Hydrol. Earth Syst. Sci. Discuss., 12, 10289–10330, 2015
www.hydrol-earth-syst-sci-discuss.net/12/10289/2015/
doi:10.5194/hessd-12-10289-2015

Hydrology and
Earth System
Sciences
Discussions
Open Access

Evaluation of five hydrological models across Europe and their suitability for making projections under climate change

W. Greuell[1], J. C. M. Andersson[2], C. Donnelly[2], L. Feyen[3], D. Gerten[4], F. Ludwig[1], G. Pisacane[5], P. Roudier[3], and S. Schaphoff[4]

[1]Earth System Sciences (ESS) group, Wageningen University and Research centre, Droevendaalsesteeg 3, 6708 PB Wageningen, the Netherlands
[2]SMHI – Swedish Meteorological and Hydrological Institute, Hydrological Research, 601 76 Norrköping, Sweden
[3]Climate and Risk management unit, Institute for Environment and Sustainability (IES), Joint Research Centre (JRC), European Commission (EC), Via Enrico Fermi, 2749, 21027 Ispra VA, Italy
[4]Potsdam Institute for Climate Impact Research (PIK), Research Domain I: Earth System Analysis, Telegrafenberg A62, 14473 Potsdam, Germany
[5]Italian National Agency for New Technologies, Energy and Sustainable Economic Development (ENEA), Lungotevere Thaon di Revel, 76 – 00196 Rome, Italy

Received: 21 August 2015 – Accepted: 11 September 2015 – Published: 9 October 2015

Correspondence to: W. Greuell (wouter.greuell@wur.nl)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

The main aims of this paper are the evaluation of five large-scale hydrological models across Europe and the assessment of the suitability of the models for making projections under climate change. For the evaluation, 22 years of discharge measurements from 46 large catchments were exploited. In the reference simulations forcing was taken from the E-OBS dataset for precipitation and temperature, and from the WFDEI dataset for other variables. On average across all catchments, biases were small for four of the models, ranging between −29 and +23 mm yr^{-1} (−9 and +8 %), while one model produced a large negative bias (−117 mm yr^{-1}; −38 %). Despite large differences in e.g. the evapotranspiration schemes, the skill to simulate interannual variability did not differ much between the models, which can be ascribed to the dominant effect of interannual variation in precipitation on interannual variation in discharge. Assuming that the skill of a model to simulate interannual variability provides a measure for the model's ability to make projections under climate change, the skill of future discharge projections will not differ much between models. The quality of the simulation of the mean annual cycles, and low and high discharge was found to be related to the degree of calibration of the models, with the more calibrated models outperforming the crudely and non-calibrated models. The sensitivity to forcing was investigated by carrying out alternative simulations with all forcing variables from WFDEI, which increased biases by between +66 and +85 mm yr^{-1} (21–28 %), significantly changed the inter-model ranking of the skill to simulate the mean and increased the magnitude of interannual variability by 28 %, on average.

1 Introduction

Over the last decades a large of number of distributed large-scale hydrological models have been developed (see e.g. Donnelly et al., 2015), while land-surface schemes linked to river routing schemes have also emerged. An important motivation for their development is to use them as a tool for assessing the impact of e.g. climate, land-use or management changes on the water cycle. Recently, ensemble studies of such changes using multiple hydrological models have emerged as the state-of-the-art (e.g. Hagemann et al., 2013; Schewe et al., 2014; Van Vliet et al., 2015) because this allows estimation of the uncertainty in the results. In such multiple model studies, it is valuable to have knowledge about the absolute and relative skill of the individual models.

There exist a number of recent studies evaluating different aspects of an ensemble of large-scale hydrological models. For the USA, Xia et al. (2012) made a comprehensive evaluation of land-surface schemes at multiple catchment scales. For catchments in Europe, Gudmundsson et al. (2011, 2012) conducted a comprehensive evaluation of multiple global models, looking at percentiles of discharge including their interannual variability and seasonal variation. Using the same observation data set, Stahl et al. (2011) examined a number of simulated flow indices based on anomalies and Prudhomme et al. (2011) evaluated the performance of multiple models for extremes defined as regional flood and deficiency indices. A limitation of these studies is that the database of observations is limited to catchments mostly $< 500\,\mathrm{km}^2$ with a fairly uneven spatial distribution across Europe. These catchments are significantly below the grid scale on which the models were run, hence validation is based on assumptions for downscaling model grid results to catchment scale. Moreover, at these observational scales runoff generation processes dominate over routing processes so the skill of simulating routing cannot be assessed. The results of evaluation studies can be also limited by the scale of the input forcing data which in some of the mentioned studies was significantly larger than the catchment scale. Both Xia et al. (2012) and Gudmundsson et al. (2011) found that the ensemble mean performed better than any one model across their study domains.

In this paper we will discuss a systematic and extensive evaluation and intercomparison of discharge simulations by five hydrological models across Europe. The study is extensive in the sense that various aspects of the hydrograph are investigated, namely the mean, the interannual variability, the annual cycle, and low and high discharge. Contrary to the studies mentioned above, which are limited to evaluation at small scales, we limit the evaluation to large scales ($> 9900\,\mathrm{km}^2$), taking data from the stations that are situated closest to the mouth of large rivers. The first aim of the paper is to demonstrate by how much the models diverge or converge on the simulation of the various aspects of river discharge and to learn which aspects are generally well or badly captured.

The second aim of this paper is to use results of the evaluation to assess the suitability of hydrological models to assess climate change impacts. Previous papers that explored performance-based weighting of climate change projections computed weighting factors from the success of the models to reproduce the frequency of circulation types (Déqué and Somot, 2010) or by combining a set of performance metrics (Hesselbjerg et al., 2010). We propose that the skill of a hydrological model to simulate interannual variability in discharge is a good measure for its ability to make projections of climate changes in mean discharge. The reasoning for this proposal is that differences between e.g. drier vs. wetter years and warmer vs. colder years can be considered as an analogue for the projected climate changes. Hence, we assume that if a model is well capable of simulating the difference in discharge between dry and a wet years (or a warm and a cold years, etc.), it can be expected that the model is also well capable of simulating the climate change impacts on discharge. Thus, by evaluating the skill of the models to simulate interannual variability, we assess the relative skill of the different models to make projections of future change.

The hydrological models used in this study are not fully coupled to the atmosphere. Instead, they need to be forced by near-surface atmospheric variables like precipitation, air temperature and down-welling radiative fluxes. In an evaluation study one should thus aspire to perform model simulations forced by data sets that provide the best possible approximation of the near-surface climate during the evaluation period. In this study we exploited the E-OBS and the WFDEI data sets, two widely used observation-based data sets of excellent quality. In the reference validation simulations, we combined E-OBS precipitation and temperature with other variables from the WFDEI data

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
◀| |▶
◀ ▶
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

set. In alternative simulations all forcing variables, including precipitation and temperature, were taken from WFDEI. We analyse the differences between the two sets of simulations in order to quantify the sensitivity of our results to the uncertainty in the meteorological forcing.

2 Hydrological models

The following five continental-scale hydrological models participated in the present study: E-HYPE, Lisflood, LPJmL, VIC and WBM. Lisflood, VIC and WBM focus on the simulation of the terrestrial hydrological cycle while E-HYPE is broader in the sense that water quality is also simulated. LPJmL simulates both the water and carbon cycle coupled with vegetation dynamics. References and model characteristics most relevant for the analysis are listed in Table 1 and shall briefly be discussed here. Descriptions apply to the versions of the models used for the present study.

LPJmL, VIC and WBM can and have been applied globally, whereas E-HYPE is limited to Europe and Lisflood to Europe and Africa. Three of the models perform their simulations on a regular lat-lon grid, although at different resolutions. Lisflood was operated on a 5 km × 5 km grid and E-HYPE's simulation cells are sub-basins of irregular shape (median area of 215 km^2). The models compute evapotranspiration with widely differing concepts. WBM exploits an equation proposed by Hamon (1960), which only takes daily mean temperature as input, so variations in radiation are ignored. E-HYPE and Lisflood consider radiation in a simplified manner by estimating evapotranspiration with the equation of Hargreaves and Samani (1985). In this equation radiation is set equal to extraterrestrial radiation multiplied by a term representing atmospheric radiative transfer, which is parameterised as a function of daily maximum and minimum temperature. Variations in net surface radiation are directly considered in the method of Priestley and Taylor (1972), which is employed by LPJmL. However, outgoing fluxes are not computed by LPJmL itself but taken from the forcing. The most complex method, the one proposed by Penman and Monteith (see Shuttleworth, 1993), is used by VIC.

Title Page
Abstract Introduction
Conclusions References
Tables Figures
◀| |▶
◀ ▶
Back Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

This method closes the energy balance and internally calculates the outgoing radiative fluxes and the ground heat flux. Note that all methods have tunable parameters, varying from a single parameter in the Hamon equation, which summarizes effects of, among others, radiation, wind speed and humidity, to a larger set of parameters in the equations of Penman–Monteith, which affect e.g. stomatal resistance to transpiration and the reflection of short-wave radiation.

The models also differ in their treatment of surface runoff, soil and ground water, and river routing, see Table 1. Only E-HYPE and Lisflood compute the retarding effect of lakes on river routing while the other three models do not consider this effect. Another noticeable point is the correction that E-HYPE applies to the precipitation forcing. The motivation for this is to correct for relatively low precipitation gauge densities at high elevations and for undercatch in the observations. As a result EHYPE's precipitation is increased with respect to the precipitation forcing in the other models.

