Hydrol. Earth Syst. Sci. Discuss., 4, 4125–4173, 2007
www.hydrol-earth-syst-sci-discuss.net/4/4125/2007/

Hydrology and Earth System Sciences Discussions

Papers published in *Hydrology and Earth System Sciences Discussions* are under open-access review for the journal *Hydrology and Earth System Sciences*

Value of river discharge data for global-scale hydrological modeling

M. Hunger and P. Döll

Institute of Physical Geography, University of Frankfurt, Frankfurt am Main, Germany

Received: 29 October 2007 – Accepted: 30 October 2007 – Published: 15 November 2007

Correspondence to: M. Hunger (m.hunger@em.uni-frankfurt.de)

Abstract

This paper investigates the value of observed river discharge data for global-scale hydrological modeling of a number of flow characteristics that are required for assessing water resources, flood risk and habitat alteration of aqueous ecosystems. An improved version of WGHM (WaterGAP Global Hydrology Model) was tuned in a way that simulated and observed long-term average river discharges at each station become equal, using either the 724-station dataset (V1) against which former model versions were tuned or a new dataset (V2) of 1235 stations and often longer time series. WGHM is tuned by adjusting one model parameter (γ) that affects runoff generation from land areas, and, where necessary, by applying one or two correction factors, which correct the total runoff in a sub-basin (areal correction factor) or the discharge at the station (station correction factor). The study results are as follows. (1) Comparing V2 to V1, the global land area covered by tuning basins increases by 5%, while the area where the model can be tuned by only adjusting γ increases by 8% (546 vs. 384 stations). However, the area where a station correction factor (and not only an areal correction factor) has to be applied more than doubles (389 vs. 93 basins), which is a strong drawback as use of a station correction factor makes discharge discontinuous at the gauge and inconsistent with runoff in the basin. (2) The value of additional discharge information for representing the spatial distribution of long-term average discharge (and thus renewable water resources) with WGHM is high, particularly for river basins outside of the V1 tuning area and for basins where the average sub-basin area has decreased by at least 50% in V2 as compared to V1. For these basins, simulated long-term average discharge would differ from the observed one by a factor of, on average, 1.8 and 1.3, respectively, if the additional discharge information were not used for tuning. The value tends to be higher in semi-arid and snow-dominated regions where hydrological models are less reliable than in humid areas. The deviation of the other simulated flow characteristics (e.g. low flow, inter-annual variability and seasonality) from the observed values also decreases significantly, but this is mainly due to the better representation

of average discharge but not of variability. (3) The optimal sub-basin size for tuning depends on the modeling purpose. On the one hand, small basins between 9000 and 20 000 km^2 show a much stronger improvement in model performance due to tuning than the larger basins, which is related to the lower model performance (with and without tuning), with basins over 60 000 km^2 performing best. On the other hand, tuning of small basins decreases model consistency, as almost half of them require a station correction factor.

1 Introduction

Hydrological models suffer from uncertainties with regard to model structure, input data (in particular precipitation) and model parameters. In catchment studies, time series of observed river discharge are widely used to adjust model parameters such that a satisfactory fit of modeled and observed river discharge is obtained. Parameter adjustment, i.e. model calibration or tuning, leads to a reduction of model uncertainty by including the aggregated information about catchment processes that is provided by observed river discharge. River discharge is a unique hydrological variable as it is the final outcome of a large number of (vertical and horizontal) flow and transfer processes within the whole catchment of the discharge observation point. River discharge measured at one location therefore reflects system inflows (like precipitation), outflows (like evapotranspiration) and water storage changes (e.g. in lakes and groundwater) throughout the whole upstream area. Measurements of all other hydrological variables, e.g. evapotranspiration and groundwater recharge, at any one location reflect only local processes, and a large number of observations of these quantities within a catchment would be necessary for characterizing the overall water balance of the catchment. Discharge observations are available for many rivers of the world. Measurement errors are considered to be small (except in the case of floods) as compared to the errors in areal precipitation estimation where interpolation errors add to measurement errors (Moody and Troutman, 1992; Hagemann and Dümenil, 1998; Adam and Lettenmeier, 2003).

Even though the value of discharge information is widely recognized in catchment-scale hydrological modeling, and thus models are calibrated against measured discharge to improve model performance, continental- or global-scale modeling of river discharge rarely makes use of river discharge observations. The low density of precipitation and other input data at these large scales, which increases model uncertainty, makes it imperative to take advantage of the integrative information provided by measured river discharge.

Land surface modules of climate models do not use river discharge data at all (except for validation), and the computed river discharge values are generally very different from observed values even when the models are driven by observed climate data (e.g. Oki et al., 1999). Döll et al. (2003) reviewed how river discharge information was taken into account by continental- and global-scale hydrological models. This ranges from no consideration at all in earlier years (Yates, 1997; Klepper and van Drecht, 1998) over global tuning of some model parameters (Arnell, 1999) to basin-specific tuning of parameters to measured river discharge. Within the latter group, the global WBM model was tuned to long-term average discharge at 663 stations not by adapting model parameters but by multiplying, in basins with observed discharge, model runoff by a correction factor which is equal to the ratio of observed and simulated long-term average discharge (Fekete et al., 2002). The only global models for which basin-specific tuning of parameters has been done are the VIC (Nijssen et al., 2001) and the WGHM (WaterGAP Global Hydrology Model) model (Döll et al., 2003).

Using time series of observed monthly river discharge at downstream stations of 22 large river basins world-wide, Nijssen et al. (2001) adjusted four VIC model parameters individually for each basin. Even after calibration, simulated long-term average discharges still showed an absolute deviation from the observed values between 1% and 22% for 17 out of the 22 basins. For the Senegal basin, VIC overestimated discharge by 340%, while for Brahmaputra, Irradwaddy, Columbia, and Yukon, deviations of 50–100% were not reduced due to obvious under- or overestimation of precipitation. Excluding those five basins, basin-specific tuning reduced the relative root-mean-

square error of the monthly flows from 62% to 37% and the mean bias in annual flows from 29% to 10%. Please note that in the version of VIC used by Nijssen et al. (2001), the impact of human water consumption on river discharge was not yet taken into account, which may explain the overestimation of 22% in the Yellow River. Haddeland et al. (2006) modeled the effect of irrigation and reservoirs on river discharge in VIC but did not recalibrate the model. Döll et al. (2003) used observed river discharge at 724 stations world-wide to force WGHM to model long-term average river discharge at these stations with a deviation of less than 1%. This provided a best estimate of renewable water resources. They adjusted one model parameter only but had to introduce, in many basins, two types of correction factors to achieve this goal, even though river discharge reduction due to human water consumption was taken into account. Döll et al. (2003) agreed with Nijssen et al. (2001) in their conclusion that two main reasons for the need of corrections factors are unrealistic precipitation data and problems in modeling important hydrological processes in semi-arid and arid areas. In these areas, evaporation from small ephemeral ponds, loss of river water to the subsurface, and river discharge reduction by irrigation are likely to influence the water balance strongly. In WGHM, only the latter is modeled albeit with a high uncertainty as, for example, modeled irrigation requirements may overestimate actual irrigation water consumption in case of water scarcity.

While global-scale information on precipitation has not become significantly more reliable during the last years, additional information on river discharge has been compiled by the Global Runoff Data Centre (GRDC) in Koblenz, Germany (http://grdc.bafg.de). New station data became available, and time series length for some of the old stations increased. In the most recent version of WGHM (WGHM 2.1.f), which also takes into account improved data on irrigation areas, we took advantage of this new information and used observed discharge at 1235 instead of 724 (in WGHM 2.1d, Döll et al., 2003) stations to tune the model. Almost all of the additional stations are located upstream of the WGHM 2.1d stations, i.e. zero-order river basins are now divided into smaller sub-basins than before (Fig. 1).

In this paper, we analyze the value of this additional discharge information for improved representation of observed river discharge by the global hydrological model WGHM. Obviously, long-term average discharge at the new stations will be represented better due to tuning, but to what extent is the simulation of other flow characteristics like inter-annual variability of annual flows, seasonality of flows and low flows improved both at the new stations and the respective downstream stations?

Besides, with more stations available, the question of optimal station density for tuning arises. Even though large areas of the globe still suffer from very limited discharge information (e.g. parts of Africa, Asia and South America) so that any additional information should be valuable, in other regions (e.g. in Europe and North America) available station density is high compared to the 0.5° by 0.5° spatial resolution of WGHM. On the one hand, if station density is chosen too coarse, existing spatial heterogeneities of the tuning parameters would remain unrepresented (Becker and Braun, 1999). On the other hand, larger sub-basins might be advantageous insofar as they hold a better chance for (model and data) errors to balance out. For example, gridded 0.5° precipitation used as model input (Mitchell and Jones, 2005) is based, for almost all areas on the globe, on much less than one station per grid cell, and the poor spatial resolution leads to increased errors of basin precipitation for smaller basins which might make it impossible even for the optimal model to simulate basin discharge correctly. Thus, with decreasing sub-basin size, we may expect that fewer sub-basins can be forced to simulate the observed long-term average discharge by only adjusting the model parameter, i.e. without using correction factors. At the same time, increased station density is expected to allow an improved modeling of downstream station discharge, as (long-term average) inflow into the downstream sub-basins is equal to observed values. A priori, it is not clear how these two effects balance.

To determine the value of integrating the additional river discharge information into WGHM, two variants of WGHM 2.1f were set up: V1, where WGHM 2.1f was tuned against the old 724-station dataset used for tuning WGHM 2.1d as described in Döll et al. (2003), and V2, where WGHM 2.1f was tuned against the new 1235-station dataset.

V2 represents the standard for WGHM 2.1f. Simulation results of model variants V1 and V2 are compared in order to answer the central questions of this study:

- Does increased river discharge information promote tuning WGHM by only one model parameter?
- To what extent does tuning against more discharge observations improve model performance?
- What is the impact of basin size on model performance and basin-specific tuning?

