins the study undertaken for North America by Lohmann et al. (2004). In general, however, to the authors’ knowledge, there have been very few attempts to systematically validate runoff extremes from large-scale models against observed data in Europe or elsewhere. An appraisal of the performance of large-scale models in replicating historical hydrological extremes is, therefore, a necessary precursor to
assessing the suitability of such models for projecting characteristics of hydrological extremes into the twenty-first century.

This paper complements other model intercomparison studies by examining the performance of three large-scale hydrological models in reproducing observed streamflow in Europe, with a focus on the extreme (high and low) flows. Drought and flood characteristics vary spatially and temporally throughout Europe, and it is important to consider this variability when characterizing observed hydrological extremes at the European scale. This study therefore uses published drought and flood catalogues based on the regional deficiency index (RDI) and the regional flood index (RFI)—time series indicating the regional extent of hydrological extremes. Hannaford et al. (2011) applied the RDI methodology to produce “drought catalogues” for 23 European regions. For high flows, Parry et al. (2010) applied the same conceptual approach to propose the RFI methodology as a proxy for large-scale flood events. In addition to the existence of published historical European catalogues, the RDI and RFI methodologies have several distinct advantages for evaluating global model performance in terms of hydrological extremes: (i) the RDI and RFI historical catalogues are based on small, undisturbed catchments—as argued by Stahl et al. (2010), who used the same dataset for a pan-European trend assessment, this dataset provides a natural response and a level of detail that is not normally captured by large basins, against which global model performance is conventionally judged; (ii) the method focuses on regional analysis, meaning that local-scale extremes (which could be driven by specific combinations of local meteorological events and catchment properties that are not necessarily representative of the whole region) are smoothed out, while large-scale generating mechanisms are captured—furthermore, the spatial scale of the regional time series is consistent with that of the global models; and (iii) RDI and RFI are derived from time series anomalies: when applied to global models, the anomalies are based on the internal variability of each model-simulated flow, hence eliminating the effects of systematic biases in the models.

The analyses of simulated RDI and RFI thus concentrate on evaluating whether large-scale high and low flow events are reproduced by the global models at the right time and with the right intensity, duration, and variability, compared to the observed data. Note, however, that while historical RDI and RFI have been shown to be powerful indicators of regional extremes, (i) they are not extremes per se, as they are defined here as 10% anomalies; and (ii) they are subject to limitations (for discussion see Hannaford et al. 2011) because of the uneven distribution of hydrological records used to derive the drought and flood catalogues. Because of these limitations, the RDI and RFI from even a “perfect” model might not be identical to the regional observed catalogues, but the comparison of simulated indices with observed indices should provide a solid basis on which to consistently evaluate performance of multiple global models.

This paper compares outputs from three contrasting large-scale hydrological models, all driven by the same meteorological forcing data, with observed regional high and low flow catalogues over the period 1963–2000, in an attempt to assess the validity of their reproduction of regional runoff “extremes” in Europe. These global models are part of the Water Model Intercomparison Project (WaterMIP) (Haddeland et al. 2011). The current study uses the European RDI and RFI historical catalogues as a benchmark to assess and compare how well large-scale hydrological models reproduce large-scale hydrological phenomena in Europe. A complete explanation of the results in terms of the formulation and process descriptions of the models, and extending to other models, are beyond the scope of this paper. Future work, which is already underway, will examine these issues in more detail.

2. Data and methodology

a. Observed data

The streamflow dataset used to validate model outputs is based on an updated version of the European Water Archive—a dataset of daily streamflow records for 10 countries across Europe, the derivation of which is discussed by Stahl et al. (2010). The dataset comprises catchments with minimal anthropogenic disturbances on flow regimes, monitored by gauging stations regarded to have good hydrometric performance with records from 1961 to 2005. An additional selection criterion was that all catchments should be smaller than 1000 km² (to ensure a likelihood of minimal disturbance), although in a small number of cases larger catchments were included if influences could be proved to be negligible. Additional data from undisturbed catchments were sourced from the U.K. benchmark catchments (Hannaford and Marsh 2008) and from Banque Hydro in France (Prudhomme and Sauquet 2007). The total dataset consists of 579 gauging stations. The distribution of stations over Europe is somewhat uneven (see Fig. 1) with high densities of stations in some areas (e.g., Germany) and limited data in areas that are heavily affected by anthropogenic disturbances (e.g., northern France and the Benelux countries). No data were available across the majority of southern or eastern European countries.
b. Simulated data

For this study, outputs from three of the WaterMIP models were used [Joint U.K. Land Environment Simulator (JULES), Water Global Assessment and Prognosis (WaterGAP), and Max Planck Institute Hydrological Model (MPI-HM)], covering the range of model types discussed in Haddeland et al. (2011). JULES is the land surface component of a climate model, while WaterGAP has been developed to study water resources, and MPI-HM is an intermediate type. Their main characteristics, including input meteorological variables and the schemes used for evapotranspiration, runoff generation, and snowmelt processes are summarized in Table 1, while brief descriptions of the models are provided below.

All global models have the same 0.5° spatial resolution and were run over the period 1963–2001 using the same meteorological input data: the Water and Global Change (WATCH) forcing data (WFD; Weedon et al. 2010). As the models have originally been developed for global simulation, it is difficult to accurately assess the minimum scale they can work on. This is likely to depend on the accuracy of the input data but also on the catchment characteristics and dominant processes in the catchment (which can vary in time even at the same location). In particular, processes with substantial variability at smaller scales than the 0.5° spatial resolution of the input data used here may not be well captured—perhaps more significantly for processes dependant on orography (e.g., in the Alps) and/or meteorology (e.g., convective storms) than for the energy balance processes, for which all models have a representation of subgrid variability. For the study, our main assumption is that runoff simulations at 0.5° are generally reasonable.
The simulations considered “naturalized” conditions when direct anthropogenic effects such as dams and water abstraction were not included in the models. This is consistent with the use of observations from undisturbed catchments. For comparison with the observed data, total runoff (the sum of surface and subsurface flows) was used as reference data to generate simulated RDI and RFI daily time series. Runoff, rather than routed discharge, was considered so that the individual anomalies of all grids in the regions, and hence their spatial pattern, were accounted for without any smoothing added by the hydrological routing. WaterGAP applies a correction factor on runoff to match observed river discharge in major rivers across the globe and evapotranspiration is adjusted accordingly. Neither JULES nor MPI-HM were calibrated for this exercise, although they may have been calibrated for previous studies.

