Validation of River Flows in HadGEM1 and HadCM3 with the TRIP River Flow Model

PETE FALLOON, RICHARD BETTS, ANDREW WILTSHIRE, RUTGER DANKERS, CAMILLA MATHISON, AND DOUG MCNEALL

Met Office Hadley Centre, Exeter, United Kingdom

PAUL BATES AND MARK TRIGG
School of Geographical Sciences, University of Bristol, Bristol, United Kingdom

(Manuscript received 12 October 2010, in final form 29 March 2011)

ABSTRACT

The Total Runoff Integrating Pathways (TRIP) global river-routing scheme in the third climate configuration of the Met Office Unified Model (HadCM3) and the newer Hadley Centre Global Environmental Model version 1 (HadGEM1) general circulation models (GCMs) have been validated against long-term average measured river discharge data from 40 stations on 24 major river basins from the Global Runoff Data Centre (GRDC). TRIP was driven by runoff produced directly by the two GCMs in order to assess both the skill of river flows produced within GCMs in general and to test this as a method for validating large-scale hydrology in GCMs. TRIP predictions of long-term-averaged annual discharge were improved at 28 out of 40 gauging stations on 24 of the world’s major rivers in HadGEM1 compared to HadCM3, particularly for low- and high-latitude basins, with predictions ranging from “good” (within 20% of observed values) to “poor” (biases exceeding 50%). For most regions, the modeled annual average river flows tended to be exaggerated in both models, largely reflecting inflated estimates of precipitation, although lack of human interventions in this modeling setup may have been an additional source of error. Within individual river basins, there were no clear trends in the accuracy of HadGEM1 versus HadCM3 predictions at up- or downstream gauging stations. Relative root-mean-square error (RRMSE) scores for the annual cycle of river flow ranged from poor (>50%) to “fair” (20%–50%) with an overall range of 20.7%–1023.5%, comparable to that found in similar global-scale studies. In both models, simulations of the annual cycle of river flow were generally better for high-latitude basins than in low or midlatitudes. There was a relatively small improvement in the annual cycle of river flow in HadGEM1 compared to HadCM3, mostly in the low-latitude rivers. The findings suggest that there is still substantial work to be done to enable GCMs to simulate monthly discharge consistently well over the majority of basins, including improvements to both (i) GCM simulation of basin-scale precipitation and evaporation and (ii) hydrological processes (e.g., representation of dry land hydrology, floodplain inundation, lakes, snowmelt, and human intervention).

1. Introduction

River flow is a useful indicator of freshwater availability, and can thus be used to evaluate likely impacts of climate change on water resources and flooding. There have been a number of studies of changes in river flow at the global scale (e.g., Arora and Boer 1999; Arnell 1999b, 2003; Hagemann and Dumenil 1998; Hirabayashi et al. 2008; Milly et al. 2005; Nijssen et al. 2001a,b) using either stand-alone hydrological models driven by climate data output from general circulation models (GCMs) or river-routing schemes driven by runoff produced by GCMs. The standard form of the previous version of the Hadley Centre GCM, the third climate configuration of the Met Office Unified Model (HadCM3; Gordon et al. 2000), used a very simple river-routing scheme in which surface and subsurface runoff were instantaneously advected from the land surface through outflow points to the ocean. A basinwide accounting scheme such as this is reasonably effective at annual time steps and even at
monthly time steps in basins with main stems shorter than several 100 km, where delays are negligible and temporal signals may be dominated by runoff generation (Fekete et al. 2002). However, this approach was unable to capture the time lags between runoff generation on the land surface and outflow to the ocean (Falloon et al. 2007), which is a significant process in long rivers and snowmelt-dominated catchments.

Embedding dynamic river-routing schemes in GCMs is advantageous since (i) river flow represents an integration of upstream catchment processes and therefore indicates the state of the surface water budget over large areas, and therefore (ii) river flow is useful for validating hydrology in large-scale models as it is relatively easy to measure on the ground, (iii) it enables the time lag between runoff generation and outflow to the ocean to be reproduced, and (iv) river flow is important for a realistic thermohaline circulation in a coupled atmosphere–ocean GCM (Arora and Boer 1999; Miller et al. 1994). The latest version of the Hadley Centre GCM, the Hadley Centre Global Environmental Model version 1 (HadGEM1; Martin et al. 2006), includes the dynamic river-routing scheme Total Runoff Integrating Pathways (TRIP; Oki and Sud 1998). TRIP comprises two components—a 1° gridded river-routing network and a simple advection scheme for river routing. We have recently investigated the impact of possible future climate change on river flows in HadGEM1 (Falloon and Betts 2006). Comparison of annual and monthly TRIP river flow outputs with observed river flow-gauge data has shown good agreement both using an independent runoff dataset (Oki 1997; Oki et al. 1999), and the land surface scheme used to produce runoff as input to TRIP in HadGEM1 has been shown to reproduce observed changes in continental-scale (but not basin scale) runoff during the twentieth century (Gedney et al. 2006), driven by climate, CO₂, aerosol, and land use forcings. In addition, a recent intercomparison of twentieth-century TRIP river flows driven by runoff from several different land surface models largely showed good skill (Morse et al. 2009) although there was a general overestimation of flows, possibly because of extractions (not accounted for in the models). Overestimation of flows was particularly noted for the Congo, Nile, and Niger, which are also studied here. Globally, irrigation water abstraction is 5%–6% of the annual discharge to oceans (Döll and Siebert 2002; Wisser et al. 2010) and this is concentrated mainly in semiarid and arid climate zones. The anthropogenic influence on flows in the Congo is, however, fairly small (Döll et al. 2009). For the Nile and Niger, water lost to dry land aquifers is an important process not commonly represented in global-scale hydrological models, while both irrigation and evaporation loss from Lake Nasser explain key water losses for the Nile.

However, TRIP has not been extensively validated when driven directly by HadGEM1 or HadCM3 data at the basin scale. The aim of this study is to validate TRIP-generated river flow values from major river basins using data from the Global Runoff Data Centre (GRDC) to (i) assess the skill of a global river-routing scheme driven by GCM outputs in general, (ii) assess whether flow-gauge data can be usefully employed to compare differences in large-scale hydrology between GCMs (HadGEM1 and HadCM3 in this case), and (iii) to identify areas for future model development. Since the standard version of HadCM3 did not include the TRIP river-routing scheme, a version of TRIP identical to that used in HadGEM1 was incorporated into HadCM3: although HadGEM1 was the first operational Met Office Hadley Centre GCM to include TRIP, development of the TRIP river-routing scheme was begun in the earlier GCM, HadCM3. This enabled a direct comparison of TRIP when driven by runoff outputs from the two GCMs. We have not compared predictions from HadCM3 with the original (simple) runoff scheme against HadCM3 predictions using TRIP (i.e., effectively comparing unrouted runoff from HadCM3 with TRIP-routed runoff). This has been described by Falloon et al. (2007)—the inclusion of TRIP in HadCM3 considerably improved river flow simulations, particularly for seasonality.

2. Methods and materials

2a. Climate models

The version of the Hadley Centre GCM used in the Intergovernmental Panel on Climate Change’s (IPCC) Fourth Assessment Report (Solomon et al. 2007) is HadGEM1, although simulations using the previous version HadCM3 were also included in Solomon et al. (2007). HadGEM1 and HadCM3 are described in detail by Johns et al. (2003, 2006) and Martin et al. (2006). HadGEM1 has many improvements over HadCM3, including improved horizontal latitude–longitude resolution (atmosphere: 1.25° × 1.875° versus 2.5° × 3.75°; ocean: 1.0° × 1.0° versus 1.25° × 1.25°) and substantially improved representations of physical processes including advection, boundary layer processes, gravity wave drag, microphysics, and sea ice schemes, plus major changes to convection, land surface, and cloud schemes, and inclusion of aerosols; a detailed comparison of the two models is given by Martin et al. (2006) and Johns et al. (2006). Compared to HadCM3, the developments to HadGEM1 have led to a substantial improvement in the processes represented, including the hydrological cycle, particularly for oceanic surface freshwater fluxes (Rodriguez et al. 2010). The hydrological cycle in HadGEM1 is stronger
than that in HadCM3, with global mean precipitation being ~0.15 mm day$^{-1}$ higher in HadGEM1. Compared to observations, HadGEM1 has too much annual precipitation over the Southern Ocean and high latitudes of the North Atlantic and North Pacific; over land, HadGEM1 is too wet over India and too dry over Southeast Asia, Indonesia, and the coast of western South America. These errors are broadly similar to HadCM3, although over land HadCM3 is too wet over Amazonia, too dry over central Africa, and the errors seen over India and western South America are less pronounced than in HadGEM1 (Johns et al. 2006). Both HadCM3 and HadGEM1 appear to have an overly strong hydrological cycle (Pardaens et al. 2003; Johns et al. 2003). In HadGEM1, runoff is produced by the Met Office Surface Exchange Scheme version 2 (MOSES2; Essery et al. 2001), while HadCM3 employed the original version of MOSES (MOSES1; Cox et al. 1999). Cox et al. (1998, 1999), Essery et al. (2001, 2003), and Smith et al. (2006) provide detailed descriptions of MOSES1 and MOSES2. Below, we describe key differences between the MOSES1 and MOSES2 land surface schemes (Smith et al. 2006), with particular attention to processes relevant to hydrology.