In the next section, we shortly present WGHM 2.1f, focusing on model improvements since WGHM 2.1d (Döll et al., 2003), and discuss the discharge data used for tuning. Besides, we describe the indicators of model performance that we used to assess the value of the additional river discharge information. In Sect. 3, we show the results of the comparison of the two model variants and answer the above research questions, while in Sect. 4, we draw conclusions.

2 Methods and data

2.1 Model description

WaterGAP (Döll et al., 1999; Alcamo et al., 2003) was developed to assess water resources and water use in river basins worldwide under the conditions of global change. The model, which has a spatial resolution of 0.5° geographical latitude by 0.5° geographical longitude, has been applied in a number of studies dealing with water scarcity and water stress (Smakhtin et al., 2004; Alcamo et al., 2007) and the impact of climate change on irrigation water requirements as well as on droughts and floods (Döll, 2002; Lehner et al., 2006). WaterGAP combines a global hydrological model with several global water use models, taking into account water consumption by households, industry, livestock and irrigation. It is driven by monthly 0.5° gridded climate data. WGHM,

the hydrological model of WaterGAP, is based on spatially distributed physiographic characteristics such as land cover, soil properties, hydrogeology and the location and area of reservoirs, lakes and wetlands. A daily water balance is calculated for each of the 66 896 grid cells, considering canopy, snow and soil water storages. Runoff generated within a cell contributes to river discharge after passing groundwater or surface water storages. River discharge of one grid cell integrates local inflow and inflow from upstream cells, taking into account reduction of discharge by human water consumption as computed by the WaterGAP water use models. Discharge is routed to the basin outlet in two-hour time steps through a river network derived from the global drainage direction map DDM 30 (Döll and Lehner, 2002). WGHM is tuned based on observed river discharge at stations around the world such that the tuning parameter is adjusted individually for each sub-basin (see Sect. 2.2). In untuned basins, the value of the tuning parameter is determined based on multiple regression, with long-term average temperature, fraction of surface water area and length of non-perennial rivers as predictor variables. Model results include monthly time series of surface runoff, groundwater recharge and river discharge. Compared to version 2.1d of WGHM described by Döll et al. (2003), the current version 2.1f comprises enhancements in several modules as well as updates for a number of input datasets.

Computation of river discharge reduction by human water consumption. All four water use model (domestic, industrial, irrigation, livestock) have been updated and provide time series of water withdrawal and water consumption from 1901 until 2002. Input data for the domestic water use model have been improved in particular for Europe (Flörke and Alcamo, 2004). The industrial water use model has been revised to distinguish water for cooling thermal power plants and manufacturing water use, as these two uses differ significantly in spatial distribution, driving forces and their consumption-to-withdrawal ratio (Vassolo and Döll, 2005). The current computation of irrigation water use includes an update of the "Global map of irrigation areas" (Siebert et al., 2005) that is the main model input. The map is based on the combination of up-to-date subnational irrigation statistics with geospatial information on the position and extent of

irrigation schemes. In river basins with extensive irrigation, changes in irrigation areas can be assumed to significantly influence river discharge.

The water required for consumptive water use is subtracted from river or lake storage. As water requirements cannot be satisfied in any cell at any time, WGHM permits to extract the unsatisfied portion from a neighboring cell. Before model version 2.1f, one neighboring cell, from which additional water could be extracted, was predefined for each cell. From the eight surrounding cells, the one with the highest long-term average discharge (1961–1990) was selected based on previous model tuning rounds. In WGHM 2.1f, the allocation is done dynamically during runtime at each time step to allow a more flexible fulfillment of demand. In case of a deficit in water supply for anthropogenic use, the model at each time step selects the neighboring cell with the highest actual water storage in rivers and lakes as donor cell. However, this dynamic allocation of water withdrawal from neighboring cells could not be implemented in the tuning run for technical reasons, and like in former model versions, the donor cell has to be determined based on the long-term average discharge as simulated by the untuned model. This restriction can lead to discrepancies between modeled and observed average discharge, particularly in very small basins where water use dominates the water balance.

Climate input and surface water data. Version 2.1f uses an updated set of climate information extracted from data of the Climate Research Unit (Mitchell and Jones, 2005). The new climate time series cover the time span from 1901 to 2002, extending the former data (1901 to 1995) by seven years. As in version 2.1d, precipitation data are not corrected for observational errors, which are expected to lead to an underestimation of precipitation by globally 11% and by up to 100% in snow-dominated areas (Legates and Willmott, 1990). GLWD, the Global Lake and Wetland Database (Lehner and Döll, 2004), provides information on freshwater bodies for WGHM. For version 2.1f, it has been supplemented by 64 additional reservoirs.

Snow modeling. In WGHM, snow accumulation and melting depends on daily temperatures that are derived from monthly data using cubic splines. Accumulation is

assumed to occur at temperatures below 0°C and melting above this value. In former versions, this resulted, in most grid cells, in one winter period where all precipitation fell as snow, and there was no melting at all. The snow balance simulation has been improved by refining the spatial resolution of the snow module (Schulze and Döll, 2004). In WGHM 2.1f, the snow water balance is computed no longer for the whole 0.5° grid cell but for 100 sub-grids per 0.5° cell, taking into account the effect of elevation (based on 30" elevation data) on temperature (−0.6°C/100 m). This provides a more differentiated temperature distribution within the 0.5° cells and allows for simultaneous snow accumulation and melting in one cell if the mean temperature is close to 0°C. The new snow algorithm resulted in an improved modeling of monthly river discharge in more than half of the 40 snow-dominated test basins, and the improvement was most significant in mountainous basins. Modeling efficiency of monthly river discharge in the 40 basins increased from 0.26 to 0.42 (Schulze and Döll, 2004).

Modeling of lakes and wetlands. Computation of the water balance of lakes and wetlands has been improved by making evaporation a function of water level (water storage), reflecting the dependence of surface area, from which evaporation occurs, on the amount of stored water. Please note that the lakes and wetlands taken into account in WGHM are based on maps, and their areas are likely to represent the maximum extent (Lehner and Döll, 2004). Like in former versions of WaterGAP, an active storage volume of 5 m and 2 m (multiplied by a constant lake or wetland area as available from maps) is assumed for lakes and wetlands, respectively, as there is a lack of data about lake and wetland water volume as a function of area available at the global scale (Döll et al., 2003). Outflow is modeled as a function of water storage. Wetlands, but not lakes, are assumed to disappear if storage is zero, with evaporation and outflow being zero, too.

In former versions, lake storage could vary between 5 m (then all inflow directly becomes outflow) and 0 m (then there is no outflow), but also reach very negative values, if the water balance is negative due to high evaporation and small inflows. Evaporation from lakes only depended on potential evaporation and the constant surface area, and

was thus likely to be overestimated in case of very low sea levels that go along with a decline of surface area. As a consequence, some lakes, particularly in semi-arid and arid regions, showed long-term downward trends of lake storage in former WGHM versions. In some cases, e.g. Lake Malawi, this precluded outflow from these lakes even for a number of relatively wet years.

To avoid this implausible behavior of lake storage dynamics in WGHM 2.1f, maximum evaporation is reduced as a function of lake storage level by multiplying it with a lake evaporation reduction factor r, which is computed as

$$r = 1 - \left(\frac{|S - S_{\max}|}{2 \cdot S_{\max}} \right)^{p} \quad (1)$$

with S actual lake storage [m^3], S_{max} maximum lake storage [m^3] and p reduction exponent [–]. Thus, evaporation reduction depends on actual lake storage. If S equals S_{max}, no reduction is applied, and if S equals $-S_{max}$, evaporation is reduced to zero. Therefore, lake storage cannot decline below $-S_{max}$. The exponent p is set to 3.32 such that evaporation is reduced by 10% for S=0. The new approach mainly affects lakes with low or highly variable inflow and high potential evaporation which are mostly found in semi-arid or arid regions. During dry season the water balance of these lakes is predominantly controlled by evaporation and actual storage regularly drops below zero. With the new approach, such lakes are prevented from dropping to unrealistic low levels, such that outflow can occur in wet years even after extensive dry periods. Comparisons between simulated and observed discharge at stations downstream of large lakes and reservoirs, e.g. Lake Malawi, showed that the new approach also leads to a better representation of average outflow. Lakes with higher and more constant inflow are hardly affected as their storage levels mostly vary within the positive range.

In contrast to lakes, water storage in wetlands cannot become negative in the model. In former versions of WGHM, wetland surface area and thus evaporation was assumed to be independent of water storage until, abruptly, evaporation was set to zero at S=0. Thus, the likely decline in surface area and thus evaporation with decreasing water

storage in the wetland was not taken into account. Recognizing a generally stronger decline of surface area with declining water levels in the case of wetlands as compared to lakes, in WGHM 2.1f, the following wetland evaporation reduction factor is introduced:

$$r = 1 - \left(\frac{|S - S_{\max}|}{S_{\max}} \right)^{p} \quad (2)$$

with S actual wetland storage [m^3], S_{max} maximum wetland storage [m^3] and p wetland reduction exponent (p=3.32). Wetland evaporation is reduced by 10% when the actual storage is half of the maximum storage and becomes zero when the storage is empty. The new algorithm has little effect under wet conditions, as evaporation is hardly reduced with an actual storage exceeding 50% of maximum storage. However, impacts are significant under dry conditions. As a consequence of reduced evaporation, drying up of wetlands by evaporation becomes slower, while replenishment by inflow becomes faster. The outflow curve is smoother, as complete desiccation, with outflow becoming zero, is less likely.