JULES (Best et al. 2011; Clark et al. 2011) is the land surface scheme used in the climate models of the Met Office. It includes mechanistic descriptions of the processes that control the exchanges of energy, momentum, water, and carbon between the land surface and the atmosphere. The energy balance of the surface is calculated on a time step of one hour or less and there is a multilayer snow model. Fluxes of water and heat in the soil are represented using four soil layers with a total depth of 3 m. In the configuration used in this exercise, runoff can occur through two main mechanisms: infiltration excess surface runoff and drainage through the bottom of the soil column. Drainage is calculated as a Darcian flux assuming zero gradient of matric potential (Best et al. 2011).

MPI-HM was developed at the Max Planck Institute for Meteorology based on research from Hagemann and Dümenil (1998) and Hagemann and Dümenil-Gates (2003). In this model, lateral flow movements are considered at the global scale. A simplified land surface scheme (SL) is used in combination with a hydrological discharge model (HD). SL uses daily time series of precipitation and temperature input, and the soil is represented as a single layer. HD describes overland flow, river flow, and base flow movement based on the spatial distribution of topography, slope, lakes, and wetlands within each grid box.

The hydrological modeling system WaterGAP has been developed to simulate the distribution and availability of water resources at the global scale. WaterGAP combines a global hydrological model (Döll et al. 2003) with several global water use models (Flörke and Alcamo 2004; Alcamo et al. 2003; Döll and Siebert 2002). The global hydrological model simulates the continental water cycle incorporating spatially distributed physiographic characteristics and climate factors. The global water use model estimates water withdrawals and water consumption for households, manufacturing, energy production, livestock, and irrigation, taking into account basic socioeconomic factors. WaterGAP was calibrated against more than 1560 stations worldwide based on observed discharge data provided by the Global Runoff Data Centre (GRDC) 2009 and the WFD dataset. The minimum drainage basin area considered for calibration was 9000 km², and intermediate catchment areas were generally larger than 20 000 km². Depending on the data availability for each gauging station, the model was calibrated to discharge time series of 30 years. Because of long computing times, WaterGAP was calibrated for one parameter: the runoff coefficient (see Table 1).

---

**Table 1. Main characteristics of models considered in the study after Haddeland et al. (2011).**

<table>
<thead>
<tr>
<th>Model name</th>
<th>Model time step</th>
<th>Meteorological forcing variables</th>
<th>Energy balance</th>
<th>Evapotranspiration scheme</th>
<th>Runoff scheme</th>
<th>Snow scheme</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPI-HM</td>
<td>Daily</td>
<td>$P$ and $T$</td>
<td>No</td>
<td>Thornthwaite Saturation excess/Beta function</td>
<td>Degree day</td>
<td></td>
<td>Hagemann and Dümenil-Gates (2003), Hagemann and Dümenil (1998)</td>
</tr>
</tbody>
</table>

---

$^a$ Model name in bold is classified as a land surface model; the others are global hydrological models.

$^b$ Rainfall rate ($R$), snowfall rate ($S$), precipitation ($P$; rain or snow distinguished in the model), mean daily air temperature ($T$), maximum daily air temperature ($T_{max}$), minimum daily air temperature ($T_{min}$), wind speed ($W$), specific humidity ($Q$), longwave radiation flux (downward) ($LW$), longwave radiation flux (net) ($LW_n$), shortwave radiation flux (downward) ($SW$), and surface pressure ($SP$).

$^c$ Bulk formula: bulk transfer coefficients are used when calculating the turbulent heat fluxes.

$^d$ Beta function: runoff is a nonlinear function of soil moisture.
c. Large-scale deficiency and exceedance index time series

The methodology used in this paper follows that of the European catalogue of regional droughts in Europe (Lloyd-Hughes et al. 2010; Hannaford et al. 2011) and is based on the RDI concept, first introduced by Stahl (Lloyd-Hughes et al. 2010; Hannaford et al. 2011) and is the European catalogue of regional droughts in Europe. The construction of the regional time series of RDI follows three steps (the alternative methodology for RFI is given in parentheses):

1) Definition of daily deficiency (exceedance) index time series, subsequently referred to as DI and EI, respectively. For each available river flow record, the flow on a given day is compared to a daily varying low (high) flow threshold, and is replaced by a single index equal to 1 when the flow is less (greater) than or equal to the threshold, and 0 otherwise [Eq. (1)]. By construction, the resulting DI (EI) time series are binary. The use of a local daily varying threshold removes the influence of streamflow seasonality and geographical disparity between records. In this study, the flow exceeded 90% (10%) of the time, Q90 (Q10), is used as a threshold and is defined as follows: for a given Julian day \( d \), the daily varying Q90(\( d \)) [Q10(\( d \))] is calculated by ranking all historical flow values across the 31 days centered on day \( d \) (i.e., 15 days either side). The 31-day window increases the size of the sample and ensures a more robust estimate of the flow duration curves (ranked flows) and thus Q90(\( d \)) and Q10(\( d \)):

\[
\begin{align*}
\text{DI}(d) &= 1 \quad \text{if} \quad Q(d) \leq Q90(d) \\
\text{DI}(d) &= 0 \quad \text{if} \quad Q(d) > Q90(d) \quad \text{and} \\
\text{EI}(d) &= 1 \quad \text{if} \quad Q(d) \geq Q10(d) \\
\text{EI}(d) &= 0 \quad \text{if} \quad Q(d) < Q10(d).
\end{align*}
\]

2) Sites experiencing a deficiency (exceedance) at the same time are grouped into homogeneous regions using cluster analysis (Stahl 2001; Prudhomme and Sauquet 2007). The clustering was performed by applying an agglomerative (hierarchical) technique using the Ward method and a binary Euclidean distance measure. Tests in France showed that groups are relatively robust to the measurement period. Small manual adjustments were made to ensure that regions are spatially continuous in their extent (Hannaford et al. 2011; Prudhomme and Sauquet 2007). Note that work by Gudmundsson et al. (2011) suggests that there is little difference in the spatial structure of clusters grouped according to annual Q95 and Q90 (Q5 and Q10) for low (high) flow. The homogeneity of a cluster \( c \), \( H(c) \), is empirically calculated [Eq. (2)] as the departure between the theoretical cumulative density of RDI (RFI) and the empirical cumulative distribution \( F(x) \); by construction, 90% of the RDI (RFI) series should be equal to 0, and 10% should be equal to 1, if all stations had identical DI (EI) time series:

\[
H(c) = \int_0^1 |F(x) - p|, \tag{2}
\]

with \( H(c) \) the homogeneity index, \( F(x) \) the cumulative distribution of the RDI (RFI) of cluster \( c \), and \( p = 0.9 \). \( H(c) \) close to zero suggests high homogeneity.