### b. Land surface schemes

MOSES1 was originally developed for a GCM to calculate the surface-to-atmosphere fluxes of heat and water and to update the surface and subsurface variables that affect these fluxes (Cox et al. 1999). There are four soil layers, each with a temperature and moisture content, and the four soil layers have thicknesses from the surface downward of 0.1, 0.25, 0.65, and 2.0 m. On the surface, there are lying snow and canopy water stores. The canopy water is the rainfall intercepted by plant leaves that is available for free evaporation. The method of partitioning precipitation into canopy interception and throughfall is described by Dolman and Gregory (1992). The total moisture flux from the surface is made up of evaporation from the canopy water store, transpiration by vegetation, bare soil evaporation, and sublimation from the lying snow surface. Canopy evaporation and sublimation are subject to aerodynamic resistance only, whereas transpiration and bare soil evaporation are subject to both the aerodynamic resistance and a surface resistance. For vegetated surfaces, the surface resistance is set to an interactive canopy or “bulk stomatal” resistance; the reciprocal conductance is calculated by a photosynthesis model (Cox et al. 1998) and depends on surface air temperature, humidity deficit, incident radiation, and vegetation type. Both the canopy and bare soil surface conductances also include dependencies on the soil moisture content via a soil moisture availability factor (Cox et al. 1999). The aerodynamic resistance is a function of surface roughness, wind speed, and atmospheric surface layer stability. MOSES assumes an energy balance at the land–atmosphere interface so that the net downward radiation at the surface is equal to the sum of the sensible and latent heat fluxes from the surface into the atmosphere, the conductive heat flux from the surface to the center of the top soil layer, and the latent heat required for any surface snowmelt. When snow is lying, melting occurs when the surface temperature or the snow layer temperature exceeds the freezing point of water. Sufficient snowmelt occurs to ensure that the surface energy balance is satisfied with the surface temperature equal to the freezing point. The soil thermodynamics in MOSES (Cox et al. 1999) are represented by diffusive heat exchanges between the soil layers and by the heat transported between the layers by the fluxes of moisture. The soil thermal characteristics are functions of soil moisture (liquid and frozen). Soil water phase changes are also represented and the associated latent heat is included in the thermodynamic calculations. The soil hydrology component of MOSES is based on a finite difference form of the Richard’s equation (Richards 1931). The vertical discretization for soil hydrology is the same as that for the thermodynamics (i.e., the positions of the moisture levels and layers coincide with those of the temperatures). The prognostic soil moisture content is the sum of frozen and unfrozen components, and the diffusive fluxes of soil moisture between the layers are given by the Darcy equation; hydraulic conductivity and soil water suction are given in terms of their saturation values by the formulae of Clapp and Hornberger (1978). The top boundary condition for the MOSES soil hydrology is the water flux, which enters the soil at the surface, determined as the sum of (throughfall and snowmelt) minus surface runoff. The lower boundary condition assumes free drainage (Cox et al. 1999). Transpiration through plants extracts soil moisture directly from each soil layer via roots and bare soil evaporation depletes moisture from the top soil layer. The ability of roots to access moisture in each soil layer is determined by a root density distribution; root density is assumed to follow an exponential distribution with depth (Thornley and Johnson 1990).

MOSES2 (Essery et al. 2003) employs a tiled model of subgrid heterogeneity in the MOSES land surface scheme, and includes seasonally varying vegetation (Martin et al. 2006). Whereas MOSES1 (Cox et al. 1999) used effective parameters to calculate a single surface energy balance for each grid box, MOSES2 treats subgrid land-cover heterogeneity explicitly. Separate surface temperatures, shortwave and longwave radiative fluxes, sensible and latent heat fluxes, ground heat fluxes, canopy moisture contents, snow masses, and snowmelt rates are computed for each surface type in a grid box. These are then
aggregated to form a grid-square mean with weightings equal to the fractions of each type in the grid square. Air temperature, humidity, and wind speed on atmospheric model levels above the surface and soil temperatures and moisture contents below the surface are treated as homogeneous across a grid box. Nine surface types are recognized in MOSES2 (as applied in HadGEM1): broadleaf trees, needleleaf trees, C3 (temperate) grass, C4 (tropical) grass, shrubs, urban, inland water, bare soil, and ice. Except for those classified as land ice, a land grid box can be made up from any mixture of the first eight surface types. Fractions of surface types within each land surface grid box are read from an ancillary file (as in the version of HadGEM1 used here) or modeled by the Top-down Representation of Interactive Foliage and Flora Including Dynamics (TRIFFID) dynamic vegetation model (Cox 2001). With each type of vegetation is associated a canopy height, a snow-free roughness length, and a canopy water capacity, which are parameterized functions of the leaf area index (LAI), rather than read from ancillaries (as in MOSES1; Essery et al. 2001). Each vegetated tile also has an exponential root density distribution depending on the plant type. Evaporation from the bare soil tile is drawn from the top soil layer only using a reformulated conductance equation, which also applies to soil beneath sparse vegetation. Lakes and urban surfaces are represented very simply: lake tiles are smooth and perpetually wet, and urban tiles are rough and have a small surface water capacity from which evaporation can occur. There is no water extraction from the soil layers for urban tiles. Neither MOSES1 nor MOSES 2 simulates groundwater explicitly. Other features in MOSES 2.2 (not included in MOSES1) include (Essery et al. 2001) new vegetation maps; an optional spectral albedo scheme that calculates separate diffuse and direct beam albedos in visible and near-infrared bands for vegetation tiles, with snow aging parameterized using a prognostic grain size; extension of the Penman–Monteith elimination of the surface temperature from the surface energy balance to include upward longwave radiation; reformulated canopy heat capacity and fractional coverage calculations in the optional canopy model; a new implicit numerical scheme for updating temperatures and moisture contents of soil layers; and inclusion of increments due to snowmelt or limited moisture availability within the implicit calculation of surface heat and moisture fluxes.

A direct comparison of MOSES1 and MOSES2 has not been performed, but the performance of MOSES 2 is discussed in climate simulations by Essery et al. (2003) and in mesoscale forecasts by Best et al. (2000). The introduction of surface tiling within MOSES2 (i.e., comparing aggregated and tiled versions of MOSES2) was not found to give a clear improvement in the simulated climate but offered more flexibility in the representation of heterogeneous land surface processes. Using the same meteorological forcing in offline simulations using versions of the MOSES2 surface model with and without tiling, the tiled model gave slightly lower winter temperatures at high latitudes and higher summer temperatures at midlatitudes (Essery et al. 2003). When coupled to a GCM, reduced evaporation in the tiled version of MOSES2 led to changes in cloud cover and radiation at the surface that enhanced these differences.

c. TRIP river-routing model and input data

HadGEM1 includes the TRIP river-routing model (Oki and Sud 1998). TRIP uses a simple advection method to route total (surface and subsurface) runoff along prescribed river channels, and does not consider evaporation from within the river channel. The river channels are represented by two datasets that give the direction and sequence of the flow of water at 1° resolution. The model requires a universal meander coefficient and an effective velocity to be set globally. The value of the meander coefficient used here was 1.4 (dimensionless), set according to Oki and Sud (1998), representing the ratio of actual river length to modeled river length in the TRIP routing network. Gridded river-routing networks need to reproduce both the maximum river length (achieved using the meander coefficient) and the length distribution within the basin that can be described by the width function (Rinaldo et al. 1995; Veneziano et al. 2000). To reproduce the width function realistically, several hundred grid cells may be necessary (Fekete et al. 2001), implying that 0.5° resolution may be more appropriate for monthly flow routing. Although a 0.5° resolution routing dataset exists for TRIP (and several alternatives are available—e.g., Döll and Lehner 2002; Vorösmarty et al. 2000), the 1° version was included in HadCM3 and HadGEM1 and we note that increased routing resolution would be a beneficial future development, particularly because GCMs currently in development have horizontal atmospheric resolution approaching 0.5° (e.g., Walters et al. 2011). In our simulations, effective velocity was set to the globally optimal value of 0.5 m s⁻¹ determined by Oki et al. (1999). Chapelon et al. (2002) assessed the sensitivity of TRIP river flow simulations to effective velocity and meander coefficient values for the Amazon basin. Within the range of values employed, they found predictions to be slightly more sensitive to effective velocity (0.25–1.0 m s⁻¹) than to the meander coefficient (1.0–1.8). However, tuning TRIP parameters did not result in significant improvements in the annual cycle of simulated discharge. Therefore, we have not included parameter sensitivity experiments in this study. TRIP runs on a daily time step, and is coupled to HadGEM1 by interpolating surface and subsurface...
runoff values (produced by the MOSES2 land surface scheme) from the GCM grid to the 1° TRIP grid. Runoff at coastal outflow points is then regridded to the atmosphere grid and added to the ocean as a freshwater flux (for a more detailed description, see Martin et al. 2006). Outputs from TRIP include total inflow, total outflow, and water storage for each grid box. Most of the river basin sizes represented on the 1° TRIP grid are within 20% of published values, with a root-mean-square error of approximately 10% (Oki and Sud 1998).