Differences in calibration methods between the models are important for the interpretation of the analyses of this paper. None of the parameters affecting the hydrological cycle directly were tuned in LPJmL and WBM, so for this study these models can be considered as non-calibrated models. Calibration of VIC, Lisflood and E-HYPE varies significantly. VIC and E-HYPE use calibration parameters that do not vary regionally across Europe. In VIC six surface and soil parameters were tuned assuming no variation across land use types, while in E-HYPE evapotranspiration, snow, runoff generation and routing parameters are soil and land use dependent. Also, in E-HYPE specific parameters are calibrated for the larger lakes across Europe. Lisflood is the only model with regionally varying calibration parameters determined by calibration for individual catchments. It is finally relevant to inform which forcing was used to calibrate the models. E-HYPE v2.1 was calibrated with Global Precipitation Climatology Centre (GPCC) corrected ERA-INTERIM forcing (Donnelly et al., 2015) but for this study (v2.5) the forcing was adjusted slightly with E-OBS. Lisflood was calibrated to a high-resolution (5 km) interpolated data set of meteorological observations composed at the JRC (Ntegeka et al., 2013) and VIC exploited a data set composed by Nijssen et al. (2001).

3 Data

3.1 Atmospheric forcing

Simulations were forced with gridded time series of variables that have been composed to make a best estimate of the surface meteorology during the period of the simulations (1970–2010). Values of the various forcing variables were taken from three widely used gridded data sets:

1. E-OBS (version 9, see Haylock et al., 2008) provided precipitation, and minimum, mean and maximum temperature for 1970–2010. The E-OBS data are based on daily station data collected in the ECA&D (European Climate Assessment and Data) archive. To compute grid cell means, the ECA&D data were interpolated (Hofstra et al., 2009) in all three dimensions (longitude, latitude and surface elevation). The station density of ECA&D varies hugely across countries and regions, with e.g. very dense networks in Ireland and the Netherlands but an order of magnitude sparser networks in countries like Austria and Switzerland (~ 15 and ~ 30 stations, respectively, for 1971–2000) where a considerable part of Europe's runoff originates. To produce E-OBS precipitation, no correction for undercatch was applied. We downloaded E-OBS data at a resolution of $0.25^\circ \times 0.25^\circ$ and aggregated these data to a resolution of $0.5^\circ \times 0.5^\circ$.

2. We also used precipitation, minimum, mean and maximum temperature, specific humidity, wind speed, incoming short- and long-wave radiation and net radiation from WATCH forcing data ERA-Interim (WFDEI, Weedon et al., 2014) for 1979–2010. Daily variation of WFDEI precipitation is derived from the ERA-Interim reanalysis. Monthly totals of the WFDEI precipitation are bias-corrected towards gridded observations. This is done twice, once with Global Precipitation Climatology Centre (GPCC) data (Adler et al., 2003) and once with Climatic Research Unit (CRU) data (Mitchell and Jones, 2005), leading to two version of the WFDEI precipitation data. Unless mentioned otherwise, the GPCC-version is used here.

Even though there is considerable variability in the GPCC station density across Europe, this variability is smaller than in the ECA&D precipitation station network. For the present study it is very important to notice that in a final step WFDEI precipitation was corrected for undercatch by multiplication with the factors provided by Adam and Lettenmaier (2003). These factors are given on a grid of 0.5° × 0.5° and per month. A drawback of the gridded GPCC precipitation data is that they have been produced without any consideration of the elevation of the precipitation gauges in the gridding procedure. Gauges tend to be relatively more numerous at lower elevations and since precipitation tends to increase with elevation, this could result in an underestimation of the precipitation in the WFDEI precipitation data. WFDEI data sets of most other (not precipitation) variables were, like precipitation, derived by combining monthly observations with daily ERA-Interim reanalyses.

3. Since WFDEI data start from 1979 only, we used its precursor the WATCH forcing data (WFD, see Weedon et al., 2011) for 1970–1978. Simulations for 1970–1978 only serve as spin up of the hydrological models.

Figure 1 compares precipitation averaged over 1979–2000 between E-OBS and WFDEI. On average across the domain WFDEI precipitation exceeds E-OBS precipitation by 104 mm yr^{-1}. The difference varies spatially but WFDEI exceeds E-OBS precipitation almost everywhere. Exceptions are some regions with much relief and high annual precipitation amounts like parts of the Alps and South Norway where E-OBS exceeds WFDEI precipitation. Removing the undercatch correction from the WFDEI data reduces the domain-averaged difference to 5 mm yr^{-1}, so the domain-averaged difference is almost entirely due to the undercatch correction applied to the WFDEI data.

3.2 Discharge measurements

The discharge observation stations used in this study (Table 2 and Fig. 2) were selected using the following criteria:

- Catchment upstream from the station has a catchment area $> 9900\,\mathrm{km}^2$. The main argument for the threshold is to minimize contributions to the differences between simulations and observations caused by discrepancies between the catchment areas of both types of data.
- Time series is complete for the period 1979–2000.
- If more than one station on a river complies with the first two criteria, the most downstream station is selected.

The primary data source were daily values from the archive of the Global Runoff Data Centre (GRDC, http://www.bafg.de/GRDC/EN/Home/homepage_node.html), resulting in time series from 39 stations after application of the three selection criteria. Seven of these stations are situated along tributaries of larger rivers (e.g. the Main as a tributary of the Rhine) which themselves are also represented in the selection. Twenty-three of the selected rivers are located in Fennoscandia. In order to have specific information for the Alps, GRDC data from two stations situated near the foot of this mountain range, namely Rheinfelden along the Rhine and Chancy along the Rhone, were added. Data from rivers on the Iberian Peninsula, in France, Italy and Poland were lacking in the GRDC data set. To fill this gap, data for seven rivers in these countries produced by Dai et al. (2009) were added to the validation data. These are monthly values, mostly observed, but all of the seven time series were incomplete for 1979–2000 when taking observations only. Missing parts of the time series were completed by Dai et al. (2009) using hydrological model simulations corrected towards the observations.

Based on our own expert knowledge, we subdivided the 46 resulting time series into two groups. The first contains all 26 rivers that are negligibly affected by reservoir regulation. The second part contains those 20 rivers of which the discharge is judged to be substantially affected on time scales of a year and shorter by human interference. We used the data from all 46 stations for the evaluation of the mean discharge while only the data from the first group were exploited to evaluate interannual variability, the annual cycle, and low and high discharge.

4 Simulations

To quantify the sensitivity of the results to the uncertainty in the forcing, two simulations for the period 1970–2010 were carried out at daily or 3 hourly time step with each of the hydrological models:

1. The "E-OBS simulation", in which the forcing consists of precipitation and temperature from E-OBS and of other variables from WFD (1970–1978) and WFDEI (1979–2010).
2. The "WFDEI simulation", for which all forcing variables, including precipitation and temperature, were taken from WFD (1970–1978) and WFDEI (1979–2010).

So, the two simulations differ in the precipitation and temperature forcing only. All models were evaluated using daily values of the discharge. For the analysis all discharge simulations were delivered on a lat-lon grid with a resolution 0.5°. For this purpose the output from the models operated at higher (Lisflood and WBM) or variable (E-HYPE) resolution was aggregated or resampled. All models were run in naturalized flow mode meaning river regulation, irrigation and other anthropogenic influences were not simulated. Domain boundaries were located at 25° W, 40° E, 33 and 72° N.

5 Evaluation metrics

Most metrics of this paper quantify skill of simulating discharge across all observation stations. Unless otherwise mentioned these statistics are computed by weighting the contribution from each station by the area of the catchment upstream from the station (A). In the equations of this section the subscripts s and m represent the simulations and measurements, respectively. Angle brackets ($\langle\ \rangle$) denote the mean across all stations.

Equations for widely used metrics like the bias and the correlation coefficient are considered familiar to the reader. Equation (1) provides the expression for R_σ, the ratio of the standard deviations in the simulated and the measured discharge statistic X (can be mean, 5th or 95th percentile of all daily values in this paper):

$$R_\sigma = \sqrt{\frac{\sum\limits_{\mathrm{i}} A_{\mathrm{i}}(X_{\mathrm{is}} - \langle X_{\mathrm{s}}\rangle)^2}{\sum\limits_{\mathrm{i}} A_{\mathrm{i}}(X_{\mathrm{im}} - \langle X_{\mathrm{m}}\rangle)^2}} \tag{1}$$

where the subscript i denotes an individual station.

The first metric used to evaluate the skill of simulating interannual variability is ⟨CRMSEann⟩, the root of the station-averaged CMSEann, the latter being the centred mean square error of annual discharge for a single station:

$$\mathrm{CMSEann} = \frac{\sum\limits_{j=1}^{n} [(q_{j\mathrm{s}} - Q_{\mathrm{s}}) - (q_{j\mathrm{m}} - Q_{\mathrm{m}})]^2}{n-1} \tag{2}$$

where j is the year index, n the number of years (22 in this analysis), q annual discharge and Q the temporal mean of q. In words, the CMSEann is, in a scatter plot with annual values of simulated and observed discharge values for a single station, a measure for the distance of the points to the 1 : 1 line. It is important to notice that the bias is eliminated by taking annual anomalies and that hence a potential bias does not contribute to this metric. From the CMSEann of all individual stations, ⟨CRMSEann⟩ is computed by taking the root of the weighted mean:

$$\langle \mathrm{CRMSEann}\rangle = \sqrt{\langle \mathrm{CMSEann}\rangle}. \tag{3}$$

The second metric used to evaluate the skill of simulating interannual variability is the station-averaged ratio of the standard deviations of annual discharge in the simulations and the measurements $\langle\sigma_\mathrm{s}/\sigma_\mathrm{m}\rangle$, with the contribution from each station equal to:

$$\frac{\sigma_\mathrm{s}}{\sigma_\mathrm{m}} = \frac{\sqrt{\sum_{j=1}^{n}(q_{j\mathrm{s}} - Q_\mathrm{s})^2}}{\sqrt{\sum_{j=1}^{n}(q_{j\mathrm{m}} - Q_\mathrm{m})^2}}. \tag{4}$$

This metric quantifies whether a simulation has skill in the simulation of the magnitude of the interannual variability but it is not sensitive to having the simulated anomalies in the correct sequence. Note that CRMSEann can be written as a unique function of σ_s, σ_m and the correlation coefficient between the annual anomalies of the measurements and the simulations (Taylor, 2001).