3) Regional deficiency (flood) index time series are defined for each homogeneous region as the arithmetic mean of the DI (EI) series of all sites within the regions for each day of record [Eq. (3)]. The RDI(\( d \)) [RFI(\( d \))] represents the proportion of catchments in the region that experience low (high) flows that fall below (rise above) the threshold on the same Julian day \( d \). A maximum value of 1 represents a flow deficiency (exceedance) at all sites in the region, thus indicating a large-scale low (high) flow event:

\[
\text{RDI}(d) = \frac{1}{M} \sum_{i=1}^{M} \text{DI}_i(d) \quad \text{and} \quad \text{RFI}(d) = \frac{1}{M} \sum_{i=1}^{M} \text{EI}_i(d), \tag{3}
\]

with \( M \) the number of records in the region.

Historical RDI and RFI catalogues are derived from the 579 streamflow records discussed earlier and based on 23 homogenous regions (Fig. 1) modified from Hannaford et al. (2011); note that southeast Great Britain is a single geographic region instead of two sub-regions divided according to groundwater influence as it was in the previous study.

Simulated RDI and RFI time series are derived using an identical procedure as outlined above but applied to the total runoff (the sum of surface and subsurface runoff) with thresholds (Q90 and Q10) calculated for each grid cell. Once simulated DI and EI time series are generated for each grid cell, simulated RDI and RFI daily time series are calculated [using Eq. (3)] by considering all cells whose centroid is within the geographical boundaries of the 23 European regions.

It should thus be noted that the use of the RDI and RFI means that the procedure does not assess the absolute magnitude of runoff compared to observed discharge, but evaluates whether the periods of simulated and observed high and low flows occur at the same time and with the same spatial extent as observed.
3. Results

Three complementary analyses were undertaken to assess the ability of the large-scale models to reproduce large-scale hydrological extremes:

- Statistical analyses of RDI and RFI time series derived from river flow observations (the historical catalogues) against those derived from simulated runoff. Three goodness-of-fit measures are used and shown in Table 2 for six case study regions:

  (i) relative mean error (RME; %) as a measure of the overall bias in the model;
  (ii) ratio between the standard deviation of simulated and observed RDI/RFI as a measure of the overall reproduction of the spread in the series; and
  (iii) Spearman (ranked) correlation (rho; Spearman 1944) between observed and simulated RDI–RFI as a measure of the overall reproduction of the timing of the events. This measure is preferred to the Pearson correlation analysis as it does not assume a Gaussian distribution for the time series.

- Graphical analyses of the complete RDI and RFI time series for six contrasting regions. This allows rapid identification of whether the major historical large-scale events have been simulated.

- For two specific events (the high flows of 2000 and the drought of 1975/76), detailed comparisons of the RDI and RFI time series. This allows the temporal structure of the simulated RDI and RFI to be highlighted at a very fine resolution and enables the identification of differences between the three models during notable high–low flow periods.

a. Statistical measures

Below we discuss each measure of goodness of fit, with values for six selected regions being shown in Table 2.

1) RELATIVE MEAN ERROR

While, by construction, DI (EI) time series for each cell are populated by 10% of 1 and 90% of 0, only a region in which the DI (EI) time series are identical within all cells has RDI (RFI) comprising 10% of 1 and 90% of 0. In reality, RDI (RFI) time series have a value of 1 only when the whole region experiences an anomaly on the same day (i.e., when DI–EI of all stations or all grid cells belonging to that region are equal to 1 for that day). The magnitude of RDI (RFI) expresses the spatial coherence of the low (high) flow event, with the duration (e.g., time with index above threshold) depicting the length of the event. The mean relative error hence measures whether a model tends to overestimate the spatial coherence [i.e., there are more high RDI (RFI) values than observed; positive RME] or underestimate it (negative RME) and by how much.

Generally, RME is positive for RDI and negative for RFI (Table 2), suggesting all models tend to overestimate the spatial coherence and length of drought events but underestimate high flow events. The model with overall RME of smallest magnitude was MPI-HM for both RDI (2.43% overall) and RFI (−2.42%). RME is of similar magnitude and sign for JULES and WaterGAP, and larger for RDI (overall +7.98 and +9.27, respectively) than for RFI (overall −4.22 and −3.18, respectively). There is a marked difference in the performance by regions (not shown), with southern France, the French Alps, the High Alps, and East Germany amongst the regions with the largest RME (between 15% and 22% for RDI for JULES and WaterGAP), while northwest and southwest Great Britain, northern France, and north Germany show the lowest RME for JULES and WaterGAP (less than 10% for RDI) but not always for MPI-HM.

2) REPRODUCTION OF THE SPREAD

The variability of RDI (RFI) time series depicts the level of spatial coherence of low (high) flow events: as already discussed, the lowest variability would be achieved for a region where low (high) flow anomalies occur simultaneously. With spatial coherence of events reducing, their number increases as different areas of the region experience a flow anomaly at different times (by definition, 10% of days have DI (EI) associated with a flow anomaly). The larger the standard deviation is, the less coherent, the more numerous, and, by extension, the shorter the events are. For RDI, all three models overestimate the standard deviation of the time series in all regions but northwest Great Britain and southern Scandinavia, and for WaterGAP the French Southern Alps (not shown). The largest overestimation of the variability is found for MPI-HM but exceeds 30% for all three models in four regions (albeit different ones) and is greater than 10% for 14 (JULES and MPI-HM) or 15 regions (WaterGAP). In contrast, overall time series variability is better reproduced for RFI, and within ±10% for 7 (JULES), 9 (WaterGAP), and 10 (MPI-HM) regions. This could be because high flow events are already shorter and less spatially coherent than low flow episodes. High flow events are also often predominantly driven by precipitation, which is a direct input to the models, and so a good estimate of precipitation will tend to force realistic variability of high flows. In general, all of the models (and in particular MPI-HM) tend to generate high flow episodes that are shorter and not
TABLE 2. Goodness-of-fit assessments for the RFI and RDI series and two periods of low flow (1976) and high flow (2000) for six European regions. RDI is always equal to 0 in 2000 in southeast Great Britain as modeled by JULES: number of stations = NS; homogeneity index = $H$; Spread = ratio of standard deviation of simulated series over standard deviation of observed series; Rho = Spearman correlation; Rho 76 = estimate for 19 Apr 1975 to 19 Jun 1977; Rho 00 = estimate for 1 Jan to 31 Dec 2000. Bold values are significant at the 95% level (only for Rho).