2.2 Model tuning against observed river discharge

WGHM is tuned against river discharge observed at gauging stations around the world. For each station, 30 years of discharge data were used (or fewer years if less than 30 years of data were available). If the discharge data contained more than 30 years, the 30 year period that corresponded best with the period from 1961 to 1990 was selected, as WaterGAP climate input is most reliable for this time span. The goal of model tuning is to adjust the simulated long-term average discharge at the outflow point of the subbasin to the observed long-term average discharge (Döll et al., 2003).

2.2.1 Tuning factors

In order to avoid overparameterization (Beven, 2006) and to make tuning in a large number of sub-basins feasible, only the soil water balance is tuned by adjusting one model parameter, the runoff coefficient. The runoff coefficient γ determines the fraction of effective precipitation (precipitation or snowmelt) P_{eff} [mm/d] that becomes runoff from land R_l [mm/d] at a given soil water saturation:

$$R_l = P_{\text{eff}} \left(\frac{S_s}{S_{s\max}} \right)^{\gamma} \quad (3)$$

with S_s soil water content within the effective root zone [mm] and $S_{s\max}$ total available soil water capacity within the effective root zone [mm]. γ is adjusted in a sub-basin specific manner, i.e. all grid cells within the inter-station area are given the same value. The values of γ are allowed to range only between 0.3 and 3. However, for many basins, observed long-term discharge cannot be simulated with a deviation of less than 1% by adjusting γ. This is due to a number of reasons, among them errors in input data (e.g. precipitation and radiation), errors in the estimation of human water consumption and neglecting important processes like river water loss to subsurface and evaporation of runoff e.g. in small ephemeral ponds. Besides, the water balance of lakes and wetlands remains unaffected by adjusting the model parameter, but can be very important for the water balance of a basin. In these cases, an areal correction factor CFA is computed which adjusts total runoff (the sum of runoff from land and surface water bodies) of each cell in the sub-basin equally. As there are sub-basins that contain both cells with positive (precipitation > evapotranspiration) and negative (evapotranspiration > precipitation) cell water balance, CFA can take two values symmetric to 1.0 within one sub-basin. If it is necessary to increase runoff in a basin, a CFA greater than one (e.g. 1.2) is used for cells with positive mean water balance and CFA is set to the corresponding value below one (e.g. 0.8) for cells with negative water balance. In former model versions, a CFA range from 0 to 2 was allowed, which however may lead to problems particularly in small and/or dry downstream basins, where observed inflow and

outflow are very similar. In some of these cases, CFA was set to zero, impeding runoff generation at every single time step, which is not plausible. To avoid this unwanted effect, CFA is restricted to a range from 0.5 to 1.5 in WGHM 2.1f.

CFA does not suffice to simulate observed long-term average river discharge in all sub-basins if the impact of errors and misrepresentations mentioned above is too strong. Furthermore, even minor errors of discharge measurement may inhibit that sub-basin runoff can be adjusted by CFA in small sub-basins at middle or lower reaches of rivers with comparatively high discharge. Thus an additional station correction factor CFS is required for several basins to assure correct average inflow into downstream subbasins. CFS simply corrects discharge at the grid cell where the gauging station is located such that the simulated long-term average discharge at that grid cell is equal to the observed value (Döll et al., 2003).

Please note that in basins where correction factors are used, the dynamics of the water cycle are no longer modeled in a consistent manner. Where CFA is used, cell runoff from all grid cells within a basin is adjusted such that the sum of grid cell runoff is equal to the difference between the long-term average discharge of the basin's station and the next upstream station(s), but cell runoff is no longer consistent with soil water storage or evapotranspiration. In basins with CFA, the model serves to interpolate measured discharge in space and time. For these basins, application of CFA in model simulations allows a more realistic simulation of runoff, discharge and water storage dynamics in groundwater and surface waters.

When, in addition, CFS is required, discharge becomes discontinuous along the river, from the cell downstream of the station to the cell where the station is located. Grid cell runoff remains unaffected by CFS and thus discharge is inconsistent with runoff. The advantage of using CFS is that the long-term inflow to downstream stations is set to the observed value, which increases the chance of adequately simulate downstream discharge.

2.2.2 Observational data

WGHM 2.1f was tuned against discharge observed at 1,235 gauging stations. These data were provided by the Global Runoff Data Center (GRDC) in Koblenz, Germany. In this paper, the resulting model variant is called V2. Variant V1 was tuned against the discharge dataset that was used for tuning WGHM 2.1d (and 2.1e), consisting of 724 stations. Both station sets had to be co-registered with the drainage direction map DDM30 (Döll and Lehner, 2002), which required considerable checking and some adjustment of geographical location. The V1 and the V2 station data were selected according to the same rules (Döll et al., 2003; Kaspar, 2004):

- minimum basin size area of the most upstream station: 9000 km^2
- minimum inter-station basin area: 20 000 km^2
- minimum length of observed time series of monthly river discharge: four years

In V2, 133 of the 1235 stations have a time series length of less than 10 years, 245 stations of 10–19 years, 375 of 20–29 years, and for 482 stations, 30 years of discharge were used for tuning. Figure 1 shows the location of tuning stations in variants V1 and V2. Of the 724 V1 stations, 627 were kept in V2. 97 V1 stations were not considered in the new dataset, as stations with longer or more recent time series were available in the vicinity. The remaining 608 stations that are used in V2 were not yet included in V1. Please note that in case of 102 of the 627 stations that are in both V1 and V2, the available discharge time series have changed significantly. At 83 stations, time series length has increased by more than 20% (V1 average: 14 years, V2 average: 25 years), while for the remaining 19 stations, the time period of the tuning years shifted to more recent years by more than 20% of the tuning period (average shift: 10 years towards present).

V2 represents a distinct densification of stations especially in North America and northern Asia. Densification is low in Europe as V1 already includes a relatively dense

station net there. In South America, most new stations are located in Brazil, and in Australia, in the Murray-Darling basin. In central and southern Asia, the Aral lake basin has been particularly densified, and in Africa, the Congo basin. The total basin area covered by V2 (69.9 million km^2 or 48.7% of the global land area without Greenland and Antarctica) exceeds the area covered by V1 by about 3.4 million km^2 or 2.4% of the total land area. The largest additional areas are located within the Niger (Africa), Paraná (South America) and Khatanga (Siberia) basins as well as in northern Canada and Alaska.

2.2.3 Technical constraints to tuning

Despite tuning, long-term average observed and simulated discharges differ by more than 2% in case of 29 of the 724 stations of V1 and in case of 83 of the 1235 stations of V2. Of the 627 stations that are common to V1 and V2, 31 stations are concerned. This problem is due to two technical constraints in the tuning procedure of WGHM. First, in normal model runs, water consumption requirements can be fulfilled by taking water from a neighboring cell which even may be located outside the basin where the requirement exists. This could not be implemented in the tuning process and leads to discrepancies particularly in small, narrow and water scarce basins with intensive water use. This applies to around 90% of the affected basins in V2. Most of them are located in the semiarid regions of the USA and Mexico, while a few others can be found in central and southern Asia. Besides, model initialization in tuning runs starts 5 years before the specific tuning period of a station. The two model runs V1 and V2 examined in this study, however, were started in 1901 and thus generally have a longer forerun until they reach the evaluation period, i.e. the tuning period. As a consequence, discrepancies in the fill level of the basins' water storages can occur at the beginning of the evaluation period. The variations are mostly negligible as at least five years ahead of the evaluation period are identical in both cases. However, in eight V2 basins located in Alaska and Siberia that are dominated by surface water bodies, discrepancies in discharge are noticeable.

2.3 Indicators of model performance

In order to characterize model performance and quality, it is assessed how well the model simulates six observed river flow characteristics (Table 1). Certain flow characteristics are particularly relevant for specific water management fields like water supply (in particular long-term average flow, low flows, variability of annual and monthly flows), flood protection (high flows) and ecosystem protection (seasonality of flows, low flows). Time series of simulated (S) and observed (O) monthly river discharge values are compared with respect to these flow characteristics, and the goodness-of-fit is quantified by indicators.

A common measure for the goodness-of-fit in hydrology is the modeling efficiency E, or the Nash-Sutcliffe coefficient (Nash and Sutcliffe, 1970):

$$E = 1.0 - \frac{\sum_{i=1}^{n} (O_i - S_i)^2}{\sum_{i=1}^{n} \left(O_i - \bar{O}\right)^2} \tag{4}$$

It is defined as the mean squared error normalized by the variance of the observed data subtracted from unity. Thus it represents model success with respect to the mean as well as to the variance of the observations. While a coefficient of one represents a perfect fit of simulated and observed time series, values below zero indicate that the average of observed discharge would still be a better estimation than the model. The problem with using E to compare two variants is that one cannot distinguish whether the higher E-value is due to a lower mean error or to a better representation of the variance.

To overcome this problem, in this study two measures are applied that allow a distinct evaluation of the model with respect to the simulation of the variance and the mean. The first measure is the well known coefficient of determination (R^2) with a range from zero to one, which describes how much of the total variance in the observed data is

explained by the model:

$$R^2 = \left\{ \frac{\sum_{i=1}^{n} \left(O_i - \bar{O}\right)\left(S_i - \bar{S}\right)}{\left[\sum_{i=1}^{n} \left(O_i - \bar{O}\right)^2\right]^{0.5} \left[\sum_{i=1}^{n} \left(S_i - \bar{S}\right)^2\right]^{0.5}} \right\}^2 \tag{5}$$

In analyses of time series, R^2 evaluates linear relationships between the observed and the modeled data. It is not sensitive to systematic over- or underestimations of the model, concerning magnitude of the modeled data (mean error) as well as its variability (Legates and McCabe, 1999; Krause et al., 2005). Besides, R^2 – like the coefficient of efficiency E – tends to be sensitive to outliers, which may lead to a bias in model evaluation towards high flow events and has to be considered regarding the results. Nevertheless, R^2 is assumed to provide fundamental information on how well the sequence of higher and lower flows in an observed discharge time series is represented by the model.