Finally, simulation of the mean annual cycle for each station is quantified by the Centred Root Mean Square Error of monthly means:

$$\mathrm{CRMSEmth} = \sqrt{\frac{\sum_{k=1}^{12}[(\hat{q}_{k\mathrm{s}} - Q_\mathrm{s}) - (\hat{q}_{k\mathrm{m}} - Q_\mathrm{m})]^2}{(12-1)}} \tag{5}$$

where k is month number and $\hat{q}$ mean (over all years) monthly discharge. This metric measures the difference between simulated and observed mean monthly discharge, averaged over the 12 months of the year. Again the bias is intentionally eliminated by subtracting mean discharge so that a potential bias does not contribute to this metric.

To obtain insight into the performance of the hydrological models, we chose to evaluate different aspects of the hydrograph like bias, the interannual variability and the annual cycle separately. This logically led to not considering commonly used criteria like the Nash–Sutcliffe efficiency (Nash and Sutcliffe, 1970) and the Kling–Gupta efficiency (Gupta et al., 2009). These efficiencies measure the overall performance of a simulation without specifying the contributions of the various aspects of the hydrograph to the overall performance.

6 Evaluation

6.1 Introduction

Throughout the analysis we consider discharge by dividing measured and modelled discharge values by the area of the catchment area upstream from the station. However, for measurements and simulations we used different sources to determine the catchment area, namely the metadata of the observations and the sum of the area of the upstream cells or basins that contribute to the modelled discharge, respectively. By doing so, effects of differences between observed and modelled catchment area, which can be significant (Donnelly et al., 2012), were eliminated.

As already indicated, domain-average statistics were calculated by weighting contributions from stations with the area of the catchment. As a result and taking all 46 stations, the Danube contributes 26 % to the total while contributions of the five regions are 8 % for the Mediterranean, 2 % for the Alps, 11 % for Fennoscandia south of 62° N, 12 % for Fennoscandia north of 62° N and 67 % for Central Europe (broadly from the Thames to the Rhone to the Danube to the Neman). Limiting the data set to rivers with little regulation (group 1; 26 rivers), these contributions are 36 % (Danube), 0 % (Mediterranean), 1 % (Alps), 8 % (South Fennoscandia), 6 % (North Fennoscandia) and 85 % (Central Europe).

Evaluation of the modelled discharge was limited to 1979–2000. The starting year (1979) was defined by the starting year of WFDEI. The end year was set to 2000 since by taking a later end year the number of stations would seriously diminish due to time series becoming incomplete.

6.2 Mean discharge

The mean discharge is evaluated in Fig. 3 and Table 3. For all models the correlation coefficient, which is a measure of a model's skill to reproduce the spatial pattern of observed discharge, is high (between 0.942 and 0.966). The following deviations from perfect skill are noticeable:

1. Although the station-averaged bias is small for four out of the five models, varying from −29 to +23 $\mathrm{mm\,yr^{-1}}$ (−9 to +8 %), WBM has a substantially larger bias (−117 $\mathrm{mm\,yr^{-1}}$; −38 %).

2. As the high value of the ratio of the standard deviations in the x and the y direction ($R_\sigma = 1.29$) indicates, E-HYPE tends to produce an overestimate of discharge that increases with discharge itself while the overestimate tends to disappear or to become an underestimate in regions with low discharge. This tendency is seen in the points representing the rivers of Central Europe and South Fennoscandia and confirmed by the "high discharge rivers" of the Alps, at least partly due to the built-in precipitation correction applied in E-HYPE. In the Lisflood simulation the stations of Central Europe show the same tendency towards high R_σ but here the tendency is not confirmed by the Alpine rivers. Whereas VIC and WBM have an R_σ that is slightly less than 1.0, R_σ is only 0.77 in LPJmL. The low value of LPJmL is mainly due to the North Fennoscandian and the Alpine rivers, perhaps because this model did not adjust the complex flow regimes of these rivers by calibration.

3. In the plots for all models there tend to be two groups of points scattered around two lines, one for rivers of North Fennoscandia and one for the remaining rivers. The line for North Fennoscandia is, with respect to the line for the other regions, shifted towards lower simulated discharge. This also occurs for E-HYPE despite its built-in precipitation correction. The fact that all models, despite huge differences in the method of computing evapotranspiration, show the same behaviour, suggests that precipitation is underestimated in E-OBS in North Fennoscandia.

6.3 Interannual variability

The skill of the models to simulate interannual variability in discharge is illustrated in the Taylor diagrams of Fig. 4 and quantified by the station mean Centred Root Mean Square Error (⟨CRMSEann⟩) listed in Table 3. For one of the models as an example (LPJmL), Fig. 5 shows a scatter plot of the standard deviation of simulated annual discharge (σ_s) vs. the standard deviation in the measured annual discharge (σ_m), with each circle representing a basin.

The most important result is that the models do not differ much in their skill to simulate interannual variability, with ⟨CRMSEann⟩ ranging between 24 mm yr^{-1} (Lisflood) and 29 mm yr^{-1} (WBM).

Figure 5 shows that LPJmL is well capable of simulating the magnitude of the observed interannual variability. All points are close to the 1 : 1 line, with σ_s/σ_m ranging between 0.64 (Kymijoki) and 1.34 (Neris), the mean value across all basins ($\langle\sigma_s/\sigma_m\rangle$) being 0.89 (see Table 3). Also, the variation across the entire range of σ_m, as quantified by the correlation coefficient (0.945), is well captured by LPJmL. Similar scatter plots of σ_s vs. σ_m for the other models (not shown here) demonstrate that all models are quite capable of simulating the magnitude of the observed interannual variability, with $\langle\sigma_s/\sigma_m\rangle$ varying between 0.75 (Lisflood) and 1.09 (VIC), see Table 3.

Somewhat contradictory, for Lisflood $\langle\sigma_s/\sigma_m\rangle$ deviates more from the ideal value of one than for all other models, while it has the best ⟨CRMSEann⟩ of all models. The low value of $\langle\sigma_s/\sigma_m\rangle$ for Lisflood can also be seen in the Taylor diagram for this model, in which the majority of the basins are plotted below the circle of $\sigma_s = \sigma_m$. For the other models the basins are generally closer to the circle of $\sigma_s = \sigma_m$. In the Taylor diagrams, WBM exhibits a larger spread of the points than the other models do, especially in the radial direction σ_s/σ_m, so in WBM simulated interannual variability tends to have a larger spread around measured interannual variability than in the other four models. Nevertheless, ⟨CRMSEann⟩ is of the same order of magnitude as for the other models. We checked whether this was due to weighting individual stations by catchment area since e.g. the Danube with its large weight (36 %) is relatively well situated in the Taylor diagram of WBM. So, we recomputed ⟨CRMSEann⟩ giving equal weight to all stations. It was found that qualitatively this did not affect the conclusion: the ⟨CRMSEann⟩ of the five models remains within a rather narrow range (35–43 $\mathrm{mm\,yr^{-1}}$) and WBM's ranking in terms of ⟨CRMSEann⟩ (third) remains the same.

The best simulation of the interannual variability, in terms of the distance to the point of perfect simulation in the Taylor diagrams, was achieved for three Central European rivers, namely Rhine, Moselle and Weser. For LPJmL and WBM these three rivers form the top 3 while for all models they rank within the top 7. This high skill for all models could be caused by high quality of the forcing data within the catchments of these rivers, especially of the interannual variability of precipitation, and/or by relatively small deviations from naturalized flow.

6.4 Annual cycle

Twenty-two year mean modelled annual cycles were compared with the measurements for all of the 26 rivers with small human impacts. The six rivers shown in Fig. 6 are representative for the entire set, reflecting different flow regimes. The skill of reproducing the measured annual cycle is quantified by the Centred Root Mean Square Error of the monthly values (CRMSRmth). Though the bias was intentionally eliminated in the calculation of CRMSRmth, this was not done in the figures since that would remove the reference of no flow.

Large differences between the models in the skill of simulating the annual cycle are found. Relative to the other models, Lisflood performed the best, ranking as the number 1 in 17 out of the 26 cases and as the number 2 in all other cases. E-HYPE performs well in Fennoscandia, being the best or second best for 8 out of the 10 rivers in that region, but tends to show less performance elsewhere, e.g. for the Danube, which may be related to a single parameter describing routing at all scales in E-HYPE. For 8 out of the 10 rivers in Fennoscandia, LPJmL has the poorest simulation of the annual cycle among all models while that model is among the top 2 for only two out of the 26 rivers.

We speculate that these results are linked to too early and rapid melting of the snow pack. The annual cycles simulated by VIC are of intermediate quality. VIC tends to overestimate the amplitude of the annual cycle and in North Fennoscandia the spring peak is too late and lasts too long. Despite being the only model with a large bias in the E-OBS simulation, WBM takes a mid-position in the simulation of the bias-corrected annual cycle. For most river basins, WBM is quite correct in simulating the timing of the annual extremes but the amplitude is too small.

As to regional differences in skill, simulated annual cycles are generally best for rain-fed Central-European rivers like the Moselle, the Main and the Weser. The simulation of the Alpine rivers (Inn and Rhone) is more problematic. In the Alps WBM and LPJmL peak too early, probably because snowmelt occurs too early and/or too fast in these models, which in some basins also leads to an underestimation of discharge in summer. E-HYPE's overestimations of runoff from the Alps (see Fig. 3) appears to occur especially in summer. The annual cycle of the Losna (South Norway) is quite well simulated by most models but the other two Fennoscandian examples given in Fig. 4 are illustrative of some typical issues. The Vuoksi (Finland) runs through huge lakes along much of its course, so in the observations seasonal variations are almost absent. This is captured quite well by E-HYPE and Lisflood, which consider the effect of the lakes on river flow, but not by the other models which lack a description of this effect. Torneälven, situated in Swedish Lapland, has a very steep peak in May–June caused by snow melt and has a small but non-zero base flow in winter. Its annual cycle is well simulated by Lisflood and E-HYPE. VIC and WBM, however, severely underestimate winter discharge and VIC produces a spring/summer peak that is too wide. A sensitivity experiment demonstrated that for VIC these issues were largely due to inadequate simulation of frozen soils.