<table>
<thead>
<tr>
<th>Region</th>
<th>S.E Great Britain</th>
<th>N.W Spain</th>
<th>W &amp; C France</th>
<th>High Alps</th>
<th>E Germany &amp; Czech</th>
<th>N.W Scandinavia</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RFI</td>
<td>RDI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NS</td>
<td>H</td>
<td>Spread</td>
<td>Rho</td>
<td>Rho 76</td>
<td>Rho 00</td>
</tr>
<tr>
<td></td>
<td>37</td>
<td>8.30</td>
<td>10.70</td>
<td>7.40</td>
<td>8.80</td>
<td>8.20</td>
</tr>
<tr>
<td></td>
<td>Rho 76</td>
<td>0.69</td>
<td>0.52</td>
<td>0.49</td>
<td>0.39</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>Rho 00</td>
<td>0.73</td>
<td>0.67</td>
<td>0.50</td>
<td>0.39</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>37</td>
<td>37</td>
<td>48</td>
<td>35</td>
<td>37</td>
<td>37</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Region</th>
<th>S.E Great Britain</th>
<th>N.W Spain</th>
<th>W &amp; C France</th>
<th>High Alps</th>
<th>E Germany &amp; Czech</th>
<th>N.W Scandinavia</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RME</td>
<td>Spread</td>
<td>Rho</td>
<td>Rho 76</td>
<td>Rho 00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-4.30</td>
<td>1.25</td>
<td>0.61</td>
<td>0.69</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-8.16</td>
<td>1.11</td>
<td>0.54</td>
<td>0.52</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-11.67</td>
<td>0.91</td>
<td>0.52</td>
<td>0.49</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-1.65</td>
<td>1.03</td>
<td>0.39</td>
<td>0.34</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-4.69</td>
<td>0.92</td>
<td>0.47</td>
<td>0.30</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-1.48</td>
<td>0.94</td>
<td>0.62</td>
<td>0.62</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td></td>
<td>37</td>
<td>8.47</td>
<td>1.23</td>
<td>0.77</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>1.45</td>
<td>0.50</td>
<td>0.46</td>
<td>0.27</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10.70</td>
<td>1.07</td>
<td>0.61</td>
<td>0.55</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7.40</td>
<td>1.07</td>
<td>0.59</td>
<td>0.59</td>
<td>0.79</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8.80</td>
<td>1.17</td>
<td>0.68</td>
<td>0.68</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8.20</td>
<td>1.09</td>
<td>0.62</td>
<td>0.62</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10.70</td>
<td>1.04</td>
<td>0.62</td>
<td>0.62</td>
<td>0.84</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Region</th>
<th>S.E Great Britain</th>
<th>N.W Spain</th>
<th>W &amp; C France</th>
<th>High Alps</th>
<th>E Germany &amp; Czech</th>
<th>N.W Scandinavia</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RME</td>
<td>Spread</td>
<td>Rho</td>
<td>Rho 76</td>
<td>Rho 00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-7.69</td>
<td>1.16</td>
<td>0.59</td>
<td>0.65</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-9.28</td>
<td>1.02</td>
<td>0.49</td>
<td>0.23</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-12.50</td>
<td>0.89</td>
<td>0.55</td>
<td>0.41</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-2.59</td>
<td>1.07</td>
<td>0.47</td>
<td>0.34</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-3.72</td>
<td>0.91</td>
<td>0.48</td>
<td>0.30</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-2.71</td>
<td>0.93</td>
<td>0.65</td>
<td>0.67</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8.18</td>
<td>1.24</td>
<td>0.77</td>
<td>0.91</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>13.81</td>
<td>1.18</td>
<td>0.43</td>
<td>0.43</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10.99</td>
<td>1.04</td>
<td>0.77</td>
<td>0.43</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td></td>
<td>21.13</td>
<td>1.38</td>
<td>0.50</td>
<td>0.77</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>16.63</td>
<td>1.09</td>
<td>0.64</td>
<td>0.50</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.90</td>
<td>1.00</td>
<td>0.69</td>
<td>0.64</td>
<td>0.44</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Region</th>
<th>S.E Great Britain</th>
<th>N.W Spain</th>
<th>W &amp; C France</th>
<th>High Alps</th>
<th>E Germany &amp; Czech</th>
<th>N.W Scandinavia</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RME</td>
<td>Spread</td>
<td>Rho</td>
<td>Rho 76</td>
<td>Rho 00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-1.99</td>
<td>1.33</td>
<td>0.56</td>
<td>0.59</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-13.28</td>
<td>1.09</td>
<td>0.48</td>
<td>0.50</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-9.29</td>
<td>0.95</td>
<td>0.50</td>
<td>0.45</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-1.07</td>
<td>1.15</td>
<td>0.41</td>
<td>0.34</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-4.89</td>
<td>0.96</td>
<td>0.40</td>
<td>0.30</td>
<td>0.46</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>1.06</td>
<td>0.51</td>
<td>0.52</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6.65</td>
<td>1.22</td>
<td>0.63</td>
<td>0.74</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18.57</td>
<td>1.42</td>
<td>0.60</td>
<td>0.56</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.73</td>
<td>1.03</td>
<td>0.67</td>
<td>0.58</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10.08</td>
<td>1.33</td>
<td>0.51</td>
<td>0.50</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>13.94</td>
<td>1.21</td>
<td>0.53</td>
<td>0.50</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.90</td>
<td>1.02</td>
<td>0.59</td>
<td>0.62</td>
<td>0.57</td>
<td></td>
</tr>
</tbody>
</table>
as spatially coherent as those observed in Europe. Note that JULES is the only model that underestimates RFI variability by over 10% in two regions (southern France and western/central France), suggesting that it generates high flow episodes that are too long and over too large a spatial domain in those regions.

3) CORRELATION BETWEEN OBSERVED AND SIMULATED DATA

In addition to the relative magnitude and duration of high and low flow events, a global model should be able to accurately simulate their timing (onset and decay). Spearman’s rank (Spearman 1944), which does not require time series to follow a normal distribution, was used to examine correlations between observed and simulated RDI and RFI. Figures 2 and 3 map the correlation coefficients between historical catalogues and simulated time series respectively for high flows (RFI) and low flows (RDI) for each of the 23 European study regions alongside background information on station density used to derive the historical catalogues (number of stations divided by the number of grid cells within a given region) and region homogeneity \( H \) (see Eq. (2); called RDIarea in Hannaford et al. 2011). All plotted correlations are significant at \( \alpha = 0.05 \).

Correlation coefficients are generally greater for low flows than for high flows across most regions and models. This is likely to reflect the difference in the temporal structure of RDI and RFI, as droughts and low flow episodes tend to develop slowly and persist for longer than floods and high flow events, which are more frequent but quicker to subside. This large day-to-day
variability of RFI compared to RDI will tend to result in smaller correlation coefficients as a result of any errors in modeled timing of the event.