Since the standard version of HadCM3 does not include TRIP, a version of HadCM3 was developed incorporating the TRIP river-routing scheme to allow rigorous comparison of river flow predictions between the two GCMs. Monthly river outflow values were then taken from a long-term baseline run of HadGEM1 and the version of HadCM3 including TRIP and averaged over 10-year periods. The baseline simulations were “control” experiments, with forcings representing preindustrial climates. For instance, the HadGEM1 control experiment applied forcings representative of the year 1860 for greenhouse gases $[\text{GHGs} - \text{CO}_2 = 286.2 \text{ ppmv, } \text{CH}_4 = 805.5 \text{ ppbv, } \text{N}_2\text{O} = 286.2 \text{ ppbv, and chlorofluorocarbons (CFCs)} = 0]$, aerosol precursor emissions ($\text{SO}_2$, dimethyl sulfate, sea salt, and soot), ozone concentrations, and the land surface. Full details are given by Johns et al. (2003, 2006) and Stott et al. (2006).

d. Flow-gauge data and assessment metrics

We selected 40 gauging stations on 24 major river basins with at least 20 years of monthly records from the Global Runoff Data Centre (GRDC 2010)—see Fig. 1. The basins studied all have areas exceeding 300 000 km$^2$ (Miller et al. 1994; Oki and Sud 1998; Hagemann and Dumenil 1998). The scale of the basins is such that at all stations there is a single annual (i.e., monomodal) flood pulse. The gauging stations were chosen to cover rivers at high, mid-, and low latitudes, and at each latitude band at least two rivers with several up- and downstream stations were selected. This allowed us to assess where improvements in downstream river flow predictions had been achieved at the expense of upstream predictions and vice versa, and also the potential to capture differences in behavior between subcatchments (e.g., Chapelon et al. 2002). Hence, multiple gauges within an individual river basin represent subsampling rather than replication. Long-term-averaged (10 year) values of monthly and annual TRIP-simulated river flows from the HadCM3 and HadGEM1 runs at the grid box corresponding to the location of the gauging station used for observations were then compared with the long-term-averaged (>20 years, with variable averaging periods) observations from the GRDC. We have not corrected river flow estimates for the area upstream of the gauging stations (e.g., Haddeland et al. 2011).

Note that the river flow values from the HadCM3 and HadGEM1 climate models represent preindustrial climates, while the GRDC flow-gauge data represent averages of different periods (largely during the twentieth century), which may lead to some discrepancies between the two datasets. Key components of potential discrepancies may include (i) differences in climate (although our control simulations will include some elements of natural variability) and (ii) changes in the human modification of river basins (e.g., land use change, reservoirs, floodplain loss, and channel alignments—which are not included in the version of TRIP used here; Hagemann & Dumenil 1998; Oki et al. 1999). It is possible that discrepancies arising from the different averaging periods used here may be small—in their analysis of continental
river discharge records from 1948 to 2005, Dai et al. (2009) found only small trends in most basins. Only about one-third of the 200 largest rivers showed statistically significant trends over this period (Dai et al. 2009) with the rivers having downward trends (45) outnumbering those with upward trends (19). However, the aim of our study was to develop ways of assessing river flows in GCMs, a key application of which is during the development cycle of the GCM itself, which usually only includes “unforced” control simulations representing preindustrial climates, rather than dynamic twentieth-century conditions. In this context, it is therefore necessary to evaluate GCM-simulated and observed river flows with caution given the caveats above.

Recent comparisons of the outputs of global-scale river flow and runoff schemes with flow-gauge data have employed various statistics to evaluate the different aspects of simulation skill. For instance, comparisons of the annual cycle of discharge have used either the root-mean-square error (RMSE; Smith et al. 1996) in terms of volume (Arora and Boer 1999; Oki et al. 1999) or as a relative percentage [relative root-mean-square error (RRMSE); Arora and Boer 1999] or the Nash–Sutcliffe modeling efficiency (Nash and Sutcliffe 1971) and similar equations such as the normalized flow error (Miller et al. 1994). Annual average discharge has been assessed using the difference between simulated and observed annual discharge, either as a volume (Oki et al. 1999; Chapelon et al. 2002) or as a (relative) percentage (Arnell 2003; Nijssen et al. 2001; Chapelon et al. 2002). Arora and Boer (1999) further separated the fraction of the total error associated with the annual mean and the amplitude and phase of the annual cycle of river flow. Alcamo et al. (2003a) applied the Nash–Sutcliffe modeling efficiency statistic to time series of annual average discharge values. To provide a “global” assessment across all gauges, the regression coefficient (Oki et al. 1999) and average mean (and standard deviation of) quadratic error over all stations (Chapelon et al. 2002; Miller et al. 1994) have been employed. Willmott and Matsuura (2005) discuss limitations in the use of RMSE as a model performance statistic, including that it is a function of three characteristics of a set of errors (rather than of one—the average error), and that RMSE varies with variability within the distribution of error magnitudes, the square root of the number of errors, and with average error magnitude [or mean absolute error (MAE)]. Willmott and Matsuura (2005) therefore propose the MAE as a more natural, unambiguous measure of average error. Expressing MAE as a relative percentage of observed values [mean absolute percent error (MAPE); e.g., Smith and Pavelsky 2008; Wang et al. 2009] further aids comparison across basins since it avoids potential issues with larger (apparent) errors arising from greater absolute flow values. We have used several of these means of assessing model performance relative to the observed flow-gauge data—absolute and relative (%) bias for annual average runoff; RRMSE, (squared) correlation coefficient ($r^2$), MAE, and MAPE for the annual cycle (using monthly average discharge values); and global square of the correlation coefficient ($r^2$) of annual average discharge values across all gauges.

To determine overall differences between the performance of one model and another, it is necessary to define poor or good performance in terms of the assessment statistics used. This may be done either arbitrarily, with reference to estimates of errors in the observed data, or simply by comparing error statistics with those found in other studies. For example, Nash–Sutcliffe values greater than 0.5 were considered “good” in the calibrated river flow model of Alcamo et al. (2003a), though thresholds of 0.6–0.8 are also commonly applied (McCuen et al. 2006). However, as McCuen et al. (2006) point out, the Nash–Sutcliffe efficiency coefficient is sensitive to a number of factors, including sample size and outliers, and higher Nash–Sutcliffe values may not accurately reflect better model performance unless the sampling distribution is known. RRMSE values found in previous uncalibrated global-scale river flow assessments have ranged from 19% to 1886% (Nohara et al. 2006), 32% to 712% (Nijssen et al. 2001b), and 15% to 250% (Arora and Boer 1999). Nijssen et al. (2001a,b) reported annual average runoff biases from an uncalibrated global river flow model ranging from 5% to 424%, although values of 10%–40.6% were found for nine selected basins. Errors in annual average runoff of 50% are not uncommon (Arnell 1999a,b; Meigh et al. 1999; Döll et al. 2003). Arnell (2003) compared annual average runoff biases in his unrouted global runoff simulations driven by observed climatology (between +53% and −54%) with values for the same basins from Nijssen et al. (2001a,b), which were +83% to −69% (uncalibrated) and +21% to −15.9% (calibrated). Arnell (2003) found a correlation between simulated and observed annual average runoff values across all gauges of 0.91 with a median bias of 10%, using a hydrological model driven with observed climatology. Döll et al. (2003) assessed the percentage of basins where the difference between observed and simulated long-term discharge values was either <10% or 10%–20%, although this was applied to $Q_{90}$ values (the discharge exceeded 90% of the time).

In our study, we have applied the following thresholds to qualitatively assess model performance: bias in annual average discharge, RRMSE, and MAPE (over an annual cycle of monthly values) >50% (“poor”), 20%–50% (“fair”), and <20% (“good”). We choose <20% as
our criterion for good performance as this is a reasonable upper limit for typical errors in observed gauge data (Oki et al. 1999 broadly suggest that observational errors of ~5% at the 95% confidence level might be expected; see McMillan et al. 2010 for a discussion). Hence, modeled annual or monthly discharge values within 20% of the gauged observations are what we would wish a GCM model to be able to achieve consistently, and values within 10% can be viewed as statistically indistinguishable from the observed data. However, classifying metrics into small categories such as these will likely mask the impacts of any changes in model performance within these classes. It is also worth noting that hydrologists and meteorologists might have very different views of what constitutes a useful simulation of river discharge depending on what they then wanted to do with the data. For example, to predict the feedbacks between river discharge and the thermohaline circulation might simply require that the annual runoff volume was broadly correct (e.g., Pardaens et al. 2003), whereas for most hydrologists simulating the correct timing of the monomodal flood pulse in these basins would be of at least equal importance. Correct annual runoff volumes but with significant phase errors in the monthly flow would likely indicate that the GCM was getting the right results for the wrong reasons.