6.5 Low and high discharge

Low discharges were evaluated by comparing the fifth percentile (Q_5) of the simulated and observed daily values for the catchments with negligible human interference (Fig. 7

and Table 3). High discharge was evaluated similarly by analysis of the ninety-fifth percentile (Q_{95}; Fig. 8 and Table 3). Regarding low discharges, there is a clear difference in skill between the more calibrated models (E-HYPE and Lisflood), which perform relatively well, and the non-calibrated models (LPJmL and WBM), which underestimate low discharge for all basins. In the non-calibrated models, Q_5 from the smaller among the selected (large) basins is almost negligible compared to the observations. VIC takes an intermediate position but still clearly underestimates low discharge.

Most models are better at simulating high discharge than low discharge. Only for E-HYPE there is not much difference between the two. E-HYPE overestimates Q_{95} in the Alps, again probably due to its built-in undercatch correction. For Lisflood scatter around the 1 : 1 line is clearly smaller for high discharges than for low discharges, as quantified by correlation coefficients of 0.97 and 0.90, respectively, reflecting Lisfloods original purpose for flood forecasting. For LPJmL, VIC and WBM the absolute value of the relative bias is much smaller for high discharges (+29, +19 and −18 %, respectively) than for low discharges (−72, −38 and −64 %, respectively). LPJmL and VIC overestimate high discharges, so they are too extreme both for low and for high discharges. At the same time mean discharge is quite well captured by both models (relative biases of −9 % for LPJmL and +8 % for VIC). So, both models probably simulate mean evapotranspiration relatively well but they do not produce enough delay between rainfall and runoff. This is also the case for WBM but the bias in Q_{95} remains negative due to the negative bias in the mean (−38 %).

We also evaluated less severe hydrological droughts by comparing simulated with measured Q_{10} values. Relative biases for LPJmL (−58 %), VIC (−22 %) and WBM (−61 %) were still large though slightly closer to zero than for Q_5. A further evaluation for extremes of three of the here used models is presented by Roudier et al. (submitted) who considered once-in-100-year low and high discharges. Qualitatively, our conclusions about Q_5 and Q_{95} are identical to those in Roudier et al. (2015), both in terms of inter-model differences and in terms of low vs. high discharge.

7 Sensitivity to meteorological forcing

In order to study the sensitivity of the results to uncertainty in forcing, the five hydrological models were also run with WFDEI forcing, replacing the E-OBS forcing used to obtain the results of the previous section.

Before analysing the results, which generally show significant sensitivity to the forcing, two side issues shall be mentioned. These were addressed by performing sensitivity experiments with VIC. The first experiment showed that the differences in discharge between the WFDEI simulation and the E-OBS simulations are almost entirely due to the differences in precipitation between the two forcings and hardly due to the differences in temperature. In the second experiment the GPCC precipitation version of the WFDEI forcing was replaced by the version corrected with CRU precipitation. This increased the domain-averaged discharge by only $+3\,\mathrm{mm\,yr^{-1}}$, so the effect of choosing either the GPCC or the CRU precipitation data has a negligible effect on domain-averaged discharge.

In Fig. 9 we evaluate mean (1979–2000) simulated discharge for the “WFDEI simulation” for two models (VIC and WBM). Compared to the panels of the same models for the “E-OBS simulation” in Fig. 1, the main impression is that the entire pattern of points in the scatter plots shifts upward. The same occurs for the other three models. When averaged across all basins, simulated discharge shifts upward by 68 and $85\,\mathrm{mm\,yr^{-1}}$ (21 and 28 %) across the five models (Table 3). So, the increase is substantial and does not vary much between the models and river basins. Qualitatively the increase can be explained by the domain-averaged difference between the precipitation rates of WFDEI and E-OBS ($104\,\mathrm{mm\,yr^{-1}}$), of which the largest part runs off and a smaller part evaporates. There is also a considerable effect of replacing the forcing on the ranking of the models by the absolute value of the bias, e.g whereas WBM is an outlier (rank 5) with E-OBS forcing, it ranks as the number two with WFDEI forcing.

Table 3 shows that the magnitude of the interannual variability ($\langle\sigma_s/\sigma_m\rangle$) increases for all models when E-OBS is replaced by WFDEI forcing. The average over the models increases from 0.89 to 1.14 (by 28 %) and, whereas in the E-OBS simulation $\langle\sigma_\mathrm{s}/\sigma_\mathrm{m}\rangle$ is less than unity for four of the five models, it is larger than unity for four of the five models in the WFDEI simulation. This upward shift in $\langle\sigma_\mathrm{s}/\sigma_\mathrm{m}\rangle$ is also visible in the Taylor diagrams, for instance for VIC by comparing Fig. 10 with the panel for VIC in Fig. 4. We explain this increase in $\langle\sigma_\mathrm{s}/\sigma_\mathrm{m}\rangle$ by the larger amount of precipitation, and hence larger interannual variability in precipitation, in the WFDEI data as compared to the E-OBS data. Replacing the E-OBS by the WFDEI forcing also results in an increase of ⟨CRMSEann⟩ for four of the five models. The effect on ranking the models in terms of ⟨CRMSEann⟩ is only slight, with Lisflood (nr. 1) and LPJmL (nr. 5) remaining at their position and the other three models changing their positions.

More precipitation in the WFDEI forcing also affects the mean annual cycles. The general effect of the WFDEI forcing is to enhance discharge in all models, months and for all basins, and to increase the amplitude of the annual cycle. Figure 11 shows the example of the Danube for the WFDEI simulation, which should be compared with the panel for the Danube in Fig. 6. Again, ranking of the models in terms of their skill is changed. With WFDEI forcing WBM performs best of all models for nine basins, all but one situated in Central Europe, one being the Danube, while with E-OBS forcing WBM ranked as the number 1 for only three basins.

While the evaluation of Q_5 is not seriously affected by replacing the forcing, simulated Q_{95} is enhanced. As a result, all models overestimate the basin-averaged Q_{95} with WFDEI forcing, varying from +6 % (WBM) to +47 % (LPJmL).

In summary, the effect of enhanced precipitation in the WFDEI forcing is to increase various statistics of simulated discharge in all hydrological models and across the entire domain. This includes increases in the mean, in the magnitude of the interannual variability (σ_s), in the amplitude of the mean annual cycle, in Q_5 and in Q_{95}. As a result, ranking the skill of the models to simulate the mean and the annual cycle changes significantly when one high-quality forcing data set is replaced by another one.

8 Discussion and conclusions

8.1 Model evaluation

The first aim of this paper was to carry out an extensive evaluation and intercomparison of five continental-scale hydrological models for Europe. We found considerable inter-model differences in mean discharge and in the simulations of the annual discharge cycle, of low flow and of high flow. There is less spread among the models in their skill to simulate interannual variability, e.g. ⟨CRMSEann⟩ ranges between 24 and 29 $\mathrm{mm\,yr^{-1}}$ for the E-OBS simulation. We explain this by the dominant effect of interannual variation in precipitation on interannual variation in discharge, keeping in mind that all models used the same precipitation forcing (with the exception of a precipitation correction in E-HYPE). Apparently, the large diversity between the models, for instance in their methods of computing evapotranspiration and snow processes, only have a minor effect on interannual variability.

The performance of the models to simulate the annual cycle, and low and high flow is related to the way the models have been calibrated. The most extensive calibrations were carried out for Lisflood and E-HYPE, though with strategies that differed substantially. Indeed, these two models generally produced the best simulations of the mentioned aspects of the hydrograph. LPJmL and WBM, which can be considered as uncalibrated, generally show less skill in simulating the annual cycle, low and high flow. VIC takes an intermediate position both in terms of the complexity of the calibration procedure and in terms of the quality of the simulations. So, more extensively calibration is correlated with better simulations of the mentioned aspects of the hydrograph. This relation is expected but it should be noted that, of course, the description of the physical processes in the models also affects the simulations.

8.2 Sensitivity to meteorological forcing

By replacing the E-OBS forcing with WFDEI forcing we demonstrated that the relative skill of the models to simulate the mean and the interannual variability is very sensitive to the meteorological forcing. For the case of E-OBS and WFDEI we showed that this sensitivity is caused by the differences in precipitation between the two forcings, WFDEI being higher by 104 $\mathrm{mm\,yr^{-1}}$ on average across Europe, rather than by the differences in temperature. Our study confirms the large sensitivity of modelled discharge to the uncertainty in the precipitation input found in previous studies (e.g. Fekete et al., 2004; Biemans et al., 2009). The large sensitivity to the forcing also underlines that good or bad skill scores often reflect, at least partly, the proximity of the forcing to the forcing used during calibration.

We further found that the difference between E-OBS and WFDEI, causing the largest part of the differences in the simulated discharge at the pan-European scale, is the absence (presence) of a correction for undercatch in the E-OBS (WFDEI) precipitation data. This leads to higher mean and more interannual variability in the discharge of the WFDEI simulations as compared to the E-OBS simulations. We found that compared to observed discharge, all models but one have too little mean discharge and too little interannual variability in discharge when forced with E-OBS data whereas all models but one have too much mean discharge and too much interannual variability in discharge when forced with WFDEI data. We interpret this as an indication that E-OBS underestimates precipitation and that WFDEI overestimates precipitation. This in turn would indicate that making an undercatch correction is justified but that the correction applied to the WFDEI data is too large. Note that this reasoning only provides an *indication* about the correctness of precipitation in the two data sets since models participating in this study could be biased in similar ways, e.g. by having parameter values that compensate the undercatch problem. Arguments supporting the indication though are the fact that models have been calibrated in very different and independent ways, and that the tendencies appear both in the mean and in the magnitude of the interannual variability.