As second measure, we introduced the "symmetric deviation factor" SDF which describes the mean error of discharge simulation as the ratio of observed and simulated discharge values (or vice versa). It can be applied to both time series and aggregated values. SDF is defined as

$$\mathrm{SDF} = \left\{ \begin{matrix} \frac{S}{O} & \text{for } S \geq O \\ \frac{O}{S} & \text{for } S < O \end{matrix} \right\}. \tag{6}$$

SDF ranges from plus one to infinity, with values close to one representing good fits between simulated and observed values. SDF reflects that an underestimation by a factor of 2 (S=0.5*O), for example, represents reality as well (or badly) as overestimation by a factor of two (S=2*O). In both cases, SDF is equal to 2. This understanding of goodness-of-fit is, however, not mirrored by the usually applied error measures like absolute error or relative error, which are bounded below. In case of underestimation,

the error cannot be larger than the observed value or 100%, while in case of overestimation, error values are unlimited. For the above example, the relative error would be −50% in the case of underestimation, but 200% in the case of overestimation. This asymmetric character makes interpretation difficult, in particular when these measures are averaged. SDF is symmetric and unlimited both in case of over- and of underestimation.

SDFs of long-term average, low and high flows are computed by inserting the respective simulated and observed values (one per basin and variant) in Eq. (6). SDFs of time series (annual, monthly and mean monthly flows) are determined by first calculating SDF for each year, month or the twelve monthly means of the observation period, and then computing the median; thus SDF represents the median deviation of the values. For computation of R^2, the annual, monthly or mean monthly values are inserted into Eq. (5).

For overall assessment of model performance, all indicators are averaged over stations. For R^2, the arithmetic mean was chosen, while the median was preferred for SDF, as it is not sensitive to single outliers. SDF can become very large if either the simulated or the observed discharge is very close to zero. In case that simulated or observed discharges equal zero at a certain time step, the respective value is excluded from SDF averaging.

3 Results and discussion

We will now answer the three questions posed in Sect. 1 which will help to assess the value of (additional) river discharge information in global hydrological modeling.

3.1 Does increased river discharge information promote tuning WGHM by only one model parameter?

Comparing variant V2 to variant V1, the area for which tuning was done increases by 5.1% to 69.9 million km^2, which is equivalent to 48.7% of the global land area excluding Greenland and Antarctica (Table 1). Figure 2 shows for which river basins WGHM 2.1f could be tuned by adjusting only the runoff coefficient γ, with an error of less than 2%, in case of V1 (724 stations) and V2 (1235 stations). There are two major effects of densification of river discharge information. On the one hand, in several very large basins, in particular in Siberia, that cannot be tuned with V1, the finer discretization of V2 allows tuning of at least some sub-basins (Fig. 1). On the other hand, a few V2 sub-basins of larger V1 sub-basins that can be tuned as a whole with V1 (e.g. Ganges, Congo), cannot adjusted with V2 (Fig. 2). In all world regions, there are basins, that can be tuned in V1 only and not in V2, and basins that can be tuned in V2 only and not in V1. Only in Siberia and Australia, a positive effect of densification is obvious (more stations can be tuned in V2).

In case of V1, 384 of the 724 sub-basins or 31.3 million km^2 could be tuned by adjusting only the runoff coefficient γ (Table 2). In case of V2, the number of these sub-basins increases to 546 and the area to 33.9 million km^2. The fraction of sub-basins that could be tuned decreases from 53.0% for V1 to 44.2% for V2, but the respective tuning basins area, as a fraction of total tuning basins area (total land area, except Greenland and Antarctica) increases slightly from 47.0% (21.8%) for V1 to 48.5% (23.7%) (Table 2). It has to be pointed out that tuning success or failure can not directly be linked to model performance. A highly subdivided river basin with only a few successfully tuned sub-basins might be much closer to reality than an entirely adjusted spacious basin where errors balance out by chance at the outlet.

The basin area where only γ and the areal correction factor CFA had to be adjusted increased from 247 to 300, but the corresponding basin area decreased strongly from 38.2% to 22.3%. At the same time, the area where the station correction factor CFS

had to be introduced, increased strongly from 6.9% of the land area to 14.2% (Table 2), and the number of corrected stations increased from 93 to 389. V2 basins which require CFS are mainly located in snow dominated (e.g. Alaska, northern Canada and northern Siberia) and very dry areas (e.g. northern Africa, Central Asia), where the model can not account for all essential processes of the water cycle.

One reason for the increased amount of sub-basins that can only be adjusted by CFS might be the decreased average sub-basin size in V2. CFA is adjusted by comparing simulated and observed runoff generation within a sub-basin. Observed runoff generation is determined as observed discharge at the outflow station minus the sum of discharges at upstream stations. In sub-basins that are located in middle or lower reaches of a river the relative influence of local runoff generation on total river discharge gets lower as the sub-basin area becomes only a small fraction of the total basin area.

Thus, the benefit of tuning against more discharge observations is that the basin area where long-term average discharge can be computed correctly by adjusting only the model parameter γ has increased by more than 8%, and that the number of stations (but not the percentage of stations) where this is possible also increased. Siberia, where station density is very low in V1, shows the most pronounced increase in area. However, the cost of tuning against more discharge observations is high, as the area where a station correction factor is required doubles. This means that the area with inconsistent runoff generation and discharge, and with discontinuous discharge values along the river network, doubles.

3.2 To what extent does tuning against more discharge observations improve model performance?

The question is to what extent and in which cases the adjustment of long-term average river discharge at more stations (and using changed observation time series) improves the simulation of the other five flow characteristics in Table 1. For a comprehensive answer of this question, four research questions are posed:

A) Does tuning against longer or more recent discharge time series improve model performance?

B) Does tuning against discharge at more stations improve model performance...

B1) ...within the total V1 tuning area?

B2) ...outside the total V1 tuning area?

C) To what extent does the segmentation of a station's basin into sub-basins improve model performance at that station?

D) To what extent does the segmentation of a station's basin into sub-basins improve model performance inside the basin?

These research questions are answered in Sects. 3.2.2 to 3.2.5, taking into account the 6 flow characteristics listed in Table 1.

Five question-specific subsets of the entire station dataset were generated. To answer question A, 60 stations were selected that 1) belong to both V1 and V2, 2) have the same basin in V1 and V2 and 3) comprise significantly changed time series of observed discharge (subset A). To answer questions B to D, only those stations were considered where the time series has not changed significantly from V1 to V2. Subsets B1 and B2 combined include all of these stations, except those with the same basins in V1 and V2. The resulting 747 stations are used to evaluate the overall change in model performance due to discharge observations at more stations inside V1 tuning area (subset B1: 691 stations) and outside V1 tuning area (subset B2: 56 stations). Subset C, with 117 stations, is applied to investigate the effects of finer watershed segmentation on the discharge simulation at the outflow points of the respective basins (question C). It contains only those stations of subset B that are common to V1 and V2 and that have more upstream stations in V2 than in V1. Finally, question D is answered based on subset D that includes 387 tuning stations located within zero-order basins (i.e. basin draining into the ocean or terminal internal sinks) showing a considerable

increase of station density in V2 as compared to V1, i.e. where average sub-basin size decreases by at least 50%.

To demonstrate typical effects of refined tuning on the simulation of flow characteristics and on the associated indicators Fig. 3 displays evaluation results at two exemplary discharge stations in the USA. The station at Old Hickory, Cumberland River, belongs to subsets B1 and D, i.e. it is not part of the V1 dataset and is located in a zero-order basin with significantly increased tuning station density (Fig. 3a). After tuning against long-term average discharge, the annual hydrograph of V2 primarily shows a significant shift towards the observed hydrograph, while its variance remains virtually unchanged as compared to V1. This is reflected by a decrease in average deviation of annual variability (median SDF V1: 1.31, V2: 1.11), while R^2 hardly changes. The mean monthly hydrograph of V2 additionally indicates a better representation of flow variance, which is distinctly underestimated by V1. With V2, particularly the representation of receding and rising discharges between May and December is improved. Consequently, both SDF and R^2 values of monthly flow characteristics (seasonal and monthly variability) are significantly better in V2. However, monthly variance is still underestimated by the model. This becomes evident regarding monthly Q_{90} which is improved but still overestimated, and monthly Q_{10} which is underestimated by V2.

The station at Dalles, Columbia River, belongs to subsets B1 and C, i.e. it is a tuning station in both V1 and V2 (Fig. 3b). While its sub-basin covers 192 000 km^2 in V1, it is subdivided into 8 smaller sub-basins in V2 with an average area of 24 000 km^2. In contrast to the Old Hickory Dam station, there is no general shift between simulated hydrographs of V1 and V2, as they are both adjusted against average discharge. The left hydrograph shows that changes in annual discharges are negligible which is also reflected by unchanged SDF and R^2 of annual variability. SDF values of all monthly characteristics, including seasonal and monthly variability as well as low and high flows, indicate slight improvements, while R^2 of the variability characteristics remains rather constant. Regarding the mean monthly hydrographs, representation of flows in spring and autumn becomes somewhat better, however, changes between V1 and V2 ap-

pear rather insignificant compared to the remaining discrepancy between observed and simulated hydrographs. This discrepancy is caused by assuming, in WGHM 2.1f, that man-made reservoirs behave like natural lakes.

As a first analysis step, the impact of additional discharge information on the capability of WGHM to represent long-term average discharges, i.e. renewable freshwater resources, is analyzed in Sect. 3.2.1 by looking at the spatial pattern of changes.

3.2.1 To what extent does tuning against more discharge observation improve the representation of long-term average river discharge?