3. Results and discussion

Tables 1 and 2 and Figs. 2 and 3 show the results of the validation for 40 gauging stations from 24 of the world’s major rivers. In 28 out of the 40 gauging stations on the 24 rivers investigated here, prediction of annual average river flow was improved (as assessed by percent difference from the GRDC observations) in HadGEM1 compared to HadCM3, while skill was reduced in HadGEM1 for the remaining 12 stations. The correlation coefficients ($r^2$) between simulated and observed annual average discharge values across all 40 gauges were 0.67 and 0.88 for HadCM3 and HadGEM1, respectively. When assessed only over the downstream gauges for the 24 rivers (to avoid resampling/replication errors), the $r^2$ values were 0.69 and 0.97 for HadCM3 and HadGEM1, respectively. For most regions, the modeled river flows in both HadCM3 and HadGEM1 were exaggerated (Fig. 3), either reflecting inflated estimates of precipitation (in common with other atmosphere–ocean GCMs; Nohara et al. 2006), underestimates of evapotranspiration, or the lack of abstraction in the version of TRIP used here. Of the gauging stations studied, modeled river flows were positively biased at 9 (12) out of 16 low-latitude stations, 11 (9) out of 12 midlatitude stations, and 7 (8) out of 12 high-latitude stations in HadCM3 and HadGEM1, respectively. On the other hand, under-prediction of annual river flows occurred at 7 (4) out of 16 low-latitude stations, 1 (3) out of 12 midlatitude stations, and 5 (4) out of 12 high-latitude stations in HadCM3 and HadGEM1, respectively. Using the arbitrary scoring classification described earlier suggests an overall improvement in predicted annual average flow volumes in HadGEM1 compared to HadCM3 and a notable increase in good predictions in HadGEM1, with much of this improvement occurring in the high- and low-latitude basins (Table 2). Overall, predictions of annual average flow volumes in both models ranged from poor to good. The reason for the relatively small improvement in flow predictions at low latitudes in HadGEM1 compared to HadCM3 may be the fairly small improvements in precipitation errors for many basins (see later discussion).

We also assessed model fit to the GRDC-observed annual cycle of monthly river flow values using the $r^2$ correlation coefficient, the RRMSE, and MAPE. Using the scoring classification adopted in this study for the RRMSE values suggests that predictions of the annual cycle of river flow in both models ranged from poor to fair. MAPE values for both models were generally close to RRMSE values for both models in many basins, although MAPE values were notably smaller for the upstream station on the Amazon in both models and in HadGEM1 for the Mekong (Luang Prabang), Orinoco (Tama–Tama), and Niger (Niamey and Gaya). MAPE values were notably greater than RRMSE in both HadGEM1 and HadCM3 for the Nile (Aswan Dam), in HadGEM1 for the Indus, and in HadCM3 for the Niger (Niamey), Chari, and Orange. In both models, predictions for high-latitude rivers were generally better than for low- or midlatitude rivers, as indicated by the larger $r^2$ and smaller RRMSE values for high-latitude rivers. This is likely because of the strong seasonality of the hydrological system in cold climate regions that is dominated by snowmelt in spring, which apparently is reproduced well by the two GCMs—Döll et al. (2003) and Spérna Weiland et al. (2010) also noted good simulation of seasonality for Siberian rivers. Using the scoring system for RRMSE values (Table 2) shows only a relatively small improvement in the predicted annual cycle of river flow between HadCM3 and HadGEM1, with the largest improvement occurring in the low-latitude basins. Applying the same scoring system to MAPE values (Table 2) shows very similar overall patterns to the scoring system based on RRMSE values, again indicating a small general improvement in fair simulations, arising mostly from the low-latitude basins.

Figure 3 and Fig. 2a show that a considerable improvement in flow prediction for the Amazon mouth was obtained with HadGEM1 compared to HadCM3. The annual average flow predicted by HadCM3 was approximately half the observed volume, while that in HadGEM1
Table 1. Comparison of GRDC-observed, HadCM3, and HadGEM1 long-term mean river discharge from major river gauging stations. RRMSE, MAE, MAPE, and $r^2$ values are derived from the annual cycle of river flow. Higher superscript values within a particular river basin indicate increasing distance downstream.

<table>
<thead>
<tr>
<th>River/station/latitude/longitude</th>
<th>Annual average outflow (m$^3$ s$^{-1}$ × 1000)</th>
<th>Difference in annual average outflow GRDC simulated (m$^3$ s$^{-1}$ × 1000)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GRDC HadCM3 HadGEM1</td>
<td>HadCM3 HadGEM1</td>
</tr>
<tr>
<td>Low-latitude Rivers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1) Amazon—Obidos, Brazil (1.90°S, 55.50°W)$^2$</td>
<td>176.2 87.2 196.3</td>
<td>−89.0 20.1</td>
</tr>
<tr>
<td>2) Amazon—Sao Paulo de Olievenca, Brazil (3.47°S, 68.75°W)$^1$</td>
<td>7.0 25.1 69.2</td>
<td>18.0 62.2</td>
</tr>
<tr>
<td>3) Zaire—Kinshasa, Zaire (4.30°S, 15.30°E)</td>
<td>39.5 78.6 24.0</td>
<td>39.1 −15.6</td>
</tr>
<tr>
<td>4) Nile—El Ekhsase, Egypt (29.70°N, 31.28°E)$^2$</td>
<td>1.3 13.8 12.6</td>
<td>12.6 11.3</td>
</tr>
<tr>
<td>5) Nile—Aswan Dam, Egypt (23.96°N, 32.90°E)$^1$</td>
<td>2.8 13.8 12.5</td>
<td>11.1 9.7</td>
</tr>
<tr>
<td>6) Irrawaddy—Sagaing, Myanmar (21.98°N, 96.10°E)</td>
<td>8.0 6.1 10.9</td>
<td>−1.9 2.9</td>
</tr>
<tr>
<td>7) Brahmaputra—Bhadurabad, Bangladesh (25.18°N, 89.67°E)$^2$</td>
<td>21.3 11.9 23.4</td>
<td>−9.4 2.1</td>
</tr>
<tr>
<td>8) Brahmaputra—Pantu, Bangladesh (26.13°N, 91.70°E)$^1$</td>
<td>18.1 9.6 15.6</td>
<td>−8.5 −2.6</td>
</tr>
<tr>
<td>9) Mekong—Jakse, Vietnam (15.12°N, 105.80°E)$^2$</td>
<td>9.0 6.3 12.0</td>
<td>−2.8 3.0</td>
</tr>
<tr>
<td>10) Mekong—Luang Prabang, Vietnam (19.88°N, 102.13°E)$^1$</td>
<td>3.6 4.5 9.3</td>
<td>0.9 5.6</td>
</tr>
<tr>
<td>11) Orinoco—Puente Angosta, Venezuela (8.15°N, 63.60°W)$^2$</td>
<td>30.6 13.8 28.2</td>
<td>−16.8 −2.5</td>
</tr>
<tr>
<td>12) Orinoco—Tama-Tama, Venezuela (3.12°N, 66.07°W)$^1$</td>
<td>1.2 1.4 3.7</td>
<td>0.2 2.5</td>
</tr>
<tr>
<td>13) Niger—Niamey, Nigeria (13.52°N, 2.08°E)$^2$</td>
<td>0.9 2.6 1.8</td>
<td>1.7 0.9</td>
</tr>
<tr>
<td>14) Niger—Koulikoro, Nigeria (12.87°N, 7.55°W)$^1$</td>
<td>1.4 1.0 1.5</td>
<td>−0.4 0.1</td>
</tr>
<tr>
<td>15) Niger—Gaya, Nigeria (11.88°N, 3.4°E)$^3$</td>
<td>1.0 2.8 1.9</td>
<td>1.8 0.9</td>
</tr>
<tr>
<td>16) Chari—Ndjamena, Chad (11.50°N, 15.03°E)</td>
<td>1.1 4.2 1.1</td>
<td>3.1 0.0</td>
</tr>
<tr>
<td>Midlatitude rivers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17) Mississippi—Vicksburg, United States (32.22°N, 90.90°W)$^1$</td>
<td>17.6 25.4 16.1</td>
<td>7.8 −1.5</td>
</tr>
<tr>
<td>18) Mississippi—Alton, United States (38.88°N, 90.18°W)$^2$</td>
<td>2.9 13.5 6.0</td>
<td>10.6 3.1</td>
</tr>
<tr>
<td>19) Mississippi—St. Paul, United States (44.94°N, 93.09°W)$^1$</td>
<td>0.3 0.4 0.2</td>
<td>0.1 −0.1</td>
</tr>
<tr>
<td>20) St. Lawrence—Cornwall, Canada (44.01°N, 84.78°W)</td>
<td>7.9 9.9 12.3</td>
<td>2.0 4.3</td>
</tr>
<tr>
<td>21) Huanghe (Yellow)—Huayuankou, China (34.92°N, 113.65°E)$^2$</td>
<td>1.4 4.5 3.5</td>
<td>3.1 2.1</td>
</tr>
<tr>
<td>22) Huanghe (Yellow)—Shanxi, China (34.82°N, 111.15°E)$^1$</td>
<td>1.4 4.4 3.4</td>
<td>3.0 2.0</td>
</tr>
<tr>
<td>23) Chang Jiang (Yangtze)—Datong, China (30.77°N, 117.62°E)</td>
<td>28.8 41.2 50.0</td>
<td>12.4 21.2</td>
</tr>
<tr>
<td>24) Amur—Komsomolsk, Russia (50.63°N, 137.12°E)$^2$</td>
<td>9.9 12.1 10.8</td>
<td>2.2 0.9</td>
</tr>
<tr>
<td>25) Amur—Khabarovsk, Russia (48.43°N, 135.05°E)$^1$</td>
<td>8.5 11.5 10.2</td>
<td>3.0 1.7</td>
</tr>
<tr>
<td>26) Orange—Aliwal North, South Africa (30.68°S, 26.71°E)</td>
<td>0.2 0.3 0.9</td>
<td>0.2 0.8</td>
</tr>
<tr>
<td>27) Volga—Volgograd, Russia (48.77°N, 44.72°E)</td>
<td>8.3 5.6 7.5</td>
<td>−2.7 −0.8</td>
</tr>
<tr>
<td>28) Indus—Kotri, Pakistan (25.37°N, 68.37°E)</td>
<td>2.5 2.6 6.0</td>
<td>0.1 3.5</td>
</tr>
<tr>
<td>High-latitude rivers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29) Kolyma—Kolymskaya, Russia (68.73°N, 158.72°E)</td>
<td>3.1 3.3 3.2</td>
<td>0.2 0.1</td>
</tr>
<tr>
<td>30) Lena—Kusur, Russia (70.70°N, 127.65°E)$^1$</td>
<td>16.6 12.1 14.1</td>
<td>−4.5 −2.6</td>
</tr>
<tr>
<td>31) Lena—Stolb, Russia (72.37°N, 126.80°E)$^2$</td>
<td>15.2 12.2 14.2</td>
<td>−3.0 −1.0</td>
</tr>
<tr>
<td>32) Ob—Salekhard, Russia (66.57°N, 66.53°E)</td>
<td>12.5 9.9 13.2</td>
<td>−2.7 0.6</td>
</tr>
<tr>
<td>33) Yenisey—Igarka, Russia (67.48°N, 86.50°E)</td>
<td>18.1 16.2 15.9</td>
<td>−1.8 −2.2</td>
</tr>
<tr>
<td>34) Mackenzie—Arctic Red River, Canada (67.46°N, 133.74°W)$^1$</td>
<td>9.1 11.4 10.4</td>
<td>2.3 1.3</td>
</tr>
<tr>
<td>35) Mackenzie—Norman Wells, Canada (65.28°N, 126.85°W)$^2$</td>
<td>8.3 10.7 9.6</td>
<td>2.3 1.3</td>
</tr>
<tr>
<td>36) Mackenzie—near Fort Providence, Canada (112.27°N, 117.53°W)$^1$</td>
<td>4.4 6.1 5.0</td>
<td>1.7 0.7</td>
</tr>
<tr>
<td>37) Yukon—Pilot Station, United States (61.93°N, 162.88°W)$^3$</td>
<td>6.4 8.3 7.2</td>
<td>2.0 0.8</td>
</tr>
<tr>
<td>38) Yukon—Ruby, United States (64.73°N, 155.48°W)$^2$</td>
<td>4.2 7.1 5.8</td>
<td>2.8 1.6</td>
</tr>
<tr>
<td>39) Yukon—Eagle, United States (64.79°N, 141.20°W)$^1$</td>
<td>2.4 3.9 2.8</td>
<td>1.5 0.5</td>
</tr>
<tr>
<td>40) Severnay Drina—Ust-Pinega, Russia (64.1°N, 42.17°E)</td>
<td>3.4 3.1 2.4</td>
<td>−0.2 −1.0</td>
</tr>
</tbody>
</table>