The value and range of correlations for high and low flow events are of similar magnitude for all three models. Correlations show clear regional patterns with all three models systematically reproducing large-scale hydrological extreme time series in different regions to varying extents. For example, the spatiotemporal variation of both extremes (as measured by the correlation) in the southeast Great Britain region is consistently well reproduced by the three models. Conversely, simulated RDI and RFI of the Pyrenees and French Southern Alps regions are consistently poorly correlated with the observed series, despite RME of approximately 10% and 25% for RDI and RFI, respectively, for the Pyrenees. This suggests that all models had difficulty in reproducing the hydrology of mountainous regions (and in particular the occurrence and duration of large-scale low and high flow events), perhaps linked to the spatial resolution of the models and the driving climate data (0.5°), which is relatively coarse for adequately describing the very variable topography of the regions or the weather in mountainous areas. In other high mountainous regions, such as the High Alps, poor correlation is found for the high flows (between 0.39 for WaterGAP and 0.47 for JULES) while performance is slightly better for low flows (between 0.51 for MPI-HM and 0.55 for WaterGAP). Note, however, that the larger RME for RDI than RFI suggests a systematic bias in simulating dry episodes by all models. In contrast, both regional index time series for northwest Scandinavia are relatively well reproduced (correlations between 0.59 and 0.61 for RDI and 0.51 and 0.65 for RFI), suggesting that the hydrological
processes underlying snowmelt-dominated hydrology can be relatively well reproduced by all models (also reflected by a RME of less than 5%). Haddeland et al. (2011) suggest that differences in the modeled partitioning of precipitation into rainfall and snowfall (which is generally calculated using a temperature threshold close to 0°C) is responsible for the large spread in the estimated snowfall, particularly in regions where winter temperatures are closer to 0°C, and, by extension, for the timing of the simulated runoff. MPI-HM and WaterGAP both partition precipitation depending on temperature, while JULES takes snowfall and rainfall inputs directly, so it is possible that all three models used rather different partitioning. This could explain the large differences in RDI and RFI simulated by the three models and the relatively poor performance of WaterGAP, JULES, and MPI-HM in reproducing large-scale low and high flow events in the Pyrenees and Alps compared to better performance in Scandinavia (which is much colder and hence all models tend to agree on snowfall).

With the exception of two British regions, large-scale high flow events tend to be better reproduced in western and northern European regions, influenced by a maritime climate, than in the more continental regions of central Europe. This could reflect the regional differences in flood-generating mechanisms, mainly associated with frontal precipitation (hence with relatively large spatial coherence and relatively large persistence) in maritime climates and convective precipitation (hence driven by local, short-lived intense events) in continental climates. The spatial scale of the modeling grid (0.5°) and uncertainty in the measurement of highly variable convective rainfall and its timing in the forcing data all tend limit the reproducibility of such local precipitation events and, hence, of the resulting simulated high flows.

Intermodel differences in model performance at both low and high flows are not substantial in terms of the correlation coefficients. For low flows, for which the driving meteorological conditions are more similar in different climatic regions, model performances are relatively similar. Overall, WaterGAP shows the largest correlations, followed by MPI-HM and then JULES, where correlations lower than 0.55 are found for 12 regions (this is the case for only 4 and 6 regions for WaterGAP and MPI-HM, respectively). For high flows, WaterGAP and JULES have similar correlation coefficients, both outperforming MPI-HM, but there is no model that is systematically associated with the largest correlation coefficients in all regions (5, 7, and 4 regions show correlation coefficients greater or equal than 0.55 for WaterGAP, JULES, and MPI-HM, respectively). Regionally, JULES seems to perform best in France (all regions except Pyrenees and Alps), Scandinavia, and parts of Germany, while MPI-HM performs best in Great Britain, Germany, and Scandinavia, outperforming JULES in those last two areas. There are no regions where WaterGAP systematically performs well or poorly for both high and low flows.

4) Influence of Regional Characteristics on the Simulated Regional Indices

Station density reflects how accurately the historical RDI and RFI catalogues capture the spatiotemporal development of regional hydrological extremes: the more stations and the better their distribution in the region, the more representative the regional indices. The quantity H is a measure of the region homogeneity (i.e., whether the entire region experiences a flow anomaly at the same time as historical measurements). There is no strong pattern linking model performance (regardless of the goodness-of-fit measure considered) with station density or region homogeneity (Fig. 4). This suggests that differences in regional characteristics are unlikely to significantly affect the model intercomparison and that the historical catalogues defined by Hannaford et al. (2011) and Parry et al. (2010) can be used as benchmark time series against which to compare large-scale hydrological model performance.

b. Comparing Drought and Flood Catalogues for Observed and Simulated Data

The three goodness-of-fit measures provide an overall summary of similarity between the historical catalogues and the simulated RDI and RFI time series, but they do not provide an adequate indication of the temporal and spatial variability in model performance. For a hydrological model to show skill in its simulation of extreme events, three criteria must be fulfilled:

- accurate timing of presence/absence of extreme events—this is mainly influenced by the meteorological drivers;
- flashiness/smoothness of simulated time series (and in particular of extreme events) similar to those of the observed catalogues—this reflects the rainfall-runoff transformation mechanisms; and
- spatial consistency of extreme events of similar magnitude with that observed—this shows that the spatial pattern of runoff is well reproduced by the models.

Figures 5–8 present RDI and RFI time series obtained from historical observations (left) and simulations for the period 1963–2000 for six contrasting regions: southeast Great Britain, northwest Spain, western and central France, the High Alps, East Germany–Czech Republic, and northwest Scandinavia.
FIG. 4. Scatterplots of region homogeneity (black symbols) or station density (gray symbols) against (top) RME, (middle) spread reproduction, and (bottom) Spearman correlation for (left) RDI and (right) RFI for WaterGAP (circles), JULES (squares), and MPI-HM (diamonds).
European drought occurrence is generally fairly well reproduced by the considered global models across the six regions, although with varying degrees of success (Figs. 5 and 6). In general, the timing of all major drought events is well reproduced, and periods with no major large-scale drought are also represented successfully. However, skill in reproducing drought duration or severity (in terms of regional spatial coherence of the events, as quantified by the RDI) differs between the global models.

JULES tends to overestimate the duration (and occasionally the severity) of droughts; excessive drought length is particularly notable for northwest Spain and East Germany–Czech Republic. This would be consistent with overestimation of evaporation, but poor parameter values or inadequate representation of processes could also contribute. JULES has more success in reproducing the drought characteristics of regions with long periods of observed drought because of higher catchment storage capacities, such as the groundwater-dominated southeast Great Britain, where RDI is reasonably well reproduced by JULES. This could be because the 3 m soil column and slowly responding, drainage-dominated runoff of JULES...
are closer to reality in this area than elsewhere. JULES also shows some ability in modeling the drought characteristics of snowmelt-influenced regions such as the High Alps, where the typical short duration of droughts is well reproduced, albeit with the duration of major events being slightly too long and with the peak intensity too high.