Figure 4 depicts the deviation of long-term average discharge as computed with WGHM 2.1f V1 from the observed value at V2 stations. The map shows the value of additional stations and prolonged time series. The larger the SDF, the less accurate WGHM would have computed long-term average discharge without the information included in V2, and the higher is the value of the additional discharge information. In variant V2, all SDFs should be zero. However, as described in Sect. 2.2.3, 83 sub-basins, concentrated in the semi-arid, heavily irrigated parts of the USA and Mexico, could not be tuned satisfactorily due to technical constraints in the tuning procedure. Hence, their SDF values differ from 1 not only in V1, but also in V2 and the improvements achieved by applying V2 are lower than expressed by the SDF of V1. Therefore, in Fig. 4, the values for these basins were corrected by subtracting (SDF_{V2}–1.0) from SDF_{V1}. If, for instance, SDF_{V1} equals 1.5 and SDF_{V2} equals 1.2, the corrected value would be 1.5–(1.2–1.0)=1.3.

In most regions of Europe, where the network of tuning stations has already been dense in V1, the additional discharge information in V2 does not improve model representation of long-term average discharge much. Only few sub-basins show SDF values above 1.5 (e.g. in northern Spain and Scandinavia), i.e. sub-basins where discharge computed without the additional information is off by a factor of more than 1.5. Improvements are somewhat more pronounced in eastern Europe (Volga basin), and distinctly higher in the large Siberian basins of Ob, Yenisey and Lena where the tun-

ing dataset has been significantly densified in V2. In the basin of the Tobol River, a contributory to the Ob River, SDF even reaches values above 6. In central, southern and southeastern Asia additional discharge information is scarce, and the majority of the few refined basins show SDF values above 1.5, and even above 3 in the Aral Sea basin. In Australia, performance improvements are large in the Murray-Darling basin because the number of stations has increased from 2 to 8 and the basin is strongly affected by human intervention, i.e. irrigation withdrawals and locks (reservoirs). Obviously, the impact of irrigation and reservoirs is not modeled accurately enough by WGHM. In Africa, the majority of additional tuning stations are located in the Niger and Kongo basins. The map shows SDF values between 1.1 and greater than 6 in most of their sub-basins. In southern Africa, where only the tuning time series changed (dotted sub-basins in Fig. 4) but no new tuning stations were added, SDF values remain below 1.5 except for one small basin. In the lower Paraná and upper Amazon basins as well as in some smaller South American basins, SDF is between 1.1 and 1.5, while in the Río Colorado/Río Salado basin, tuning with a more recent discharge time series leads to an even more pronounced performance. In North America, the value of additional stations is particularly high in semi-arid basins like the Colorado River and Rio Grande basins and in the western sub-basins of the Mississippi. Besides, several sub-basins of the Yukon and the Mackenzie show SDF values above 3. In all these areas the density of tuning stations increased distinctly. In the eastern, more humid parts of North America, SDF is below than 1.1 in most sub-basins.

In summary, WGHM representation of long-term average discharge (i.e. renewable freshwater resources) is strongly improved by additional discharge information in the case of large basins that have been significantly subdivided in V2, like in the large Siberian basins, the Congo basin or the Murray-Darling basin. The value of the additional discharge information tends to be higher in semi-arid and snow dominated regions where results of WGHM, and hydrological models in general, are typically less reliable (e.g. the western part of North America). Conversely, the value of additional discharge information is lower in basins where the model (including its input data like

precipitation) is more reliable and tuning station density is already high in V1 (e.g. in Central Europe). In general, the value of additional stations is higher than the value of longer time series, but the performance gains can be significant in case of formerly very short time series, e.g. for the Indus (formerly 4, now 14 years) and the Orange River (9 and 29 years, respectively).

3.2.2 Does tuning against longer or more recent discharge time series improve model performance?

Subset A used to investigate this question comprises 46 discharge observation stations with significantly extended time series (by more than 20%) and 14 stations with a tuning period shifted to more recent years (by more than 20% of the tuning period). The upper left diagram in Fig. 5 compares V1 and V2 with regard to deviation between observed and simulated discharges (determined by SDF) at 60 stations for the six flow characteristics. While results for low flows, high flows and annual variability show only very small improvements with V2, improvements are somewhat more pronounced for long-term average, seasonal variability and monthly variability. The diagram on the lower left depicts the percentage of stations where SDF improved, did not change or declined in V2 as compared to V1, according to the flow characteristics. A SDF change of at least 3% is considered to be significant. Regarding long-term average discharge, two thirds of the stations improved, whereas the rest did not change. As the model is tuned against average discharge and the evaluation period corresponds with the V2 tuning period a decline could only occur due to tuning errors. For low and high flows 60–70% of the stations show changed SDF results in V2. Improved stations are prevailing in both cases over declined stations, although results are somewhat better for high flows. Annual, seasonal and monthly variability changes are less pronounced. The majority of stations indicate no SDF change. While the ratio of improved to declined stations is clearly positive for annual and monthly variability (3.4 and 2.3), seasonal variability holds exactly the same number of improved and declined stations (13).

Diagrams on the right in Fig. 5 display the R^2 results, as a measure of goodness-of-

fit with respect to the variance. Comparing versions V1 and V2 (upper right diagram), none of the characteristics show a significant change in mean R^2. The percentage of all stations where R^2 did not change significantly (i.e. by more than 3%) ranges from 92% for seasonal variability to 97% for annual variability (lower right diagram), indicating that a significant change occurred at only 2 to 5 out of 60 stations. This indicates that the improved SDF of the time series of annual and monthly discharges and of the mean monthly discharges is almost exclusively due to shift in the long-term average discharge, but not due to better representation of the variability of flow.

To summarize, the presented results show that tuning against longer or more recent discharge time series leads to a noticeable impact regarding the deviation between modeled and simulated flow characteristics. Benefits are most pronounced for long-term average discharge, seasonal variability and monthly variability. Changed observation time series, however, have hardly any effect on the model's representation of flow variability.

3.2.3 Does tuning against discharge at more stations improve model performance within and outside the total V1 tuning area?

Subset B1 is applied to answer the first part of this question and comprises 691 tuning stations with altered sub-basin structure. It contains a number of stations that have already been part of V1 as well as all additional V2 stations that are located within the V1 tuning area and thus provides an overall evaluation of the performance changes that are associated to the densification of the tuning dataset. Median SDF is significantly improved for all flow characteristics (Fig. 6 top). The improvements are most obvious for long-term average discharge and decrease slightly towards the right of the diagram. The fraction of stations with significantly reduced deviation between simulated and observed flow characteristics is considerable. It covers more than half of the tuning stations regarding long-term average, high flows and annual variability, while the remaining flow characteristics still show 43.3% (monthly variability) to 48.4% (low flows) of improved stations (Fig. 6 bottom). The percentage of stations with declined

performance is low for all flow characteristics except low flows where it amounts to about 30% of the stations. The fraction of stations with improved performance outweighs the fraction of stations by declined performance by a factor of 1.6 (low flows) to 6 (annual variability). The positive impact of tuning long-term average discharge at more stations on simulating flow variability is very small but higher than in the case of changed time series (Fig. 5 right).

In *subset B2*, only those 56 stations are considered that are located outside the total V1 tuning area. In V1, discharge in these basins is computed with a regionalized tuning parameter γ that depends on three basin-specific characteristics (see Döll et al., 2003, for details). Thus, subset B2 provides information on how tuning changes model performance in basins where there was even no information of observed discharge further downstream. Not surprisingly, improvements of median SDF are much higher than for subset B1 (Fig. 6). On average, long-term average discharge at these ungauged stations differ, without tuning, by a factor of 1.8 from the observed value. The additional discharge information also strongly improves the simulation of high flow and annual variability. Please note, however, that the SDF of all flow characteristics for V2 except annual variability are higher than the corresponding SDFs in subset B2. Figure 7 (lower left diagram) shows that for 80–95% of the B2 basins high flow, annual variability and long-term average discharge are significantly better estimated if taking into account the additional discharge information. Low flow estimation, however, is effected negatively in most basins even though the SDF of low flows improves. This, the overall lower performance as compared to subset B1 and the strong improvement of the long-term average may be explained by the fact that most of the B2 basins are located in snow-dominated or semi-arid regions where model results and in particular low flow are generally less reliable. Like for subset B1, the positive impact of tuning long-term average discharge at more stations on simulating flow variability is very small (Fig. 7 right), with 60–70% of the stations showing no significant change of R^2. The number of stations with improved performance outweighs that with declined performance by a factor of around 1.4 for all three flow characteristics.

3.2.4 To what extent does the segmentation of a station's basin into sub-basins improve model performance at that station?

Tuning at upstream stations is expected to improve model performance at the downstream station, as tuning may make the simulated partitioning of precipitation into evapotranspiration and runoff more realistic, such that the dynamics or at least the magnitude of basin inflow are simulated better. The performance improvements are expected to be lower than for subsets B1 and B2, as discharges at the basin outflow stations themselves were used for tuning in both variants. To test this hypothesis, the model performance indicators of Table 1 are computed for *subset C*, i.e. all stations that are common to V1 and V2 and where the upstream basins have changed.

Comparing both model variants (not displayed in a figure) indicates that, even though the number of basins with improved performance is higher than the number of basins with declined performance (by factors ranging from 1.4 to 3.4) for all flow characteristics except annual variability (0.8), median SDFs of all flow characteristics hardly show any changes. As changes in the representation of flow variances are even more insignificant, it is supposed that overall the segmentation of a station's basin into sub-basins does not improve model performance at that station.

3.2.5 To what extent does the segmentation of a station's basin into sub-basins improve model performance inside the basin?