Correlation ($r^2$) between observed and simulated annual average discharge across all basins ($n = 40$) | 0.67 0.88 |

Correlation ($r^2$) between observed and simulated annual average discharge across downstream basins only ($n = 24$) | 0.69 0.97 |

overpredicted the observed runoff by −20%. This is significant since the Amazon contributes the largest flux of freshwater to the oceans globally—approximately three times that of any of the other rivers investigated here. Predicted seasonality of river flow (as assessed by the RRMSE) was also generally improved in HadGEM1 compared to HadCM3 for the Zaire, Brahmaputra, downstream Orinoco (Tama–Tama), Niger, and Chari.
(low latitudes); the Mississippi, Huanghe, Amur, and Volga (midlatitudes); and the Lena, Yukon, and Yenisey (high latitudes). However, the improvements for the Amur and Yenisey were marginal. Predicted seasonality (annual cycle of monthly flow values) was generally worse in HadGEM1 in comparison with HadCM3 for the Irrawady, Mekong, upstream Brahmaputra (Pandu), Orinoco, and Amazon (low latitudes); the Indus, Orange,

<table>
<thead>
<tr>
<th>% difference in annual average outflow GRDC simulated</th>
<th>$r^2$</th>
<th>RRMSE (%)</th>
<th>MAE (m$^3$ s$^{-1}$ × 1000); MAPE (%) in parentheses</th>
</tr>
</thead>
<tbody>
<tr>
<td>HadCM3</td>
<td>HadGEM1</td>
<td>HadCM3</td>
<td>HadGEM1</td>
</tr>
<tr>
<td>50.5</td>
<td>11.4</td>
<td>0.78</td>
<td>0.87</td>
</tr>
<tr>
<td>256.5</td>
<td>883.9</td>
<td>0.82</td>
<td>0.64</td>
</tr>
<tr>
<td>98.8</td>
<td>-39.4</td>
<td>0.02</td>
<td>0.2</td>
</tr>
<tr>
<td>1004.4</td>
<td>903.8</td>
<td>0.54</td>
<td>0.18</td>
</tr>
<tr>
<td>400.5</td>
<td>352.8</td>
<td>0.01</td>
<td>0.17</td>
</tr>
<tr>
<td>-23.8</td>
<td>36.3</td>
<td>0.89</td>
<td>0.86</td>
</tr>
<tr>
<td>-44.1</td>
<td>10.0</td>
<td>0.59</td>
<td>0.92</td>
</tr>
<tr>
<td>-46.8</td>
<td>-14.1</td>
<td>0.59</td>
<td>0.84</td>
</tr>
<tr>
<td>-30.6</td>
<td>32.7</td>
<td>0.3</td>
<td>0.45</td>
</tr>
<tr>
<td>23.4</td>
<td>155.0</td>
<td>0.75</td>
<td>0.95</td>
</tr>
<tr>
<td>-54.9</td>
<td>-8.0</td>
<td>0.93</td>
<td>0.96</td>
</tr>
<tr>
<td>12.8</td>
<td>201.8</td>
<td>0.85</td>
<td>0.72</td>
</tr>
<tr>
<td>190.7</td>
<td>104.6</td>
<td>0.64</td>
<td>0.5</td>
</tr>
<tr>
<td>-29.2</td>
<td>4.5</td>
<td>0.72</td>
<td>0.39</td>
</tr>
<tr>
<td>170.6</td>
<td>82.6</td>
<td>0.55</td>
<td>0.38</td>
</tr>
<tr>
<td>292.7</td>
<td>-0.7</td>
<td>0.92</td>
<td>0.84</td>
</tr>
<tr>
<td>44.6</td>
<td>-8.4</td>
<td>0.1</td>
<td>0.45</td>
</tr>
<tr>
<td>365.5</td>
<td>106.6</td>
<td>0</td>
<td>0.75</td>
</tr>
<tr>
<td>34.9</td>
<td>-46.0</td>
<td>0.22</td>
<td>0.06</td>
</tr>
<tr>
<td>24.7</td>
<td>54.7</td>
<td>0.11</td>
<td>0.54</td>
</tr>
<tr>
<td>213.7</td>
<td>144.8</td>
<td>0.01</td>
<td>0.38</td>
</tr>
<tr>
<td>224.3</td>
<td>149.6</td>
<td>0.02</td>
<td>0.39</td>
</tr>
<tr>
<td>43.0</td>
<td>73.5</td>
<td>0.23</td>
<td>0.47</td>
</tr>
<tr>
<td>22.1</td>
<td>9.4</td>
<td>0.56</td>
<td>0.68</td>
</tr>
<tr>
<td>35.7</td>
<td>20.5</td>
<td>0.58</td>
<td>0.66</td>
</tr>
<tr>
<td>127.6</td>
<td>548.5</td>
<td>0.63</td>
<td>0.66</td>
</tr>
<tr>
<td>-32.9</td>
<td>-9.4</td>
<td>0.36</td>
<td>0.68</td>
</tr>
<tr>
<td>4.1</td>
<td>139.0</td>
<td>0.68</td>
<td>0.21</td>
</tr>
<tr>
<td>6.6</td>
<td>1.8</td>
<td>0.94</td>
<td>0.91</td>
</tr>
<tr>
<td>-27.1</td>
<td>-15.4</td>
<td>0.09</td>
<td>0.41</td>
</tr>
<tr>
<td>-19.9</td>
<td>-6.7</td>
<td>0.07</td>
<td>0.39</td>
</tr>
<tr>
<td>-21.3</td>
<td>5.1</td>
<td>0.98</td>
<td>0.93</td>
</tr>
<tr>
<td>-10.1</td>
<td>-12.1</td>
<td>0.53</td>
<td>0.73</td>
</tr>
<tr>
<td>24.6</td>
<td>13.7</td>
<td>0.86</td>
<td>0.88</td>
</tr>
<tr>
<td>28.2</td>
<td>15.5</td>
<td>0.91</td>
<td>0.89</td>
</tr>
<tr>
<td>38.7</td>
<td>15.0</td>
<td>0.37</td>
<td>0.13</td>
</tr>
<tr>
<td>31.3</td>
<td>12.9</td>
<td>0.76</td>
<td>0.68</td>
</tr>
<tr>
<td>67.4</td>
<td>37.3</td>
<td>0.94</td>
<td>0.85</td>
</tr>
<tr>
<td>64.2</td>
<td>20.1</td>
<td>0.87</td>
<td>0.93</td>
</tr>
<tr>
<td>-6.1</td>
<td>-29.6</td>
<td>0.89</td>
<td>0.48</td>
</tr>
</tbody>
</table>
and St. Lawrence (midlatitudes); and the Mackenzie, Kolyma, Ob, and Severnay Dvina (high latitudes).