8.3 Suitability for making projections under climate change

The second aim of this paper was to use results of the evaluation to assess the suitability of hydrological models for making climate change impact projections. We propose that the skill of a hydrological model to simulate interannual variability in discharge is a good measure for the hydrological model's ability to make projections of climate changes in mean discharge. Since the skill of simulating interannual variability does not differ very much between the models, projections made by each of the five hydrological models are of about similar quality.

The argument for analysing interannual variability in discharge to assess each model's uncertainty in projected discharge changes due to climate change is that differences between years form an analogue for climate change. This approach is based on several assumptions. The first is that it is implicitly assumed that ⟨CRMSEann⟩ values derived from annual values during the evaluation period also apply when the climate change exceeds the interannual variability during the evaluation period. Secondly, we assume that year-to-year persistence of annual discharge is negligible compared to the interannual variability. Finally, the simulations done for the present study did not consider several processes that will affect changes in the water cycle at the temporal scale of climate change projections. These processes include the effects of variations in CO_2 on incoming long-wave radiation, and on evapotranspiration by changing the leaf area index and by changing stomatal conductance. So, the third assumption is that these CO_2-induced changes are small compared to the changes induced by changes in the meteorological variables considered in this study.

Despite the relatively small inter-model range in ⟨CRMSEann⟩, the question arises how values of ⟨CRMSEann⟩ could be used to assign weights to projections made by hydrological models. Such projections are indeed being produced by the models participating in the present study within the framework of the EU-project IMPACT2C. Since this minimizes the variance of the weighted average, weights should be proportional to the inverse of the squared error (Hartung et al., 2008). Applying this rule to the ⟨CRMSEann⟩, weights vary between 17 % for LPJmL and 24 % for Lisflood using the results from the E-OBS simulation. We suggest that the data used as reference for the bias corrections of the climate simulations should determine the choice of the set of weights, e.g. from the E-OBS or from the WFDEI simulation. The motivation for this is that most bias correction methods aim at correcting the climate model output so that the model reproduces the observed variability during the reference period. Since the entire projection is bias-corrected with the same s, the climate change signal is scaled on the basis of the observations used as reference for the corrections. Hence, it is the quality of the simulation forced by the reference observations that predicts the accuracy of the projected change.

Acknowledgements. This study was funded by the EU through the FP7 project IMPACT2C (Grant agreement no: 282746). We would like to thank Mark Savenije for providing the EOBS data in the desired format and Evert-Jan Bakker for giving advice about statistical issues.

References

Adler, R. F., Huffman, G. J., Chang, A., Ferraro, R., Xie, P. P., Janowiak, J., and Nelkin, E.: The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979–present), J. Hydrometeorol., 4, 1147–1167, 2003.

Biemans, H., Hutjes, R. W. A., Kabat, P., Strengers, B. J., Gerten, D., and Rost, S.: Effects of precipitation uncertainty on discharge calculations for main river basins, J. Hydrometeorol., 10, 1011–1025, 2009.

Burek, P., van der Knijff, J., and de Roo, A.: LISFLOOD Distributed Water Balance and Flood Simulation Model, revised user manual, JRC technical reports EUR 22166 EN/3 EN, European Union, printed in Italy, 2013a.

Burek, P., van der Knijff, J., and Ntegeka, V.: LISFLOOD Distributed Water Balance and Flood Simulation Model, revised user manual, JRC technical reports EUR 26167 EN, European Union, printed in Italy, 2013b.

Dai, A., Qian, T., Trenberth, K. E., and Milliman, J. D.: Changes in continental freshwater discharge from 1948 to 2004, J. Climate, 22, 2773–2792, 2009.

Déqué, M. and Somot, S.: Weighted frequency distributions express modelling uncertainties in the ENSEMBLES regional climate experiments, Clim. Res., 44, 195–201, 2010.

Donnelly, C., Rosberg, J., and Isberg, K.: A validation of river routing networks for catchment modelling from small to large scales, Hydrol. Res., 44, 917–925, 2012.

Donnelly, C., Andersson, J. C., and Arheimer, B.: Using flow signatures and catchment similarities to evaluate the E-HYPE multi-basin model across Europe, Hydrolog. Sci. J., accepted, 2015.

Fekete, B. M., Vörösmarty, C. J., Roads, J. O., and Willmott, C. J.: Uncertainties in precipitation and their impacts on runoff estimates, J. Climate, 17, 294–304, 2004.

Gudmundsson, L., Tallaksen, L. M., Stahl, K., Clark, D. B., Dumont, E., Hagemann, S., Bertrand, N., Gerten, D., Heinke, J., Hanasaki, N., Voss, F., and Koirala, S.: Comparing Large-Scale Hydrological Model Simulations to Observed Runoff Percentiles in Europe, J. Hydrometeorol., 13, 604–620, 2011.

Gudmundsson, L., Wagener, T., Tallaksen, L. M., and Engeland, K.: Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in Europe, Water Resour. Res., 48, 1–20, doi:10.1029/2011WR010911, 2012.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling, J. Hydrol., 377, 80–91, 2009.

Hagemann, S., Chen, C., Clark, D. B., Folwell, S., Gosling, S. N., Haddeland, I., Hanasaki, N., Heinke, J., Ludwig, F., Voss, F., and Wiltshire, A. J.: Climate change impact on available water resources obtained using multiple global climate and hydrology models, Earth Syst. Dynam., 4, 129–144, doi:10.5194/esd-4-129-2013, 2013.

Hamon, W. R.: Estimating Potential Evapotranspiration, PhD thesis, Massachusetts Institute of Technology, Boston, 1960.

Hargreaves, G. H. and Samani, Z. A.: Reference crop evapotranspiration from ambient air temperature, no. fiche no. 85-2517, (Microfiche collection), American Society of Agricultural Engineers, USA, 1985.

Hartung, J., Knapp, G., and Sinha, B. K.: Statistical Meta-Analysis With Applications, John Wiley and Sons, Hoboken, New Jersey, 2008.

Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D., and New, M.: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006, J. Geophys. Res.-Atmos., 113, 1–12, doi:10.1029/2008JD010201, 2008.

Hesselbjerg Christensen, J., Kjellström, E., Giorgi, F., Lenderink, G., and Rummukainen, M.: Weight assignment in regional climate models, Clim. Res., 44, 179–194, 2010.

Hofstra, N., Haylock, M., New, M., and Jones, P. D.: Testing E-OBS European high-resolution gridded data set of daily precipitation and surface temperature, J. Geophys. Res.-Atmos., 114, 1–16, doi:10.1029/2009JD01179, 2009.

Liang, X., Lettenmaier, D. P., Wood, E. F., and Burges, S. J.: A simple hydrologically based model of land surface water and energy fluxes for general circulation models, J. Geophys. Res.-Atmos., 99, 14415–14428, 1994.

Lindström, G., Pers, C., Rosberg, J., Strömqvist, J., and Arheimer, B.: Development and testing of the HYPE (Hydrological Predictions for the Environment) water quality model for different spatial scales, Hydrol. Res., 41, 295–319, 2010.

Lohmann, D., Nolte-Holube, R., and Raschke, E.: A large-scale horizontal routing model to be coupled to land surface parametrization schemes, Tellus A, 48, 708–721, 1996.

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

Nash, J. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I – a discussion of principles, J. Hydrol., 10, 282–290, 1970.

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

Ntegeka, V., Salamon, P., Gomes, G., Sint, H., Lorini, V., Zambrano-Bigiarini, M., and Thielen, J.: EFAS-Meteo: a European Daily High-Resolution Gridded Meteorological Data Set for 1990–2011, JRC technical report EUR 26408 EN, European Union, printed in Italy, 2013.

Priestley, C. H. B. and Taylor, R. J.: On the assessment of surface heat flux and evaporation using large-scale parameters, Mon. Weather Rev., 100, 81–92, 1972.

Prudhomme, C., Parry, S., Hannaford, J., and Clark, D.: How well do large-scale models reproduce regional hydrological extremes in Europe?, J. Hydrometeorol., 12, 1181–1204, 2011.

Rost, S., Gerten, D., Bondeau, A., Lucht, W., Rohwer, J., and Schaphoff, S.: Agricultural green and blue water consumption and its influence on the global water system, Water Resour. Res., 44, W09405, doi:10.1029/2007WR006331, 2008.

Roudier P., Andersson, J., Donnelly, C., Feyen, L., Gerten, W., Greuell, D., Ludwig, F., Pisacane, G., Roudier, P., and Schaphoff, S.: Projections of future floods and hydrological droughts in Europe under a +2 °C global warming, Climatic Change, submitted, 2015.

Schaphoff, S., Heyder, U., Ostberg, S., Gerten, D., Heinke, J., and Lucht, W.: Contribution of permafrost soils to the global carbon budget, Environ. Res. Lett., 8, 014026, doi:10.1088/1748-9326/8/1/014026, 2013.

Schewe, J., Heinke, J., Gerten, D., Haddeland, I., Arnell, N. W., Clark, D. B., and Kabat, P.: Multimodel assessment of water scarcity under climate change, P. Natl. Acad. Sci. USA, 111, 3245–3250, 2014.

Shuttleworth, W. J.: Evaporation, in: chapter 4, Handbook of Hydrology, edited by: Maidment, D. R., McGraw-Hill, New York, 1993.

Stahl, K., Tallaksen, L. M., Gudmundsson, L., and Christensen, J. H.: Streamflow data from small basins: a challenging test to high-resolution regional climate modeling, J. Hydrometeorol., 12, 900–912, 2011.

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res.-Atmos., 106, 7183–7192, 2001.

van Vliet, M., Donnelly, C., Strömbäck, L., and Capell, R.: European scale climate information services for water use sectors, J. Hydrol., in revision, 2015.

Weedon, G. P., Gomes, S., Viterbo, P., Shuttleworth, W. J., Blyth, E., Österle, H., and Best, M.: Creation of the WATCH forcing data and its use to assess global and regional reference crop evaporation over land during the twentieth century, J. Hydrometeorol., 12, 823–848, 2011.

Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., and Viterbo, P.: The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data, Water Resour. Res., 50, 7505–7514, 2014.