In contrast to JULES, MPI-HM tends to underestimate drought duration and the RDI plots are “noisier,” with periods of deficiency often broken by very short periods of low RDI. This could be explained by limited storage within the model, leading to runoff being generated because of saturated soils following minor precipitation events. This is consistent with the poor reproduction of drought for the southeast Great Britain region, where storage is primarily within groundwater aquifers, and where MPI-HM fails to reproduce the sustained multiseason deficits that are often observed. In contrast, in northwest Scandinavia—where storage is in the form of snowpack and ice but the resulting RDI time series is relatively flashy—MPI-HM reproduces drought characteristics reasonably well. In other regions with low storage capacity and fast-responding catchments,
MPI-HM accurately reproduces longer periods of streamflow deficiency, most notably in the western and central France region. In East Germany–Czech Republic and northwest Spain, typical drought duration is reasonably simulated, but the model generally overestimates the number and severity of major droughts.

WaterGAP shows substantial and generally realistic variation in drought characteristics across the six regions, suggesting that internal processes within the model are perhaps most appropriate for reproducing drought characteristics. Where droughts are predominantly sustained periods of low flows (e.g., in southeast Great Britain), WaterGAP simulates both duration and severity of such episodes well; where the drought regime is dominated by short and frequent low flow periods (e.g., in the High Alps and East Germany–Czech Republic), WaterGAP is also able to reproduce the drought characteristics. Note, however, that for the 1976 episode, WaterGAP overestimates the length of the drought in these regions (see next section for detailed analysis). Finally, as with JULES, WaterGAP simulates long summer droughts in northwest Spain that are not present in the observed catalog.
Investigating regionally important hydrological processes also yields some notable observations of model performance. The observed RDI for the northwest Scandinavia region shows moderate streamflow deficiencies in the January–April period, which abruptly terminate in the following months when snowmelt runoff dominates (Hannaford et al. 2011) and are followed by more sporadic deficiencies. All three models are able to reproduce this process to some extent, with JULES and WaterGAP perhaps the most accurate. Observed RDI for the January–March period for the High Alps shows distinct periods of drought (in the 1960s and early 1970s) or a lack of droughts (through the 1970s and 1980s), which again are likely to reflect interannual differences in storage as ice and snow. This is reasonably well reflected in all three models, but spring–summer droughts in this region (which may be caused by the delayed onset of snowmelt) are simulated to varying extents; JULES does not appear to capture the seasonality, whereas MPI-HM captures a tendency for short, frequent deficiencies in summer. This is consistent with the earlier description of JULES as having relatively large storage.

FIG. 8. As in Fig. 6 but for RFI.
and slowly responding runoff, whereas runoff in MPI-HM is much more responsive. Again we note that snow simulation (and, by extension, its subsequent melt) was identified by Haddeland et al. (2011) as one of the major differences between the models of WaterMIP.

The general patterns of model performance found for high flows are similar to those described for low flows. Note it is more difficult to assess how well the models reproduce the development of large-scale high flow events because such events are usually short (i.e., with RFI time series very variable in time) and there are no sustained periods of high flows across regions (Figs. 7 and 8). This “flashiness” in RFI is generally reproduced by all the global models, but they all tend to simulate occurrence of extremes even when no large-scale high flow events have been historically observed. MPI-HM exhibits the most significant temporal variability in RFI and, to a lesser extent, WaterGAP. However, they both show generally good agreement for the most significant high flow events as well as a good degree of reproducibility of the distribution of “high flow rich” episodes in the period of record.

As is the case for low flows, the modeled high flow periods generated by JULES typically overestimate duration and spatial coherence compared to the corresponding observed RFI; this is particularly notable for the High Alps. Again, it is suggested that this is a result of runoff in JULES being dominated by relatively slowly responding subsurface flow. JULES performs well in other regions, such as East Germany–Czech Republic (other than overestimation of high flows in 1987) and western and central France, where long and coherent high flow episodes are observed. In northwest Spain and southeast Great Britain, JULES also simulates the development of large-scale high flow episodes reasonably well, particularly the damaging 2000/01 floods, but still tends to overestimate high flow magnitude for most events. In Scandinavia, JULES is also able to mimic the “flashy” observed RFI, despite overestimating the persistence of the large-scale events. However, JULES also simulates short-duration high flows in other regions (such as southeast Great Britain, northwest Spain, and western and central France), predominantly clustered in June–October, even when no such temporal pattern has been observed nor simulated by WaterGAP and MPI-HM in those regions. A thorough investigation of model parameterization and model structure would be necessary to understand the simulated response of JULES, which is beyond the scope of this paper.

Akin to the patterns simulated at low flows, the simulated RFI generated from MPI-HM is noisier than the observed RFI across all the regions presented in Figs. 7 and 8 as it generally simulates short events developing very rapidly in all regions. While observed periods of high flows are often reproduced as increased oscillations of the noisy signal (e.g., in East Germany–Czech Republic in December 1974), for most regions high flows are almost randomly distributed throughout the period of record, with few discernable periods of sustained high flows other than the periods of summer high flow in northwest Spain that are relatively accurately reproduced compared to other regions. For regions with flashy RFI such as the High Alps and northwest Scandinavia, the spatiotemporal pattern simulated by MPI-HM is in agreement with observations, but this is not the case for regions with prolonged development of regional hydrological extremes (e.g., southeast Great Britain). Further investigation of model parameterization and model structure, and in particular the infiltration–saturation processes, would help to explain the systematic variability of regional indices simulated by MPI-HM.

For WaterGAP, simulated RFI shows large seasonal variation not apparent in the observed data. Occurrence of summer large-scale high flows, sustained over several months, is usually correctly simulated for southeast Great Britain, western and central France, and northwest Spain, but simulated events in autumn, winter, and early spring tend to be more frequent and shorter than observed. The cause of this seasonal duality is unclear but is unlikely to be due to the modeling of groundwater influence as the resulting smoothing would not be restricted to summer months only, but occur in all seasons. This could be a product of insufficient soil depth, so that only a “small” part of the effective precipitation can be stored in the soil. When the soil column is full of water, surface runoff is generated, simulating flashy events in winter, autumn, and spring. Another factor might be lower potential evapotranspiration estimated by WaterGAP than most other WaterMIP models, which might tend to further increase surface runoff. In general, however, WaterGAP is successful at reproducing the development of large-scale high flow events in the six regions presented here, with temporal patterns of both flashy (High Alps and northwest Scandinavia) and more persistent (groundwater-dominated southeast Great Britain) regimes captured by the simulations.