We considered that gauging stations where the difference between GRDC observations and HadGEM1 predictions of annual average flow exceeded 50% warranted further attention since these basins should be focal areas for model improvements. At low latitudes the difference between HadGEM1 predictions and GRDC observations of annual average flow was >50% at seven gauging stations. At the upstream station on the Amazon, Sao Paolo di Olivenca, HadGEM1 predictions were worse compared to HadCM3—the amount of runoff was overpredicted and the peaks in seasonality were too high. This was also the case for the upstream stations on the Mekong (Luang Prabang) and Orinoco (Tama–Tama) and the two lower-reach stations on the Niger (Niamey and Gaya). At both El Ekshasa and the Aswan Dam on the Nile, annual average river flow predictions were overpredicted and the seasonality of flow was very different to that observed. As noted by other authors, abstraction and other unrepresented processes may have a significant impact on discharge simulation for arid basins such as the Nile (see later discussion and, e.g., Oki et al. 1999; Nijssen et al. 2001b; Arnell 2003; Alcamo et al. 2003a; Döll et al. 2003; Nohara et al. 2006; Haddeland et al. 2011). At midlatitudes, the difference between HadGEM1 predictions and GRDC observations was >50% at seven gauging stations. At most of these stations, the annual average river flow predictions were overpredicted and the peaks in seasonality were too high (Alton–Mississippi, Cornwall–St. Lawrence, Huayankou, and Shannxian–Huanghe). A recent application of TRIP driven by runoff from an ensemble of GCMs (including HadCM3 and HadGEM1) also found that simulated peak discharge timing for rainy midlatitude basins shifted one to two months earlier compared to observations (Nohara et al. 2006). In our simulations the (Northern Hemisphere) winter peak in river flow was too high at Alivil North (Orange), while at Datong (Chang Jiang) and Kotri (Indus) HadGEM1 predicted two peaks in river flow compared to the one peak found in the GRDC observations. Amongst the high-latitude flow-gauge datasets considered here, the difference between HadGEM1 predictions and GRDC observations of annual average flow did not exceed 50% at any of the gauging stations. However, the errors in average annual discharge were relatively large at Ruby (Yukon), where predicted runoff was overpredicted between July and December, and at Ust-Pinega (Severnay Dvina), where the HadGEM1-predicted annual average river flow was underpredicted and the river flow peak too early and too small. Other high-latitude stations with noticeable errors were Fort Providence (Mackenzie), where annual average river flow was overpredicted and peaks in seasonality were too pronounced, and both stations on the Lena, where the simulated annual average river flow was underpredicted and seasonal peaks were not high enough. Note that Fort Providence is at the outlet of a large lake, which may explain the relatively smoothed-out seasonal pattern in the observed river discharge. Nohara et al. (2006) and Sperna Weiland et al. (2010) also found peak discharge magnitudes for high-latitude basins were not well represented by TRIP when driven by GCM data; this may suggest that GCM-simulated temperatures are too high for too many days during the cold half of the year (Sperna Weiland et al. 2010).

Within individual river basins, there were no clear trends in the accuracy of HadGEM1 versus HadCM3 predictions of annual average flow at up- or downstream gauging stations. At low latitudes, downstream HadGEM1 predictions for the Amazon and Orinoco were improved while upstream predictions were degraded and both up- and downstream predictions were improved for the
Brahmaputra and Niger, while predictions for the Mekong were worse both up- and downstream and predictions for both Nile stations remained poor. At midlatitudes, both up- and downstream predictions for the Huanghe and Amur were improved, although downstream predictions for the Mississippi were improved while upstream predictions were worse. At high latitudes, up- and downstream predictions were generally improved for the Lena, Mackenzie, and Yukon. In most of the basins mentioned above, where annual average flow prediction was improved (degraded), prediction of the annual cycle of river flow (as assessed by the RRMSE) was also improved (degraded). The one exception to this was the Mackenzie, where simulation of annual average flows was improved, while simulation of the annual cycle was degraded in HadGEM1, relative to HadCM3.

Comparisons between errors in simulated river flows and precipitation biases in the GCMs can provide useful insights into the causes of these deficiencies—and particularly in assessing whether they are caused by precipitation biases or other causes. Johns et al. (2003) and Martin et al. (2006) have assessed general precipitation biases in HadCM3 and HadGEM1, respectively. Both HadCM3 and HadGEM1 tend to overestimate global precipitation (Nohara et al. 2006). HadGEM1 has too much precipitation (annually) over the Southern Ocean and the high latitudes of the North Atlantic and North Pacific. These errors are broadly similar to those exhibited by HadCM3 (Johns et al. 2006), suggesting that both models have an overly strong hydrological cycle (Pardaens et al. 2003). Most GCMs generally have too many days with rainfall (Dai 2006; Ines and Hansen 2006), so daily rainfall totals tend to be lower and more rainfall can evaporate or infiltrate (Sperna Weiland et al. 2010). We now focus on the gauging stations where the difference between GRDC observations and HadGEM1 predictions of annual average flow exceeded 50% to determine whether the errors in simulated river flows were reflected in precipitation biases for the basins relating to these stations in HadGEM1. We also discuss simulations at high-latitude gauges where there were significant errors, since errors in HadGEM1-simulated annual average flows did not exceed 50% at any high-latitude gauging station. For most of these stations, errors in simulated river flow are reflected in the general direction of GCM precipitation biases—overestimation of rainfall in summer and winter for the Mekong, St. Lawrence, Huanghe, Chang Jiang, Indus, Mackenzie, and Yukon, and overestimation of winter rainfall for the Orange and underestimation of winter rainfall plus overestimation of summer rainfall for the Orinoco (Tama–Tama). Thus, it seems likely that precipitation biases in HadGEM1 were a contributing factor to the errors in river flow for these basins.

However, for several stations, errors in river flow were not reflected in precipitation biases, so other reasons for river flow errors must be considered—for example, evapotranspiration, runoff generation (e.g., due to errors in soil moisture storage or timing and volume of snowmelt), in the river-routing model itself, or missing processes (e.g., abstraction, lakes, dams, or land use change). At Sao Paulo di Olivenca (Amazon), HadGEM1 tends to underpredict (boreal) summer rainfall in contrast to the overprediction in river flow—it thus seems likely that the source of the error in river flows is related to either evapotranspiration, runoff production, or floodplain processes (see later discussion). At Alton (Mississippi) the precipitation biases are more complex, including overestimation of winter rainfall and a general underestimation of summer rainfall that do not match the overestimation of summer river flow, although this may partly be explained by overestimation of summer rainfall in the upper reaches of the basin. For the two downstream stations on the Niger, HadGEM1 tends to underpredict summer rainfall although the predicted winter river flow is overestimated. Rainfall was generally overestimated for the Severnaya Dvina while river flow was underpredicted, and there was a slight overestimation of summer precipitation for the Lena while late spring–early summer river flow was underpredicted. However, skill in predicting monthly flow for high-latitude basins such as the Lena is closely linked to the ability of the model to simulate snow accumulation and melt processes and seasonal changes in temperature (see, e.g., Biancamaria et al. 2009), which we have not assessed in detail here. Note also that precipitation observations (and hence GCM performance) in high-latitude regions are subject to large uncertainties because of undercatch of solid precipitation and sparse observation networks (see, e.g., Yang et al. 2005).