Willmott, C. J., Rowe, C. M., and Mintz, Y.: Climatology of the terrestrial seasonal water cycle, J. Climatol., 5, 589–606, 1985.

Wisser, D., Fekete, B. M., Vörösmarty, C. J., and Schumann, A. H.: Reconstructing 20th century global hydrography: a contribution to the Global Terrestrial Network – Hydrology (GTN-H), Hydrol. Earth Syst. Sci., 14, 1–24, doi:10.5194/hess-14-1-2010, 2010.

Xia, Y., Mitchell, K., Ek, M., Cosgrove, B., Sheffield, J., Luo, L., and Lohmann, D.: Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System project phase 2 (NLDAS-2): 2. validation of model-simulated streamflow, J. Geophys. Res.-Atmos., 117, 1–27, doi:10.1029/2011JD01604, 2012.

Table 1. Main characteristics of the hydrological models and the simulations carried out for the present study. Names of forcing variables are abbreviated as follows: p = precipitation, T_{min} = daily minimum temperature, T_{max} = daily maximum temperature, T = temperature, R_{n} = net radiative flux, S_{in} = incoming flux of shortwave radiation, L_{in} = incoming flux of long-wave radiation, wsp = wind speed and q = specific humidity.

Model (acronym)	E-HYPE	Lisflood	LPJmL	VIC	WBM
Full name	European Hydrological Predictions for the Environment	LISFLOOD	Lund–Potsdam–Jena managed Land	Variable Infiltration Capacity model	Water Balance Model
version	2.5	version of 2013	3.5	4.2.1.g	WBMplus
Time step	daily	daily	daily	3 hourly	daily
Spatial resolution	sub-basins with median size of 215 km^2	5 km	0.5°	0.5°	0.1°
Forcing variables	p, T_{min}, T_{max}, T	p, T_{min}, T_{max}	p, T, R_{n}	p, T_{min}, T_{max}, S_{in}, L_{in}, wsp, q	p, T
Evapotranspiration	Hargreaves-Samani (1985)	Hargreaves-Samani (1985)	Priestley–Taylor (1972)	Penman–Monteith (Shuttleworth, 1993)	Hamon (1960)
Snow melt	degree days	degree days with factor depending on p and season	degree days with factor depending on p; energy balance for permafrost	energy balance	function of T and p (Willmott et al., 1985)
Soil layers	up to 3	2	6	3	1
Surface runoff	function of exceedance of field capacity	within cell variable infiltration capacity	saturation excess per layer	within cell variable infiltration capacity	saturation excess
Routing	within sub-basins and between sub-basins based on Hydrosheds; weir equation for lakes	kinematic wave equations	daily volume transmission following DDM30 network	Lohmann et al. (1996)	Muskingum–Cunge
Calibration	Donnelly et al. (2015)	unpublished	not	Nijssen et al. (2001)	not
References	Lindström et al. (2010) and Donnelly et al. (2015)	Burek et al. (2013a, b)	Rost et al. (2008), Schaphoff et al. (2013)	Liang et al. (1994)	Wisser et al. (2010)

Table 2. List of observation stations used for the present study. Fen. Scan. is the abbreviation for Fennoscandia. The area is the area of the catchment upstream from the station. Mean dis. is mean discharge for 1979–2000. Data sources are described in Sect. 3.2. The column reg. contains a "no" for rivers that are hardly affected by reservoir regulation. The observations from these rivers are used for all aspects of validation considered in this paper. Discharge of rivers with "yes" in the "Reg" column is believed to be substantially affected by human interference. Data from these rivers are only used for the validation of the mean. Nr corresponds to the number in the map (Fig. 2).

River	Station	lon. (° E)	lat. (° N)	Region	Area (km^2)	Mean dis. (mm yr^{-1})	Data source	Reg.	Nr.
Danube	Svistov	25.35	43.63	Central Europe	658 340	279	GRDC	no	42
Vistula	Tczew	18.80	54.09	Central Europe	194 000	171	Dai	no	26
Rhine	Lobith	6.11	51.84	Central Europe	160 800	466	GRDC	no	31
Elbe	Neu Darchau	10.89	53.23	Central Europe	131 950	170	GRDC	yes	27
Loire	Montjean	–0.83	47.38	Central Europe	110 000	262	Dai	no	39
Oder	Hohensaaten	14.14	52.86	Central Europe	109 564	149	GRDC	no	29
Rhone	Beaucaire	4.64	43.81	Central Europe	95 590	576	Dai	no	41
Neman	Smalininkai	22.58	55.08	Central Europe	81 200	207	GRDC	no	25
Weser	Intschede	9.13	52.96	Central Europe	37 720	285	GRDC	no	28
Main	Frankfurt am Main	8.67	50.11	Central Europe	24 764	257	GRDC	no	33
Neris	Jonava	24.28	55.08	Central Europe	24 500	223	GRDC	no	24
Morava	Moravsky	16.94	48.60	Central Europe	24 129	135	GRDC	no	35
Moselle	Trier	6.62	49.73	Central Europe	23 857	414	GRDC	no	34
Havel	Rathenow	12.32	52.61	Central Europe	19 288	137	GRDC	yes	30
Vah	Sala	17.88	48.16	Central Europe	11 218	385	GRDC	no	36
Thames	Kingston	–0.31	51.41	Central Europe	9948	192	GRDC	no	32
Kemijoki	Isohaara	24.55	65.78	North Fen.Scan.	50 686	363	GRDC	yes	4
Torneälven	Kukkolankoski	24.06	65.98	North Fen.Scan.	33 930	392	GRDC	no	3
Ångermanälven	Sollefteå krv	17.27	63.17	North Fen.Scan.	30 638	551	GRDC	yes	10
Umeaälven	Stornorrfors krv	20.05	63.85	North Fen.Scan.	26 568	556	GRDC	yes	9
Indalsälven	Bergeforsens krv	17.39	62.52	North Fen.Scan.	25 761	578	GRDC	yes	12
Kalixälven	Räktfors	22.82	66.17	North Fen.Scan.	23 103	416	GRDC	no	2
Oulujoki	Meriskoski	25.52	65.02	North Fen.Scan.	22 841	369	GRDC	yes	7
Leppävesi	Vaajakoski	25.88	62.23	North Fen.Scan.	17 684	286	GRDC	yes	14
Kallavesi	Kallavesi–Konnus–Karvio	27.77	62.53	North Fen.Scan.	16 270	324	GRDC	no	11
Iijoki	Raasakka	25.43	65.32	North Fen.Scan.	14 191	389	GRDC	no	6
Tana	Polmak	28.02	70.07	North Fen.Scan.	14 160	399	GRDC	no	1
Ljungan	Skallböle krv	16.96	62.36	North Fen.Scan.	12 088	354	GRDC	yes	13
Skellefteälven	Kvistforsens krv	20.86	64.74	North Fen.Scan.	11 309	471	GRDC	yes	8
Piteälven	Sikfors krv	21.21	65.53	North Fen.Scan.	10 816	508	GRDC	no	5
Vuoksi	Tainionkoski	28.78	61.22	South Fen.Scan.	61 061	326	GRDC	no	16
Göta älv	Vargöns krv	12.37	58.36	South Fen.Scan.	46 886	368	GRDC	yes	23
Glomma	Solbergfoss	11.15	59.64	South Fen.Scan.	40 540	551	GRDC	no	21
Kymijoki	Anjala	26.82	60.70	South Fen.Scan.	36 275	270	GRDC	no	19
Kokemaenjoki	Harjavalta	22.07	61.20	South Fen.Scan.	26 117	303	GRDC	yes	18
Ljusnan	Ljusne strömmar krv	17.08	61.21	South Fen.Scan.	19 817	376	GRDC	yes	17
Dramselv	Dovikfoss	9.91	59.88	South Fen.Scan.	16 120	575	GRDC	no	20
Motala ström	Holmen	16.17	58.59	South Fen.Scan.	15 384	186	GRDC	yes	22
Losna	Losna	10.28	61.33	South Fen.Scan.	11 210	711	GRDC	no	15
Rhine	Rheinfelden	7.78	47.56	Alps	34 550	980	GRDC	yes	38
Inn	Wasserburg	12.23	48.06	Alps	11 983	948	GRDC	no	37
Rhone	Chancy	5.97	46.15	Alps	10 299	1126	GRDC	no	40
Ebro	Tortosa	0.50	40.82	Mediterranean	84 230	124	Dai	yes	44
Tagus	Almourol	–8.37	39.47	Mediterranean	67 490	142	Dai	yes	45
Guadalquivir	Alcaladelrio	–5.98	37.52	Mediterranean	46 995	55	Dai	yes	46
Minho	Fozdomour	–8.38	42.08	Mediterranean	15 457	523	Dai	yes	43

Table 3. Evaluation statistics for each model. The Centred Root Mean Square Error of annual values ⟨CRMSEann⟩ and the ratio of the interannual variability in the simulations and the measurements $\langle \sigma_s/\sigma_m \rangle$ are defined in Sect. 5. The correlation coefficient of σ_s vs. σ_m quantifies how well a model simulates inter-basin differences in the interannual variability.

	forcing	E-HYPE	Lisflood	LPJmL	VIC	WBM
		Bias (mm yr^{-1})				
Mean discharge	E-OBS	−3	−17	−29	+23	−117
	WFDEI	+63	+68	+39	+107	−47
Low discharge (Q_5)	E-OBS	−12	−3	−73	−32	−65
	WFDEI	0	+25	−63	−13	−55
High discharge (Q_{95})	E-OBS	+12	−34	+116	+89	−111
	WFDEI	+140	+102	+243	+240	+31
		Bias (%)				
Mean discharge	E-OBS	−1	−6	−9	+8	−38
	WFDEI	+20	+22	+13	+34	−15
Low discharge (Q_5)	E-OBS	−11	−3	−68	−30	−61
	WFDEI	0	+23	−59	−12	−51
High discharge (Q_{95})	E-OBS	+2	−7	+23	+17	−21
	WFDEI	+27	+20	+47	+46	+6
		Interannual variability				
⟨CRMSEann⟩ (mm yr^{-1})	E-OBS	28	24	29	25	28
	WFDEI	29	23	31	30	29
$\langle \sigma_s/\sigma_m \rangle$	E-OBS	0.92	0.75	0.89	1.09	0.81
	WFDEI	1.15	0.96	1.11	1.32	1.15
corr. coeff. σ_s vs. σ_m	E-OBS	0.93	0.96	0.94	0.92	0.91