With this question, we would like to determine the effect of a significant reduction of sub-basin size on model performance inside a zero-order river basin (like in case of the Murray-Darlin basin). *Subset D*, which is a subset of B1, includes only V2 tuning stations located within zero-order basins where average V2 sub-basin area is reduced to less than half of the V1 basin area. Differences in model performance between V1 and V2 (Fig. 8) are somewhat more distinct than in case of subsets B1 (Fig. 6). The SDFs of all six flow characteristics are higher for subset D than for subset B1 for V1, but more similar for V2. The fraction of stations with improved performance outweighs the

fraction of stations by declined performance by a factor of 1.3 (low flows) to 2.3 (monthly variability). Like for the other subsets, the positive impact of tuning long-term average discharge at more stations on simulating flow variability is insignificant. Please note for all subsets, seasonal variability, with mean R^2 values ranging between 0.63 and 0.75, is generally better modeled than annual variability (0.37–0.59) and monthly variability (0.38–0.50).

3.3 What is the impact of basin size on model performance and basin-specific tuning?

The basin sizes of the discharge stations used for tuning WGHM 2.1f V2 range from 9000 km^2 up to 1 244 000 km^2, with a mean of about 56 000 km^2. As already discussed in the introduction, basins size is an important factor with respect to model performance and tuning. To evaluate the impact of basin size, subsets B1 and B2 were merged. The new subset contains all 747 V2 stations that have an altered basin structure as compared to V1. The subset was divided into five size classes. Class boundaries and the number of associated stations are shown in the header of Table 3.

The impact of basins size on model performance of WGHM 2.1f V2 with respect to the flow characteristics of Table 1 is shown in Table 3a. Median SDF in Table 3a represents average deviation of observed and simulated discharges for five basin size classes, with lower values indicating a better model performance. Class V with the largest basins sizes (>100 000 km^2) shows the best performance for all flow characteristics except high flows (here class IV performs insignificantly better). Performance in class IV is good, too, with four out of five values better than average. While performance in class II is comparable to class IV, results are more diverse in class III and comprise three values significantly worse than average (low flows, seasonal and monthly variability). Class I (basin sizes between 9000 and 20 000 km^2) clearly performs worst with regard to all investigated flow characteristics with all values representing the minimum of all classes.

Mean R^2 is used to investigate the impact of basin size on the models representation of flow variance. Table 3a shows mean R^2 for the three variability flow characteristics

with regard to the basin size classes, with higher values indicating a better fit between observed and simulated variance. Here, class IV displays the best results with best R^2 fit for seasonal and monthly variability and above-average fit for annual variability. R^2 of the largest class (V) is significantly higher than average for seasonal variability and in the range of average for the remaining characteristics. Class III performs very well with respect to annual variability while the other results are below average. Again, results are somewhat better in class II with R^2 close to average for seasonal and monthly variability and best of all classes for annual variability. Like median SDF, mean R^2 is worst for all flow characteristics in class I with values between 5 and 15% below average.

With respect to both goodness-of-fit measures, size class I (>9000–20 000 km^2) clearly performs worst. While results are distinctly better in class II (20 000–40 000 km^2), performance decreases again for most flow characteristics in class III (40 000–60 000 km^2). The best values can be found in classes IV (60 000–100 000 km^2) and V (>100 000 km^2). The reason for the below-average performance in class I might be that sub-basins below 20 000 km^2 are too small for errors in input data to balance out. A reason for the lower performance of class III as compared to class II may be that regions with high data availability and quality like Europe and the USA are over-represented in class II. As WGHM performance strongly depends on input data quality (i.e. precipitation), model results are generally more reliable in these regions. Basins larger than 60 000 km^2 show the best model performance for all flow characteristics. Obviously, it is not important that the tuning parameter γ and the areal correction factor CFA are kept constant over the whole area, which may lead to blur spatial discrepancies in large heterogeneous catchment and decreased model performance. The dominant effect appears to be that, given the data resolution and spatial uncertainty, input data is better represented in large basins as these hold a better chance for errors to balance out.

The *impact of basin size on model tuning* is investigated in two ways. Table 3b provides the percentage of stations that could be tuned by the adjusting the model's

tuning parameter γ only as well as the fraction where either the area correction factor (CFA) or both CFA and the station correction factor (CFS) had to be applied. Table 3c lists percent changes of median SDF and mean R^2 as measures of model performance of variant V2 as compared to variant V1.

Regarding the application of tuning factors, the size classes display a very diverse behavior. Results are best in class III where nearly half of the sub-basins could be adjusted by γ only, and 67% without using station correction. In larger basins, the fraction of only γ-adjusted basins is only slightly lower (45–46%), but CFS-corrected sub-basins amount to 41% in class IV, while they only reach 32% in class V. Whereas results are somewhat worse in class II, class I shows by far the worst results. Here, less than one quarter of the sub-basins could be adjusted without using correction factors and station correction had to be applied in 91 out of 195 cases (47%).

Improvements in model performance achieved by applying V2 discharge information are generally highest in class I – except for low flows – even though performance of V2 results is significantly below average in this class. The positive effect of tuning is still significant in classes II and IV with rather low performance in V1 but reasonably good SDF values in V2. In classes III and V improvements are less pronounced. While class V already showed good results in V1, performance of class III rather remains on a low level. As seen above, the impact of tuning to long-term average discharge on simulating flow variability is very low, so that the result that the highest performance gains occur in the two largest size classes (lower part of Table 3c) is difficult to interpret.

In summary, the smallest basins (9000–20 000 km^2) appear to be less suited for tuning because correction factors have to be applied in more than 75% of the basins, with the ensuing loss of model consistency. They also show by far the lowest modeling performance with respect to the flow characteristics low flow, high flow and annual, seasonal and monthly variability even after tuning against long-term average observed river discharge. However, for these basins, tuning affords the highest performance increase, with median SDFs decreasing e.g. by 33% for long-term average discharge, such that tuning of these basins can be considered as particularly valuable if the mod-

eling goal is a better representation of observed flow characteristics.

4 Conclusions

The goal of this study was to investigate the value of observed river discharge data for global-scale hydrological modeling of a number of flow characteristics that are required for assessing water resources, water scarcity, flood risk and habitat alteration of aqueous ecosystems. To our knowledge, this has never been done before. Observed river discharge is certainly valuable for determining the quality of model results, but it can also be used to tune not only catchment-scale but also global-scale hydrological models. We think that it is essential in global-scale hydrological modeling to take advantage of the aggregated information on river basin processes and flows that is included in observed river discharge because model input data like precipitation, radiation or soil characteristics are particularly uncertain at this scale.

The global hydrological model WGHM 2.1f uses observed long-term averages of river discharge to tune the model such that simulated long-term average discharge at the observation station (grid cell) is equal to the observed value. In this study, we analyzed discharge that was computed by two model variants, V1 which had been tuned against a data set of 724 stations used in former versions of the model (Döll et al., 2003), and V2, which had been tuned against a new data set of 1235 stations, with extended time series.

WGHM is tuned against observed long-term average discharge by adjusting only one model parameter (γ) that affects runoff generation of land areas. Correction factors are applied in basins where γ does not suffice to adjust the modeled long-term average river discharge to the observed one. Tuning with the extended observed discharge data set V2 resulted in an increase of the land area that could be tuned without correction factors of more than 8%, which is mainly due to the densification of stations in Siberia. The number of stations but not the percentage where this is possible increased. However, the land area where not only the areal correction factor but also

the station correction factor had to be applied increased strongly, which is a strong disadvantage, as the application of this factor makes discharge inconsistent with runoff and leads to discontinuous discharge at the outflow of the respective sub-basin. Small basins between 9000 and 20 000 km^2 are particularly problematic, as almost half of them required a station correction factor. Only 25% of them could be tuned by only adjusting γ, while for larger basins, this was the case in more than 40%.

The impact of additional discharge information on model performance was investigated by comparing river discharge as simulated by WGHM versions V1 and V2 to observed values with respect to six flow characteristics including long-term average discharge, low flows (monthly Q_{90}), high flows (monthly Q_{10}) as well as annual, seasonal and monthly variability of discharge. In general, the value of additional stations is higher than the value of longer time series except in cases with formerly very short time series. Representation of long-term average discharge, which at least for humid regions is a good measure of renewable freshwater resources, is significantly improved by additional discharge information. The stations with the highest benefit are those new stations that are located outside of V1 basins. Without tuning, simulated values of long-term average discharge would differ from observed ones by a factor of 1.8 on average (56 stations, subset B2). When considering only the stations that are located within zero-order basins where average sub-basin size has decreased by at least 50% (387 stations, subset D), the respective value is 1.3. Large river basins that have been considerably subdivided in V2, like in the Siberian basins, the Congo basin or the Murray-Darling basin, show the highest benefits. The value of the additional discharge information tends to be higher in semi-arid and snow dominated regions where results of WGHM, and of hydrological models in general, are typically less reliable. Conversely, the value of additional discharge information tends to be lower where station density was already high in V1 and simulations are generally more reliable, like in Europe.

Looking at the other five flow characteristics, their deviation from observed values, as computed by the symmetric deviation factor SDF, decreases due to tuning against additional discharge data. Again, the basins outside the V1 basins (subset B2) show

the highest performance gains due to tuning the long-term average discharge, followed by the stations inside significantly densified basins. The stations that are included in both V1 and V2 but with additional upstream stations in V2, only show a very small increase in the performance as measured by the SDF values. All subsets show a strong correlation between decreased SDF of the long-term average discharge and the other flow characteristics. Tuning long-term average discharge does not lead to a significant improvement of the representation of flow variance. This is not even the case for subset B2, with R^2 of annual, seasonal and monthly variability increases by only 0–3%, even though here the stations with an improved R^2 outnumber those with a decreased R^2. We conclude that decreased deviation of annual and monthly discharges from observed values, which leads to lower SDF for all flow characteristics, is almost exclusively due to adjustments of the mean. It remains to be investigated if basin-specific tuning of a second model parameter which impacts flow variability is viable and useful, either using discharge characteristics in addition to long-term average discharge (as listed in Table 1) or information on large-scale (mainly seasonal) water storage variations as obtained by GRACE gravity data (Güntner et al., 2007). We think that improved modeling of storage and outflow dynamics of reservoirs, lakes and wetlands is likely to be necessary before any basin-specific calibration of a second model parameter is to be undertaken.