To further investigate the causes of discrepancies between modeled and observed river flows, we compared basin-averaged precipitation values produced by HadCM3 and HadGEM1 with values from version 2 of the Global Precipitation Climatology Project (GPCPv2; Adler et al. 2003) and calculated basin-averaged evaporation and unrouted runoff values from the two models (Table 3; Figs. 4 and 5). Fekete et al. (2004) compare six recent global gridded precipitation datasets (including GPCP) and their impact on runoff estimates; GPCPv2 was chosen here because it is the dataset being used in the development of current Met Office Hadley Centre climate models (e.g., HadGEM3; Hewitt et al. 2010). For many basins, overall errors in precipitation patterns in HadCM3 and HadGEM1 appear to explain a considerable proportion of river flow errors—for instance, large RRMSE errors (Table 3) in both precipitation and river concentrations.
flow are reported for the Brahmaputra, Chari, Huanghe, Changjiang, Orange, Indus, Lena, and Mackenzie. In other basins (for instance, the Nile, Mississippi, and Volga), relatively large errors in river flows resulted despite fairly small precipitation errors. For most basins, RRMSE and MAPE show similar error trends in precipitation for both models, with estimated MAPE values usually being close to, or slightly smaller than, the RRMSE values. Some notable exceptions were found—in the HadCM3 simulations, MAPE was much larger.
than RRMSE for the Niger and Chari, and also larger for the Irrawaddy, Brahmaputra, Huanghe, Orange, and Nile; in HadGEM1, MAPE was larger for the Brahmaputra, Indus, and Amazon basins.

Inspection of seasonal patterns of precipitation, evaporation, and runoff for the basins studied here (Figs. 4 and 5) reveals further insight into potential causes of biases in river flow in the two models. We focus mostly on river
flow errors at the downstream gauges in each basin since these may best represent the basin behavior as a whole, although there may be exceptions where anthropogenic influences or other processes affect downstream flows significantly. Errors in the seasonal cycle of precipitation also appear to explain some aspects of river flow biases for many basins and the differences between the two GCMs. For example, in HadGEM1 both precipitation and river flows were overpredicted for the Amazon, Irrawady (especially in summer), St. Lawrence, Huanghe, Changjiang,
Amur, Indus, Mackenzie, and Yukon. HadCM3 overestimates precipitation and river flow for the Zaire, while underestimates were observed in HadGEM1. For some basins, however, errors in seasonal precipitation patterns did not explain the errors in simulated river flows. The patterns of precipitation in both HadCM3 and HadGEM1 over the Nile basin are close to that observed, while large overestimates in river flow were found for both models, with significant errors in seasonality. In the Brahmaputra basin both models overestimate summer precipitation, although the error is more concentrated in the first half of the year in HadGEM1 and the second half of the year in HadCM3. In contrast, the error in river flow in HadGEM1 is largest during the second half of the year and HadCM3 underestimates river flows overall, although precipitation errors are actually smaller in HadCM3. This may suggest that HadGEM1 may be producing too much runoff and HadGEM1 too little. HadGEM1 and HadCM3 both produce precipitation patterns close to the observations for the Mekong, but river flow is underestimated in HadCM3 and overestimated in HadGEM1, both with peaks too late in the year; runoff production in HadGEM1 is greater than in HadCM3 for similar precipitation totals. Relatively large overestimates in winter river flows were found in both models for the Niger despite relatively small precipitation errors, especially for HadGEM1. For the Chari, HadGEM1 overestimates precipitation, particularly in summer, while HadCM3 estimates were similar to observations but with a slightly underestimated summer peak and small excesses in spring and autumn. However, river flows for the Chari were overestimated in HadCM3 and similar to observations for HadGEM1, potentially indicating issues with runoff generation or other errors. Summer river flows were overpredicted in
both models despite relatively small precipitation errors; very different runoff and evaporation patterns were produced by the two models. Both HadCM3 and HadGEM1 tend to overpredict precipitation over the Lena, particularly during summer, although river flows are underpredicted in both models with a delayed peak; the two models produce very different evaporation and runoff patterns, and a similar but less pronounced trend is seen for the Yenisey. As previously noted, seasonal river flow prediction in high-latitude basins is strongly related to the simulation of temperatures and snowmelt and accumulation processes (Nohara et al. 2006; Biancamaria et al. 2009; Sperna Weiland et al. 2010; Haddeland et al. 2011).

Recent work has shown that human interventions (such as dams and irrigation) can have a significant impact on the water cycle (Nilsson et al. 2005; Hansaki et al. 2006; Sperna Weiland et al. 2010). The impact of human interventions on global discharge to the oceans is small (Döll and Siebert 2002; Döll et al. 2009; Wisser et al. 2010)—for instance, global discharge may have been decreased by 2.7% because of withdrawals and by 0.8% because of dams (Döll et al. 2009). However, the impact on annual flows in individual basins may be much larger. Döll et al. (2009) found that annual flows decreased by over 10% on one-sixth of the land’s surface because of human interventions (mainly irrigation), with reductions of nearly 100% of natural discharge in some basins. The impact on seasonal flow patterns within individual basins may also be considerable, with human interventions mostly decreasing amplitude (Döll et al. 2009; Wisser et al. 2010). For these reasons, detailed validation for some rivers such as the Nile (where large errors in river flow were found) was inappropriate, since human intervention is not accounted for in the version of TRIP used here and detailed information on abstraction and human interventions would be needed to investigate this further. Difficulties in reproducing discharge values for arid region rivers such as the Nile, Niger, and Orange have been noted by other authors (Oki et al. 1999; Nijssen et al. 2001b; Arnell 2003; Alcamo et al. 2003a,b; Döll et al. 2003; Nohara et al. 2006; Haddeland et al. 2011) and attributed both to sensitivity to water use (by irrigation and dams) and missing processes (e.g., evaporation from the river surface, groundwater losses, phreatophyte evaporation, and evaporation from small ephemeral ponds). In addition, neither HadCM3 nor HadGEM1 include floodplain inundation processes or

<table>
<thead>
<tr>
<th>Basin</th>
<th>Mean annual precipitation (mm day$^{-1}$)</th>
<th>Mean annual evaporation (mm day$^{-1}$)</th>
<th>Mean annual runoff (mm day$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GPCPv2</td>
<td>HadCM3</td>
<td>HadGEM1</td>
</tr>
<tr>
<td>Amazon</td>
<td>5.2</td>
<td>4.9</td>
<td>6.7</td>
</tr>
<tr>
<td>Congo–Zaire</td>
<td>3.9</td>
<td>5.4</td>
<td>3.8</td>
</tr>
<tr>
<td>Nile</td>
<td>1.8</td>
<td>2.3</td>
<td>1.7</td>
</tr>
<tr>
<td>Irrawaddy</td>
<td>4.4</td>
<td>4.8</td>
<td>6.3</td>
</tr>
<tr>
<td>Brahmaputra</td>
<td>3.2</td>
<td>4.2</td>
<td>6.1</td>
</tr>
<tr>
<td>Mekong</td>
<td>4.1</td>
<td>4.1</td>
<td>4.4</td>
</tr>
<tr>
<td>Orinoco</td>
<td>5.1</td>
<td>4.3</td>
<td>5.6</td>
</tr>
<tr>
<td>Niger</td>
<td>2.0</td>
<td>2.5</td>
<td>1.3</td>
</tr>
<tr>
<td>Chari</td>
<td>2.3</td>
<td>3.1</td>
<td>1.6</td>
</tr>
<tr>
<td>Mississippi</td>
<td>2.3</td>
<td>2.7</td>
<td>2.2</td>
</tr>
<tr>
<td>St. Lawrence</td>
<td>2.7</td>
<td>2.6</td>
<td>2.7</td>
</tr>
<tr>
<td>Huanghe (Yellow)</td>
<td>1.3</td>
<td>2.1</td>
<td>1.9</td>
</tr>
<tr>
<td>Changjiang (Yangtze)</td>
<td>2.9</td>
<td>4.2</td>
<td>4.4</td>
</tr>
<tr>
<td>Amur</td>
<td>1.7</td>
<td>1.9</td>
<td>1.9</td>
</tr>
<tr>
<td>Orange</td>
<td>1.1</td>
<td>1.8</td>
<td>1.8</td>
</tr>
<tr>
<td>Volga</td>
<td>2.0</td>
<td>1.5</td>
<td>1.8</td>
</tr>
<tr>
<td>Indus</td>
<td>1.3</td>
<td>0.9</td>
<td>1.6</td>
</tr>
<tr>
<td>Kolyma</td>
<td>1.0</td>
<td>1.0</td>
<td>1.1</td>
</tr>
<tr>
<td>Lena</td>
<td>1.1</td>
<td>1.3</td>
<td>1.5</td>
</tr>
<tr>
<td>Ob</td>
<td>1.6</td>
<td>1.3</td>
<td>1.5</td>
</tr>
<tr>
<td>Yenisey</td>
<td>1.4</td>
<td>1.7</td>
<td>1.7</td>
</tr>
<tr>
<td>Mackenzie</td>
<td>1.1</td>
<td>1.6</td>
<td>1.6</td>
</tr>
<tr>
<td>Yukon</td>
<td>1.6</td>
<td>1.5</td>
<td>1.6</td>
</tr>
<tr>
<td>Severnay Dvina</td>
<td>2.2</td>
<td>1.7</td>
<td>2.1</td>
</tr>
</tbody>
</table>
complex wetland processes, which have significant impacts in basins such as the Amazon and Congo/Zaire (Chapelon et al. 2002; Döll et al. 2003; Haddeland et al. 2011). Understanding the causes of these errors in river flow demands particular attention for those basins where significant impacts of climate change are anticipated (Falloon and Betts 2006)—for example, large decreases in river flow are projected for the Amazon, Niger, Orinoco, and Mississippi, and large increases for the Mekong, Changjiang, and Lena.