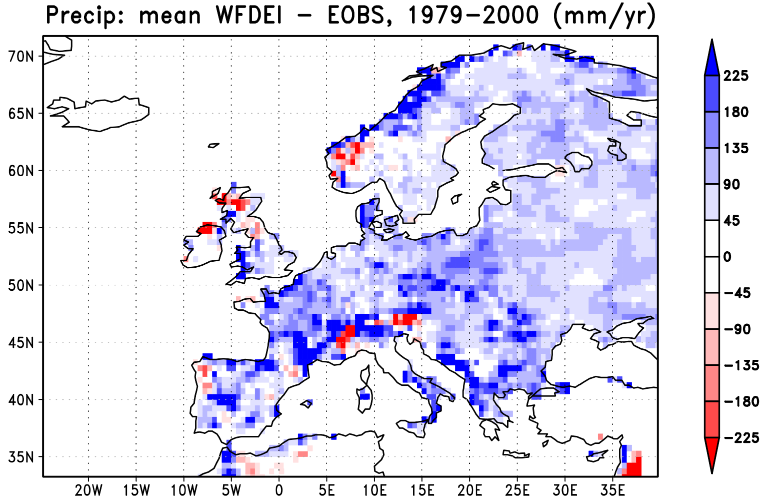

Figure 1. Difference in 1979–2000 mean annual precipitation rate between the WFDEI (GPCC version) and the E-OBS data. For white cells data are missing in one or both of the data sets.

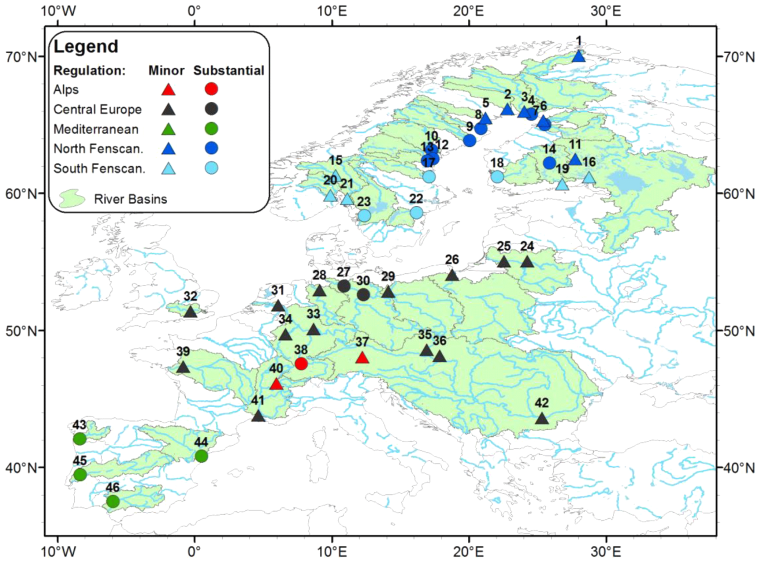

Figure 2. Map showing the stations from which discharge observations have been used in this study. Sorting of the stations and their catchments by geographic region and degree of regulation is shown by colour and shape of the symbols. More information about the stations and the basins can be found in Table 2.

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
◀| |▶
◀ ▶
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

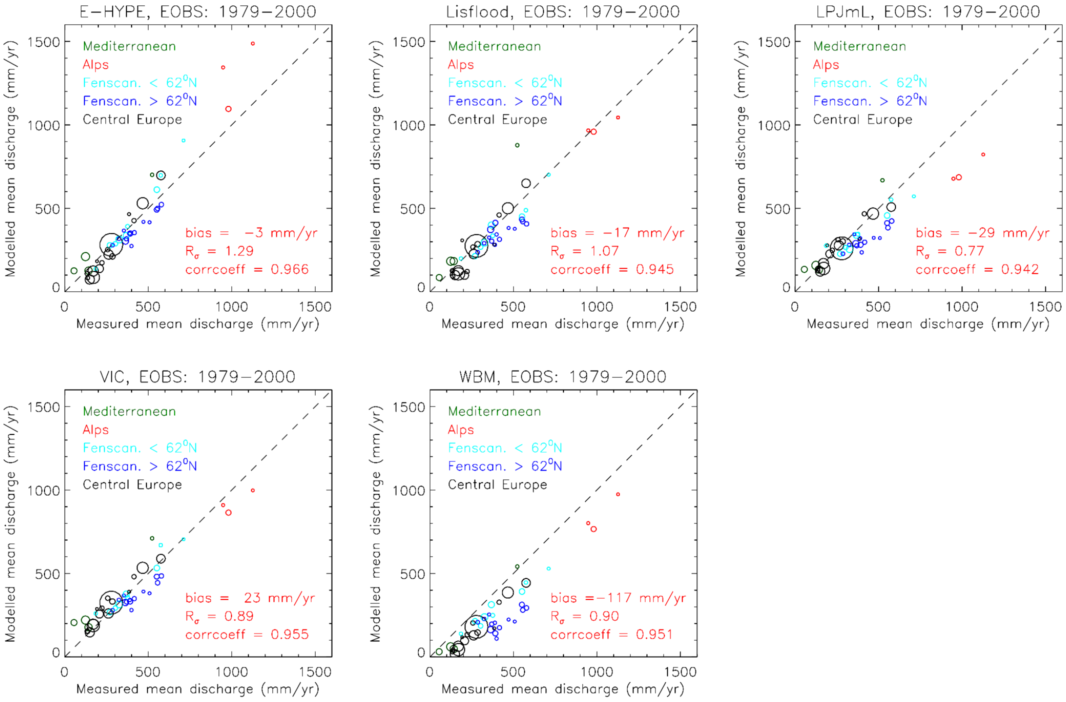

Figure 3. Modelled vs. measured 22 year mean discharge. Model calculations are from the E-OBS simulation. Each circle represents a discharge station with the area of the circle proportional to the area of the catchment upstream from the station. R_σ is the ratio of the standard deviations in the x and the y direction.

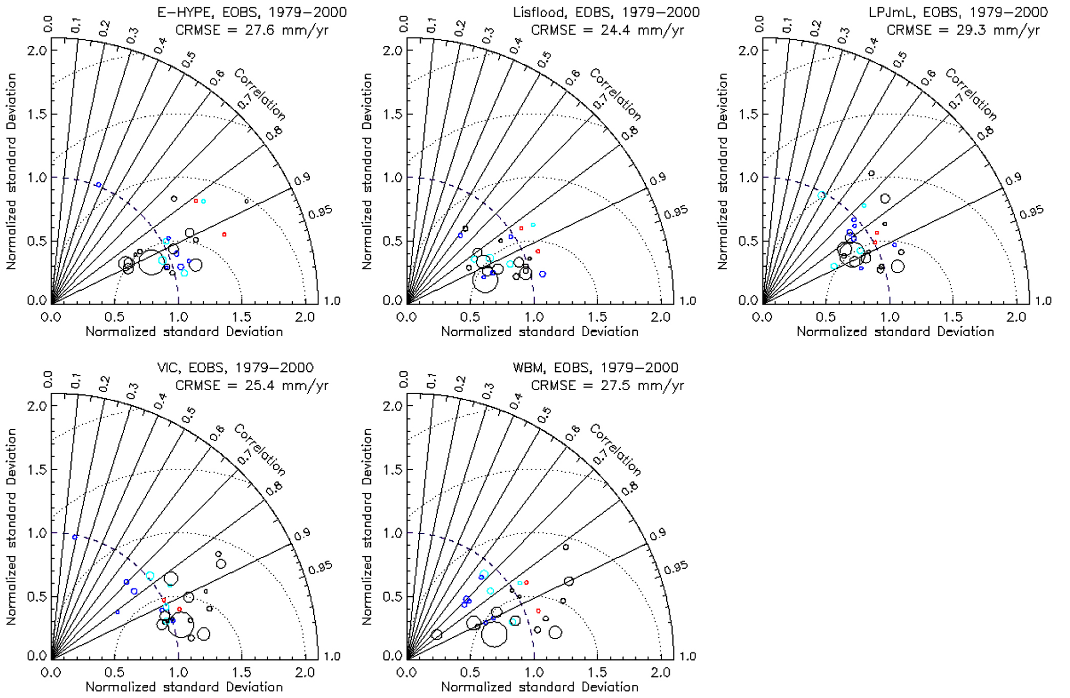

Figure 4. Taylor diagrams based on annual discharge values. Model calculations are from the E-OBS simulation. Each circle represents a discharge station, with the area of the circle proportional to the area of the catchment upstream from the station. Color codes are as in Fig. 3. The ratio of the standard deviations of simulated and measured annual discharge (σ_s/σ_m) is plotted in the radial direction, the correlation coefficient between simulated and measured annual discharge is plotted in the angular direction. The CRMSEann of each station is proportional to the distance of its circle to the point of perfect simulation ($\sigma_s/\sigma_m = 1$; corr. coeff. = 1). The stippled circles show isolines of CRMSEann multiplied by σ_m.

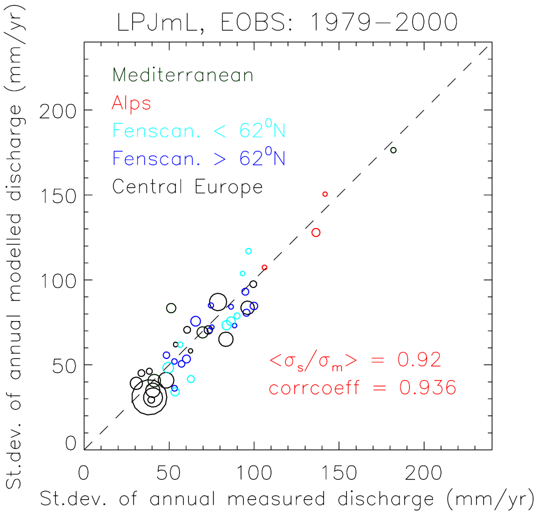

Figure 5. As Fig. 3 but for the standard deviation of annual simulated (σ_s) vs. the standard deviation of annual measured (σ_m) river discharge for LPJmL.