The optimal sub-basin size for tuning depends on the modeling purpose. Small basins below 20 000 km^2 show a much stronger improvement in model performance due to tuning than larger basins, while the improvement decreases with increasing basin size. This is related to the dependence of model performance on basin size. It is significantly lower for basins of less than 20 000 km^2 (before and after tuning) than for larger basins, with basins over 60 000 km^2 performing best. On the other hand, tuning of small basins requires the application of the station correction factor in almost half of them. Utilizing a very dense network of tuning stations thus leads to a less consistent model, but provides a significantly better spatial representation of river flow characteristics, while tuning with a network of sub-basins with more than 20 000 km^2

leads to a more consistent model which is however associated with higher uncertainty regarding the spatial distribution of discharge and renewable water resources within the sub-basins.

In conclusion, tuning of WGHM 2.1f against a new dataset of river discharge observed at 1235 stations world-wide has lead to a more realistic representation of the spatial pattern of river discharge and renewable water resources at the global scale. It better serves the modeling objective of combining the best data available to derive realistic and meaningful descriptions of terrestrial water flow characteristics. However, by forcing modeled long-term average river discharges to become equal to the respective observed values, simulation of temporal flow variability has not been improved significantly and model consistency has suffered. Unfortunately, errors in input data and the hydrological model can only be compensated to a rather limited extent by tuning against observed river discharge. Our study nevertheless shows that the value of observed river discharge data for global-scale hydrological modeling is high enough to warrant its use not only for model validation.

Acknowledgements. The authors thank the Global Runoff Data Centre in Koblenz, Germany, for collecting observed river discharge data and for making them available to the authors. They are grateful to K. Fiedler, Frankfurt University, for her contributions to model tuning.

References

Adam, J. C. and Lettenmeier, D. P.: Adjustment of global gridded precipitation for systematic bias, J. Geophys. Res.-Atmos., 108(D9), 4257, doi:10.1029/2002JD002499, 2003.

Alcamo, J., Flörke, M., and Märker, M.: Future long-term changes in global water resources driven by socio-economic and climatic changes, Hydrol. Sci. J., 52, 247–275, 2007.

Alcamo, J., Döll, P., Henrichs, T., Kaspar, F., Lehner, B., Rösch, T., and Siebert, S.: Development and testing of the WaterGAP 2 global model of water use and availability, Hydrol. Sci. J., 48, 317–338, 2003.

Arnell, N. W.: A simple water balance model for the simulation of streamflow over a large geographic domain, J. Hydrol., 217, 314–335, 1999.

Becker, A. and Braun, P.: Disaggregation, aggregation and spatial scaling in hydrological modelling, J. Hydrol., 217, 239–252, 1999.
Beven, K. J.: A manifesto for the equifinality thesis, J. Hydrol., 320, 18–36, 2006.
Döll, P., Kaspar, F., and Lehner, B.: A global hydrological model for deriving water availability indicators: model tuning and validation, J. Hydrol., 270, 105–134, 2003.
Döll, P. and Lehner, B.: Validation of a new global 30-min drainage direction map, J. Hydrol., 258, 214–231, 2002.
Döll, P., Kaspar, F., and Alcamo, J.: Computation of global water availability and water use at the scale of large drainage basins, Mathematische Geologie, 4, 111–118, 1999.
Fekete, B. M., Vörösmarty, C. J., and Grabs, W.: High-resolution fields of global runoff combining observed river discharge and simulated water balances, Global Biogeochem. Cycles, 16, 1042, doi:10.1029/1999GB001254, 2002.
Flörke, M. and Alcamo, J.: European outlook on water use, Final report, Center for Environmental Systems Research, University of Kassel, Kassel, 2004.
Güntner, A., Stuck, J., Werth, S., Döll, P., Verzano, K., and Merz, B.: A global analysis of temporal and spatial variations in continental water storage, Water Resour. Res., 43, W05416, doi:10.1029/2006WR005247, 2007.
Haddeland, I., Lettenmaier, D. P., and Skaugen, T.: Effects of irrigation on the water and energy balances of the Colorado and Mekong river basins, J. Hydrol., 324, 210–223, 2006.
Hagemann, S. and Dümenil, L.: A parametrization of the lateral waterflow for the global scale, Clim. Dynam., 14, 17–31, 1998.
Kaspar, F.: Entwicklung und Unsicherheitsanalyse eines globalen hydrologischen Modells, Dissertation, Department of Natural Sciences, University of Kassel, Kassel, 2004.
Klepper, O. and van Drecht, G.: WARibaS, Water Assessment on a River Basin Scale; A computer program for calculating water demand and satisfaction on a catchment basin level for globalscale analysis of water stress. RIVM, Bilthoven, The Netherlands, Report 402001009, 124 pp., 1998
Krause, P., Boyle, D. P., and Bäse, F.: Comparison of different efficiency criteria for hydrological model assessment, Adv. Geosci., 5, 89–97, 2005, http://www.adv-geosci.net/5/89/2005/.
Legates, D. R. and McCabe, G. J.: Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation, Water Resour. Res., 35, 233–241, 1999.
Legates, D. R. and Willmott, C. J.: Mean seasonal and spatial variability of gauge-corrected,

global precipitation, Int. J. Climatol., 10, 111–117, 1990.
Lehner, B., Döll, P., Alcamo, J., Henrichs, H., and Kaspar, F.: Estimating the impact of global change on flood and drought risks in Europe: a continental, integrated assessment, Climatic Change, 75, 273–299, 2006.
Lehner, B. and Döll, P.: Development and validation of a global database of lakes, reservoirs and wetlands, J. Hydrol., 296, 1–22, 2004.
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.
Moody, J. A. and Troutman, B. M.: Evaluation of the depth-integration method of measuring water discharge in large rivers, J. Hydrol., 135(1–4), 201–236, 1992.
Nash, J. E. 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.
Oki, T., Nishimura, T., and Dirmeyer, P.: Assessment of land surface models by runoff in major river basins of the globe using Total Runoff Integrating Pathways (TRIP), J. Meteor. Soc. Japan, 77, 235-255, 1999
Schulze, K. and Döll, P.: Neue Ansätze zur Modellierung von Schneeakkumulation und -schmelze im globalen Wassermodell WaterGAP, Tagungsband zum 7. Workshop zur großskaligen Modellierung in der Hydrologie, München, 27–28 November 2003, Kassel University Press, 145–154, Kassel, 2004.
Siebert, S., Döll, P., Hoogeveen, J., Faures, J.-M., Frenken, K., and Feick, S.: Development and validation of the global map of irrigation areas. Hydrol. Earth Syst. Sci., 9, 535–547, 2005, http://www.hydrol-earth-syst-sci.net/9/535/2005/.
Smakhtin, V., Revenga, C., and Döll, P.: A pilot global assessment of environmental water requirements and scarcity, Water International, 29, 307–317, 2004.
Vassolo, S. and Döll, P.: Global-scale gridded estimates of thermoelectric power and manufacturing water use, Water Resour. Res., 41, W04010, doi:10.1029/2004WR003360, 2005.
Yates, D. N.: Approaches to continental scale runoff for integrated assessment models, J. Hydrol., 201, 289-310, 1997

Table 1. River flow characteristics and related indicators of model quality.

River flow characteristic	Indicators
1 Long-term average flow	Median SDF[a] of arithmetic mean of annual discharge
2 Low flow	Median SDF of monthly Q_{90}^{b}
3 High flow	Median SDF of monthly Q_{10}^{c}
4 (Variability of) Annual flows	Median SDF and mean R^2 of time series of annual discharge
5 Seasonality of flow	Median SDF and mean R^2 of mean monthly discharge[d]
6 (Variability of) Monthly flows	Median SDF and mean R^2 of time series of monthly discharge

[a] SDF: Symmetric deviation factor, with SDF = simulated/observed if simulated ≥ observed, and SDF = observed/simulated otherwise.
[b] Monthly discharge that is exceeded in 9 out of 10 months.
[c] Monthly discharge that is exceeded in 1 out of 10 months.
[d] 12 values per station (January to December).

Table 2. Number and area of basins that could be tuned, in V1 and V2, by only adjusting the model parameter γ, or with applying, in addition, the areal correction factor CFA and the station correction factor CFS.

	WGHM 2.1f variant	
	V1	V2
all tuning basins	724	1235
area [10^6 km^2]	66.5	69.9
fraction of land area*	46.4%	48.7%
basins adjusted by γ only	384	546
fraction of tuning basins	53.0%	44.2%
fraction of tuning area	47.0%	48.5%
fraction of land area*	21.8%	23.7%
basins adjusted by γ and CFA	247	300
fraction of tuning basins	34.1%	24.3%
fraction of tuning area	38.2%	22.3%
fraction of land area*	17.7%	10.9%
basins adjusted by γ, CFA and CFS	93	389
fraction of tuning basins	12.8%	31.5%
fraction of tuning area	14.8%	29.2%
fraction of land area*	6.9%	14.2%

*143.4×10^6 km^2 (without Greenland and Antarctica).