The river flow simulations produced by TRIP within HadCM3 and HadGEM1 compare reasonably well with results from similar global-scale studies (see discussion in section 2). For instance, we found correlation coefficients of 0.67 and 0.88, while Arnell (2003) found a value of 0.91 for a calibrated global hydrology model driven by observed climatology data. The range in RRMSE values in our study was 20.7%–1023.5%, while Nohara et al. (2006), also using the TRIP river flow model (but driven in a stand-alone model by runoff data from a multimodel ensemble of GCMs, including HadCM3 and HadGEM1) was 19%–1886%. While some basins were simulated well, our study suggests that (i) accurately simulating monthly river discharge with GCMs remains challenging and (ii) better GCM improvements are needed to achieve this, including better GCM simulation of basin-scale precipitation and evaporation, higher-resolution routing networks, and hydrological processes (e.g., representation of dryland hydrology, floodplain inundation, lakes, snowmelt, and human intervention).

4. Conclusions

TRIP predictions of long-term-averaged annual discharge were improved at 28 out of 40 gauging stations on 24 of the world’s major rivers in HadGEM1 compared to HadCM3, particularly for low- and high-latitude basins, with predictions ranging from good (within 20% of observed values) to poor (biases exceeding 50%). Since there are no differences in the underlying TRIP river flow model used in both cases, this is likely to be the result of improved climate simulation and improvements to the annual mean runoff produced by the hydrology scheme. Hence, this provides a useful additional method for comparing the climate simulation skill of different climate models, particularly because discharge is an integrated measure representing the effect of a whole range of upstream catchment processes. However, we have not compared the impact of including or excluding the TRIP river-routing model in a GCM here.

<table>
<thead>
<tr>
<th>Table 3. (Extended)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Annual precipitation difference (GPCP model), mm day$^{-1}$</strong></td>
</tr>
<tr>
<td><strong>Annual precipitation difference (GPCP model), %</strong></td>
</tr>
<tr>
<td><strong>RRMSE (%)</strong></td>
</tr>
<tr>
<td><strong>MAE (mm day$^{-1}$); MAPE (%) in parentheses</strong></td>
</tr>
<tr>
<td>HadCM3</td>
</tr>
<tr>
<td>-0.3</td>
</tr>
<tr>
<td>1.4</td>
</tr>
<tr>
<td>0.5</td>
</tr>
<tr>
<td>0.4</td>
</tr>
<tr>
<td>1.0</td>
</tr>
<tr>
<td>0.1</td>
</tr>
<tr>
<td>-0.8</td>
</tr>
<tr>
<td>0.5</td>
</tr>
<tr>
<td>0.8</td>
</tr>
<tr>
<td>0.4</td>
</tr>
<tr>
<td>-0.1</td>
</tr>
<tr>
<td>0.8</td>
</tr>
<tr>
<td>1.3</td>
</tr>
<tr>
<td>0.2</td>
</tr>
<tr>
<td>0.7</td>
</tr>
<tr>
<td>-0.5</td>
</tr>
<tr>
<td>-0.4</td>
</tr>
<tr>
<td>0.0</td>
</tr>
<tr>
<td>0.2</td>
</tr>
<tr>
<td>-0.3</td>
</tr>
<tr>
<td>0.3</td>
</tr>
<tr>
<td>0.5</td>
</tr>
<tr>
<td>0.0</td>
</tr>
<tr>
<td>-0.5</td>
</tr>
</tbody>
</table>
For most regions, the modeled annual average river flows tended to be exaggerated in both models, largely reflecting inflated estimates of precipitation. Within individual river basins, there were no clear trends in the accuracy of HadGEM1 versus HadCM3 predictions at up- or downstream gauging stations. RRMSE scores for the annual cycle of river flow ranged from poor (>$20\%$) to fair ($20\%$–$50\%$), with an overall range of $20.7\%$–$1023.5\%$, comparable to that found in similar globalscale studies. In both models, simulations of the annual cycle of river flow were generally better for high-latitude basins than in low or midlatitudes. This is likely because...
of the strong seasonality of the hydrological system in cold climate regions that is dominated by snowmelt in spring, which apparently is reproduced well by the GCM. We found a relatively small improvement in the annual cycle of river flow in HadGEM1 compared to HadCM3, mostly in the low-latitude rivers. TRIP, in combination with HadGEM1, has been shown to provide reasonable simulations of river flow (when compared to other global-scale studies) as input to the ocean and for investigating the hydrological effects of climate change in basins where the model performance can be shown to be acceptable (i.e., where the timing of the annual flood cycle is predicted correctly and where simulated monthly discharges are within, or at least close to, likely errors in
gauge observations). However, further improvements can clearly be made for many of the major river basins studied here.

In particular, further investigation should be carried out for gauging stations with a large difference between observed and predicted annual average flow (>50%), as identified above, and where the annual cycle is poorly predicted. In particular, HadGEM1 tends to overpredict river flows at low latitudes and particular attention should be given to the Amazon, Mekong, Orinoco, Nile,

**FIG. 5.** Basin-averaged mean annual cycle of precipitation from GPCPv2 (solid black line) and HadGEM1 simulations (solid blue line), and mean annual cycle of evaporation (dotted blue line) and runoff (dashed blue line) from HadGEM1.
and Niger basins. Basins at midlatitudes with large errors (largely also overestimates) in predicted river flow included the Mississippi, St. Lawrence, Huanghe, Orange, Changjiang, and Indus. High-latitude basins with large errors (mostly underestimates) in predicted river flow included the Severnay Dvina, Yukon, Mackenzie, and Lena, although no high-latitude basins had errors in annual average flow exceeding 50%. Errors in HadGEM1- and HadCM3- predicted river flow were largely reflected in precipitation biases in HadGEM1 and HadCM3, although this was not the case for the Amazon, Mississippi, Niger, Severnay-Dvina, Lena, and Nile.
Future work should attempt to separate errors in predicted flow arising from the TRIP model and its parameterization (the model as implemented here is not calibrated on a basin-by-basin basis) from errors in predicted flow arising from discrepancies in climate and input runoff from the GCM (for instance, using higher-resolution climate models may reduce errors). In turn, this requires a more detailed approach to understanding predictive skill in the underlying processes for river flow simulation in GCMs (for instance, basin-scale precipitation, evaporation (e.g., Blyth et al. 2010), soil moisture changes, snowmelt, dryland hydrology, lakes, floodplain processes, and subgrid heterogeneity; Döll et al. 2003). A wider set of assessment metrics could also be employed, including consideration of observational errors (which would assist in judging good or poor model performance). Further work should also consider appropriate means of weighting (i) annual average flow magnitudes against seasonality within basins and (ii) assessment metrics across basins, considering flow volume and basin size. From a hydrological perspective, further focus on significant phase errors would also be desirable.

Time series (rather than long-term averages) and higher-frequency observations could also be used to validate model predictions of variability and extremes in river flow. Key (but currently missing) processes such as human interventions (e.g., irrigation and dams) need to be considered in coupled GCM–river flow models to improve predictions and allow a fairer comparison with observations. For instance, Hanasaki et al. (2006) recently developed a global reservoir scheme for the TRIP model, which improved predictions for basins such as the Nile; however, inclusion of human interventions within GCMs (rather than land surface schemes or hydrological models) may present challenges—for example, appropriate representation of changes in dams or abstraction through time. Finally, the only feedback between TRIP and HadGEM1 is through the timing of the freshwater flux into the ocean. Further research should also assess whether processes depending on the timing of this freshwater influx in HadGEM1 are better resolved by the inclusion of a river-routing scheme.

Acknowledgments. The role of PF, RB, AW, CM, and RD was supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101) and the European Union project “Water and Global Change (WATCH)” 036946-2.

REFERENCES


——, 1999b: Climate change and global water resources. Global Environ. Change, 9, S31–S49.


Richards, L., 1931: Capillary conduction of liquids through porous mediums. Physics, 1, 318–333.
Smith, J. U., P. Smith, and T. M. Addiscott, 1996: Quantitative methods to evaluate and compare soil organic matter (SOM)