As was witnessed in mountainous and snowmelt-influenced regions for modeled low flow occurrence, the models have shown an ability to reproduce seasonal snowmelt-driven streamflow responses and, in some years, the onset of high flow events in late spring, a response to the introduction of meltwater into river systems. These effects are particularly evident at high flows in northwest Scandinavia, with observed data showing a relative scarcity of high flow events in the January–April period, particularly prior to 1988, which reflects
the predominantly frozen nature of those catchments at that time of year, with river flow recessions mitigating potential high flows.

c. Simulation of notable high and low flow periods

This section investigates two major episodes of large-scale hydrological extremes in detail: the drought of 1975/76 (Zaidman et al. 2001) and the high flows of 2000 (Marsh and Dale 2002). The observed catalogues and simulated RDI and RFI time series are presented for each model for the six contrasting regions (Figs. 9 and 10 respectively), along with the correlation coefficients corresponding to the two time periods (Table 2). Two questions are posed: (i) are the onsets and termination of events reproduced at the correct time? And (ii) is the temporal development of a large-scale event (spatial coherence and temporal variability) reproduced accurately?

1) THE DROUGHT OF 1975/76

Although comparatively short in duration, the 1975/76 drought was the most spatially coherent and hydrologically intense drought on a European scale in the last 50 years. Drought developed slowly from late summer–autumn, becoming coherent from winter 1975/76 (United Kingdom) to late spring (France and Germany), with the majority of European regions exhibiting their most significant deficits in summer 1976. The drought terminated in September in most regions, although it extended into the autumn for most southern European regions (Parry et al. 2010).

Generally, WaterGAP captures the onset of the drought well, but not in all regions (Fig. 9). In contrast, MPI-HM simulates an early onset of the drought in southeast Great Britain, Germany, and the High Alps, and JULES simulates a delayed and prolonged onset in southeast Great Britain, western and central France, and the High Alps. The end of the drought is better captured by MPI-HM than by the other two models; simulated deficiencies are too prolonged in WaterGAP and JULES, especially in northwest Spain (minor drought), France, and East Germany.

In terms of drought development, MPI-HM reproduces temporal patterns of drought development reasonably well in most regions despite generally exhibiting too much variability (southeast Great Britain in particular, but also western and central France, the High Alps, and East Germany). WaterGAP tends to overestimate the magnitude of droughts (RDI too high) while JULES tends to underestimate temporal variability while overestimating RDI, producing very long, intense simulated droughts. Spain, where the 1975/76 drought was the least intense, was accurately simulated by all models as the region with the lowest drought intensity, but RDI was still largely overestimated by both WaterGAP and JULES. In Scandinavia, where only minor drought conditions were observed, low RDI was accurately simulated by all three models. In the High Alps, no model could reproduce the spatiotemporal development of the drought satisfactorily, although the shape of the peak of the drought in May and June 1976 is reasonably well captured by WaterGAP and JULES.

2) HIGH FLOWS IN 2000

During 2000, there were two major high flow events in Europe: in the spring, Germany and central Europe experienced flooding, but the rest of Europe remained unaffected (Fig. 10). In the autumn, persistent and extensive precipitation leads to widespread flooding in Great Britain (with total discharges for October–December the highest in a record dating back to 1940), and to a certain extent in France. As previously mentioned, the ability to reproduce the development of large-scale high flow episodes by models is difficult to assess because of the high variability of the high flow episodes. Nevertheless, Fig. 10 suggests that both WaterGAP and MPI-HM reproduce the start of large-scale floods (RFI) well, while they generally start later than observed in JULES (Spain, southeast Great Britain, the High Alps, and East Germany). All models struggle to simulate the termination of a flood, but this is also difficult to identify from observed RFI catalogues.

In terms of development of the event, JULES shows the least similarity with observed RFI. It simulates episodes that are slow developing, spatially coherent (i.e., high RFI), and prolonged, with a temporal variability much lower than that observed: once the model simulates an extreme, the whole region rapidly experiences the same extreme condition, with low spatial variation generating high intensity RFI for sustained periods. WaterGAP and MPI-HM, which both tend to produce time series that are more variable (i.e., less spatial coherence for flow anomalies and large day-to-day variability), simulate more realistic events. But while the general shape of the time series is reproduced reasonably well, the timing and magnitude of RFI peaks is not always synchronous with that of the observed catalogues.

4. Discussion and conclusions

For the first time, global models are assessed for their ability to reproduce large-scale hydrological extreme events using a new methodology to quantify the spatiotemporal development of low and high flow events as a proxy for hydrological extremes. Historical European
FIG. 9. RDI for the 1975/76 drought for six contrasting regions: (top to bottom) southeast Great Britain, northwest Spain, western and central France, the High Alps, East Germany–Czech Republic, and northwest Scandinavia. For each region, observed (gray) and simulated (black) 10-daily RDI are shown for (top to bottom in each region) WaterGAP JULES, and MPI-HM.
Catalogues of droughts (RDI) and high flows (RFI) were derived using river flow measurements across Europe. Using the same method, total runoff simulated by large-scale hydrological models was used to generate simulated RDI and RFI time series for the same European regions, which were directly comparable with the observed European catalogues. Since it is based on anomalies, the method evaluates to what extent the high and low flow
events are simulated at the right time (onset and cessation of events) with the same duration and with the same temporal variability and spatial consistency as observed. This is a new way for modelers to evaluate performance of their model, taking into account the intensity and spatial and temporal variability of runoff in evaluating events at the large scale. Note, however, that the analysis is not designed to evaluate the skill of global models in reproducing the magnitude (in absolute terms) of hydrological extremes.

The method was implemented for Europe using three global models, with quite different structures, from the WaterMIP project: JULES, WaterGAP, and MPI-HM. All were run with the WATCH forcing data at the same 0.5° grid resolution across Europe. Results showed that all three models have broadly comparable performance in terms of goodness-of-fit measures (RME, ratio between simulated and observed standard deviation, and Spearman correlation) between observed and simulated European RDI and RFI time series, with WaterGAP performing best in terms of correlation for both high and low flows. While RDI magnitude and variability is systematically overestimated (positive RME and SD ratio greater than 1), the shorter nature of high flow episodes is well captured by all global models (small RME and relatively good reproduction of the variability). In contrast, the timing of the high flow events and their relative magnitude is less well simulated than that of low flow events, possibly owing to the shorter nature and generally lower spatial coherence of floods, which are also often a more direct response to meteorological conditions. Correlation coefficients greater than 0.5 were found for most European regions for low flows and a majority were over 0.4 for high flows (all significant at 95% level), with the worst-performing regions consistently showing weak correlations across all models. The generally good associations for the time series are likely to be influenced by the prevalence of periods with low or zero RDI (RFI), with zeros constituting 90% of a homogenous time series by definition, but they also demonstrate that the nonoccurrence of extremes is generally simulated by all models during the correct periods. Future work could address this issue by employing statistical analyses that assess magnitude of correlation for only nonzero values.