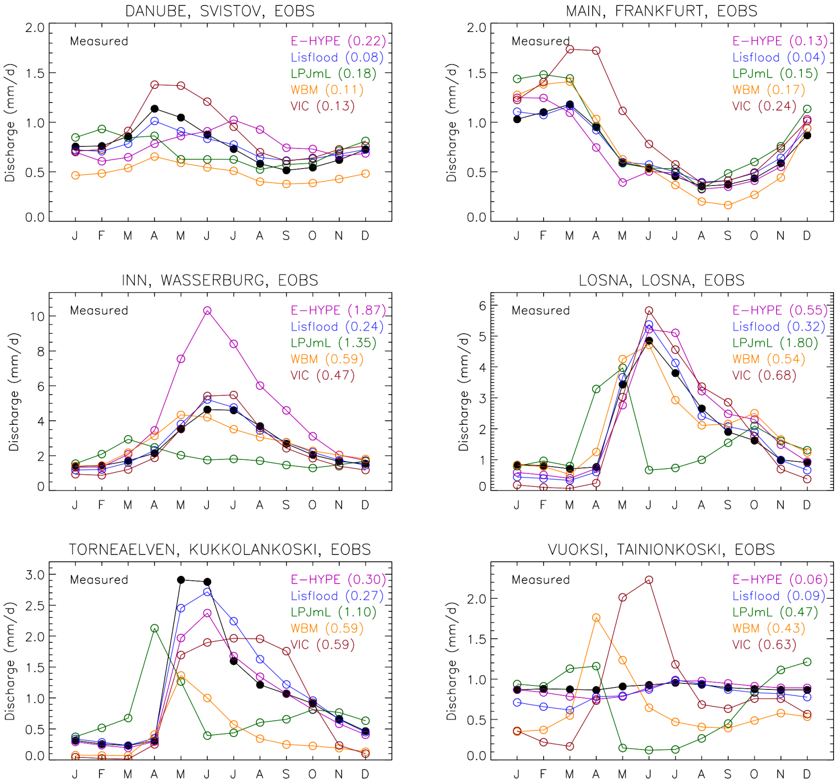

Figure 6. Comparison of the simulated 22 year-mean annual cycle of discharge with observations from six stations. Model calculations are from the E-OBS simulation. The skill of the simulations to reproduce the annual cycle is quantified by the CRMSEmth, which is given between parentheses (in mm day^{-1}).

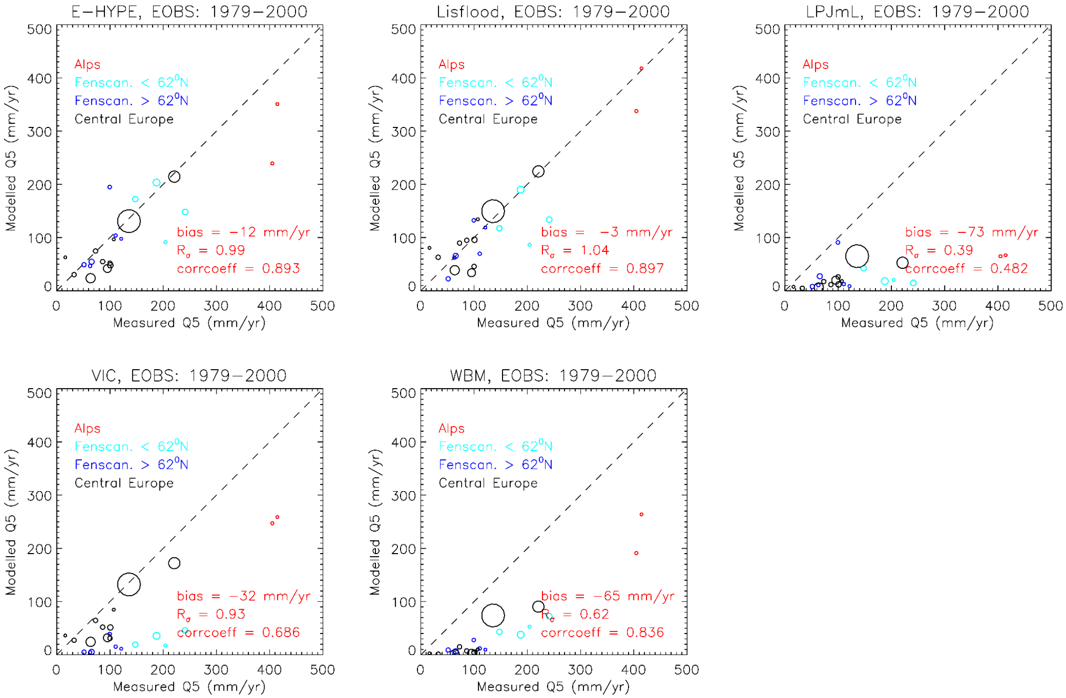

Figure 7. As Fig. 3 but for the 5th percentile of all daily discharge values.

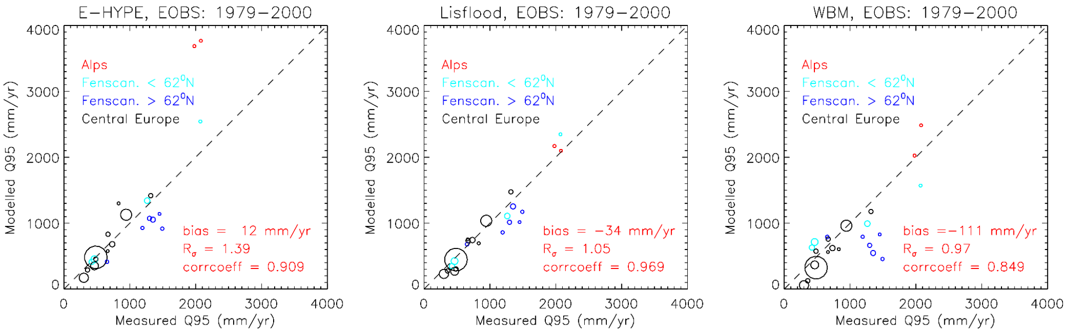

Figure 8. As Fig. 3 but for the 95th percentile of all daily discharge values.

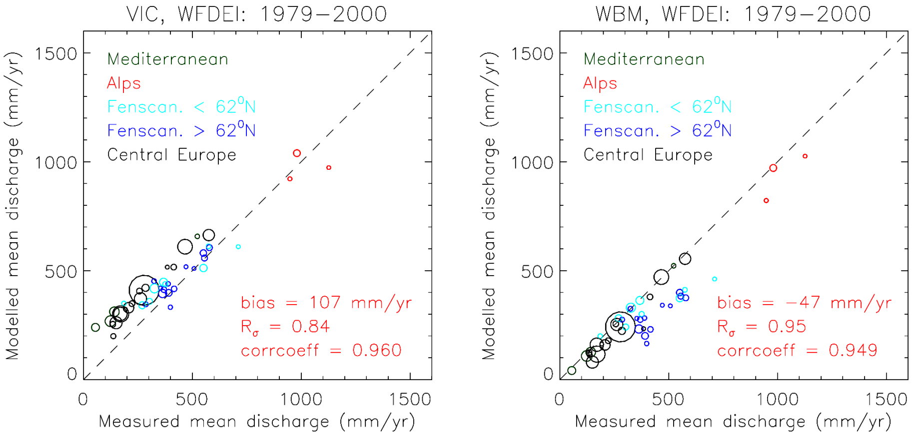

Figure 9. As Fig. 3 but for the WFDEI simulation instead of the E-OBS simulation and for a selection of two models.

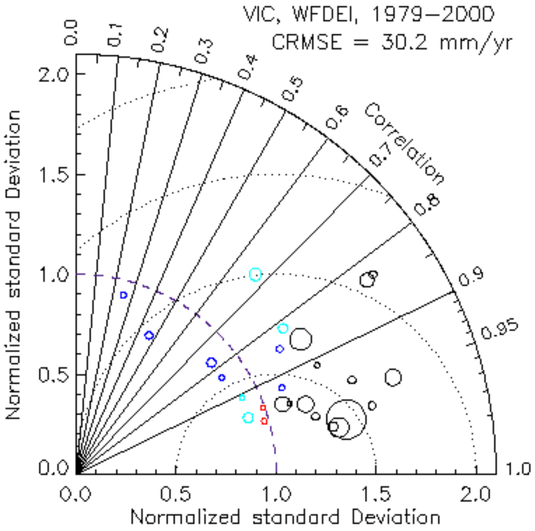

Figure 10. As Fig. 4 but for the WFDEI simulation instead of the E-OBS simulation and for VIC only.

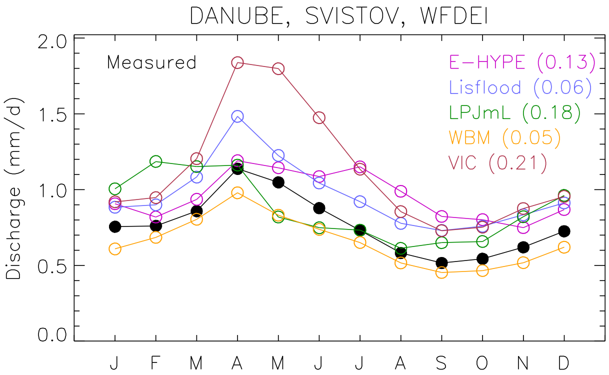

Figure 11. As Fig. 6 but for the WFDEI simulation instead of the E-OBS simulation for one exemplary gauge.