For low flows, all three global models generally capture the broad-scale characteristics of droughts in six contrasting European regions (southeast Great Britain, western and central France, northwest Spain, the High Alps, East Germany–Czech Republic, and northwest Scandinavia) with differing degrees of success. For high flows, there is some congruency between the models and observations, but this is dependent on both the model and the region under consideration. The differences in historical spatiotemporal development of floods and droughts are consistently well captured by the three considered models. However, some general weaknesses emerge, in particular for high flow periods.

Three factors can impact regional differences between observed and simulated RDI and RFI: (i) the station density influences the observed RDI and RFI to a degree, and low station density could lead to an observed index not truly representative of the regional hydrological extremes; (ii) the gridded precipitation input (WFD) might not accurately reflect the precipitation events driving hydrological extremes—this would be the case particularly for regions with large spatial variability of precipitation and weather (e.g., convective storms in continental regions and snowmelt in mountainous regions); and (iii) the rainfall-runoff processes are complex and strongly influenced by the catchment properties, and in particular by infiltration and groundwater processes that are not necessarily described adequately in the models. While the first two factors are independent of the global models and the resulting errors are shared by all simulated RDI and RFI, the third factor will be reflected by the performance measured for the different global models, as the simulated RDI and RFI time series will highlight those differences in modeling the hydrological processes.

When the simulated RDI and RFI from all three models are compared, some clear similarities and differences emerge. JULES tends to produce prolonged, highly spatially coherent events, both for high and low flows. The events develop more slowly than the historical catalogues have shown, and reach and sustain spatial coherence greater than observed. This suggests a strong spatial coherence of the model parameterization, perhaps with insufficient variability in the processes in nearby grid cells. (Note, however, that small regions such as the High Alps are described by very few grid cells and hence, by construction, would inherit low spatial variability in the simulated RDI/RFI.) This is consistent with the fact that, for the configuration of JULES used here, runoff is dominated by subsurface drainage, which gives a relatively slow response. In contrast, MPI-HM shows very high variability in the simulated RDI and RFI time series, regardless of the region, with a more rapid onset of extreme events than observed. This is particularly notable for regions with significant water storage capacity (e.g., southeast Great Britain). The attenuated hydrological regime, which leads to a prolonged development of droughts, and to a lesser extent of high flow events, is not reproduced by MPI-HM, from which simulations remain flashy. This would suggest that infiltration and groundwater storage...
are perhaps underrepresented in MPI-HM, and saturation is reached quicker for the grid cells than is observed. WaterGAP shares some of the issues relating to high variability with MPI-HM and some strong spatial coherence of simulated events with JULES, but neither is dominant. Of the three considered global models, WaterGAP is arguably the best suited to reproduce most regional characteristics of large-scale hydrological extremes, but it remains deficient in adequately reproducing their onset and, to a lesser extent, spatiotemporal variation.

It is encouraging to see the reasonable similarity between observed and simulated European RDI, suggesting that simulated total runoff gives a credible indication of regional-scale drought response. While differences between the various model outputs and observed data have been highlighted and should be accounted for in interpreting the results of any future runs, the results presented herein suggest the global models could be expected to provide relatively realistic simulations of future drought characteristics in terms of seasonality, duration, and magnitude in Europe. This is a useful finding, as it complements the range of other studies that have shown the utility of large-scale models in reproducing average runoff conditions (e.g., Haddeland et al. 2011; Lohmann et al. 2004) and suggests that low runoff extremes can also be modeled with some degree of confidence. The regional high flows shown in the RFI catalogues suggest that more caution would be required in interpreting the outputs of high runoff extremes from large-scale models. While the results suggest that the global models often capture the overall characteristics of regional high flows in Europe, the model outputs fail to realistically capture key characteristics of regionally coherent high flow periods for most European regions.

This paper has demonstrated how two regional indices, the regional deficiency index (RDI) and the regional flood index (RFI), can be used to assess how well large-scale hydrological models can reproduce the spatiotemporal development of large-scale hydrological extremes. This preliminary study has demonstrated the utility of observed RDI and RFI time series (and corresponding drought and high flow catalogues) for providing a “baseline” against which to test large-scale models. It also shown how RDI and RFI can help to highlight regional differences in hydrological regimes of global models and to evaluate their ability to reproduce specific regimes, such as groundwater-dominated or alpine hydrology, and both extremes of the hydrological regime in a temperate climate such as Europe. However, it is important to keep in mind the limitations of the RDI/RFI approach, as identified by Hannaford et al. (2011), particularly in terms of regional homogeneity and the extent to which individual catchments are representative of large regions. Future work will employ a more up-to-date dataset and will also consider observed rainfall data.

This study has highlighted some key differences in how JULES, MPI-HM, and WaterGAP simulate the spatiotemporal development of hydrological extremes in Europe, which are likely to reflect differences in model formulation and process representation. While it was outside the scope of this paper to entirely explain the differences between models in terms of the different parameterization schemes used and to provide full physical interpretation of the results, this will be an important next step to help identify geographical areas or processes where some model improvement might be possible and to test any new parameterization. The methodology used here could also be used to analyze other model variables, such as surface runoff or evapotranspiration, although the lack of suitable observations for comparison would mean that such analyses would be aimed at better understanding intermodel differences in spatial and temporal variability. While the framework developed here has proven successful and suggests that there is scope for application in other regions of the world, results regarding model performance only apply to Europe. Model behavior and the ability to reproduce hydrological processes may be very different in different climate regimes. In this vein, Haddeland et al. (2011) show that large differences in the schemes used for estimating evapotranspiration can be important when explaining differences between models and between arid and wet regions. Furthermore, other indices might be better suited than RDI and RFI in terms of their ability to capture certain aspects of complex extremes or to describe the spatial evolution of extreme events as suggested by Sheffield et al. (2009) or Lloyd-Hughes (2011).

Acknowledgments. The authors thank Dr. George Goodsell for coding the RDI and RFI calculations and three anonymous reviewers for their suggestions. This research was undertaken as part of the European Union (FP6) funded Integrated Project called Water and Global Change (WATCH) (Contract 036946).

REFERENCES