Table 3. Impact of basin size on model performance and basin-specific tuning. Model performance (a), percentage of station that are adjusted by γ, CFA and CFS (b) and percent change in model performance (c) with respect to flow characteristics according to five basin size classes (italic figures: value above average of classes, bold figures: best value).

basin size class	I	II	III	IV	V	all stations	avg. of classes
basin size (1000 km^2)	<20	20–40	40–60	60–100	>100		
no. of stations	195	301	99	64	88	747	149
(a)							
Median SDF (V2)							
long-term average discharge	1.00	1.00	1.00	1.00	1.00	1.00	1.00
low flows	1.86	1.64	1.83	1.77	***1.64***	1.71	1.75
high flows	1.26	1.22	1.19	***1.18***	1.19	1.22	1.21
annual variability of discharge	1.17	1.14	1.14	1.15	***1.14***	1.15	1.15
seasonal variability of discharge	1.56	1.45	1.54	1.46	***1.38***	1.49	1.48
monthly variability of discharge	1.79	1.67	1.72	1.59	***1.50***	1.69	1.65
Mean R^2 (V2)							
annual variability of discharge	0.44	***0.56***	0.56	0.54	0.53	0.53	0.53
seasonal variability of discharge	0.76	0.79	0.78	***0.86***	0.84	0.79	0.81
monthly variability of discharge	0.46	0.49	0.48	***0.52***	0.50	0.48	0.49
(b)							
Percentage of stations that were adjusted by							
tuning with γ only	24.6%	41.2%	48.5%	45.3%	45.5%	38.7%	41.0%
correction with CFA	28.7%	21.9%	18.2%	14.1%	22.7%	22.6%	21.1%
corrected with CFA & CFS	46.7%	36.9%	33.3%	40.6%	31.8%	38.7%	37.9%
(c)							
Percent change in median SDF: V1 as compared to V2							
long-term average discharge	***−32.5%***	*−15.2%*	−10.3%	−12.1%	−2.0%	−15.2%	−14.4%
low flows	*−7.3%*	*−12.7%*	3.1%	***−14.8%***	−1.6%	−9.8%	−6.6%
high flows	***−25.1%***	−9.1%	−4.7%	−7.2%	−3.5%	−10.7%	−9.9%
annual variability of discharge	***−23.7%***	−8.6%	−5.4%	−9.7%	−5.1%	−9.7%	−10.5%
seasonal variability of discharge	***−12.8%***	*−9.1%*	−4.0%	*−9.6%*	−4.1%	−9.4%	−7.9%
monthly variability of discharge	***−12.8%***	*−6.8%*	−1.1%	*−10.3%*	−2.2%	−6.0%	−6.7%
Percent change in mean R^2: V1 as compared to V2							
annual variability of discharge	0.7%	*1.9%*	1.4%	−1.2%	***4.4%***	2.4%	1.4%
seasonal variability of discharge	*2.3%*	0.6%	*1.5%*	***2.8%***	−0.2%	0.5%	1.4%
monthly variability of discharge	*6.7%*	3.2%	−0.1%	4.3%	***7.8%***	1.7%	4.4%

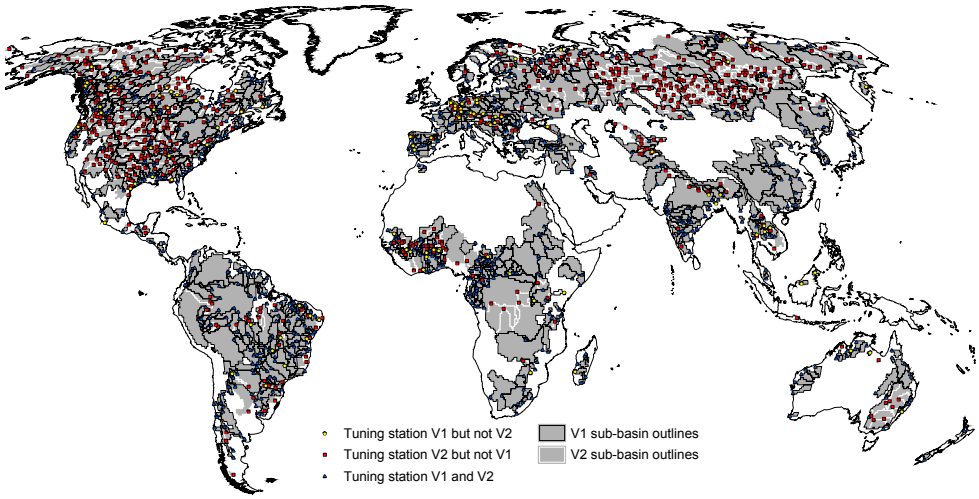


Fig. 1. River discharge observation stations used for tuning WGHM variants V1 (724 stations) and V2 (1235 stations), with their drainage basins.

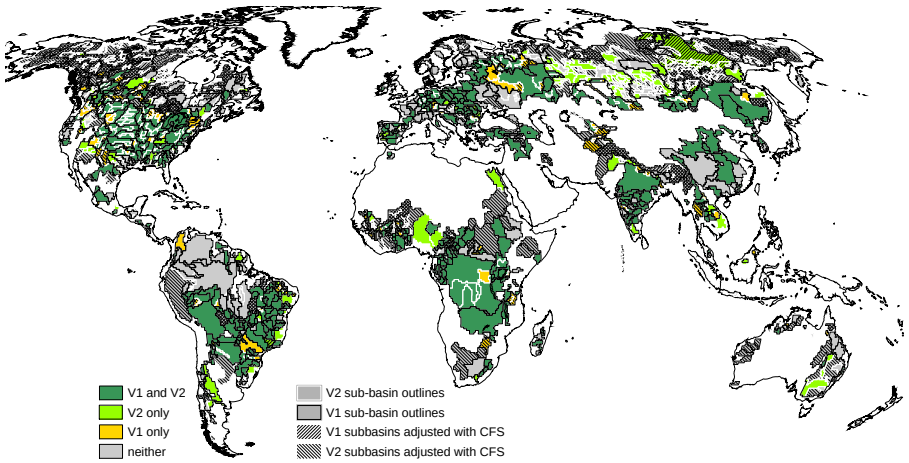
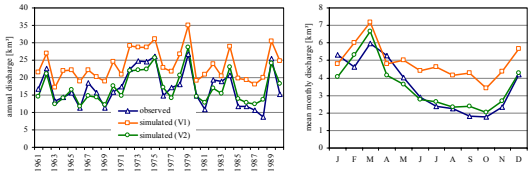


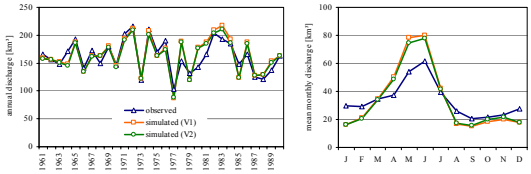
Fig. 2. Results of tuning WGHM 2.1 f variants V1 and V2. The color of the basins indicates whether each variant can compute observed long-term average river discharge at the stations by only adjusting the runoff coefficient. In the striped sub-basins, discharge needs to be adjusted by an additional station correction factor CFS.

(a) Old Hickory Dam Station (Tennessee), Cumberland River (subsets B1 and D)



	absolute values [km³]				indicator values	
flow characteristic	obs.	V1	V2	indicator	V1	V2
long-term average (annual)	17.2	23.5	17.2	SDF	1.37	1.00
low flows (monthly Q_{90})	0.49	1.07	0.66	SDF	2.19	1.35
high flows (monthly Q_{10})	2.76	3.06	2.45	SDF	1.11	1.13
annual variability				median SDF	1.31	1.11
				R^2	*0.79*	*0.81*
seasonal variability				median SDF	1.43	1.14
				R^2	*0.61*	*0.94*
monthly variability				median SDF	1.66	1.31
				R^2	*0.40*	*0.54*

(b) The Dalles Station (Oregon), Columbia River (subsets B1 and C)



	absolute values [km³]				indicator values	
flow characteristic	obs.	V1	V2	indicator	V1	V2
long-term average (annual)	162	164	162	SDF	1.01	1.00
low flows (monthly Q_{90})	7.90	4.87	5.15	SDF	1.62	1.53
high flows (monthly Q_{10})	22.5	30.0	29.0	SDF	1.33	1.29
annual variability				median SDF	1.05	1.05
				R^2	*0.77*	*0.77*
seasonal variability				median SDF	1.36	1.29
				R^2	*0.89*	*0.90*
monthly variability				median SDF	1.36	1.33
				R^2	*0.72*	*0.72*

Fig. 3. Comparison between V1 and V2 model results and observed discharges at two exemplary tuning stations. Annual and mean monthly hydrographs and indicator values with respect to the different stream flow characteristics are shown.

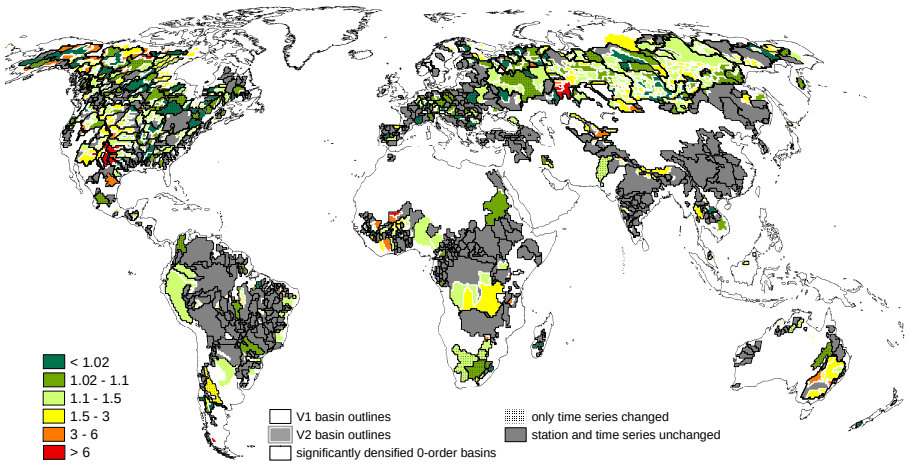


Fig. 4. Value of additional discharge information with respect to long-term average discharge (renewable water resources). The corrected basin-specific SDF of WGHM 2.1f variant V1 shows model performance at stations only considered for tuning in V2 (sub-basins where neither the station nor the discharge time series for tuning changed between V1 and V2 are shown in grey). The higher SDF is, the higher is the value of additional discharge information. For SDF=1, simulated and observed values are identical, while for SDF=2, for example, the observed value is either under- or overestimated by a factor of 2 without tuning.

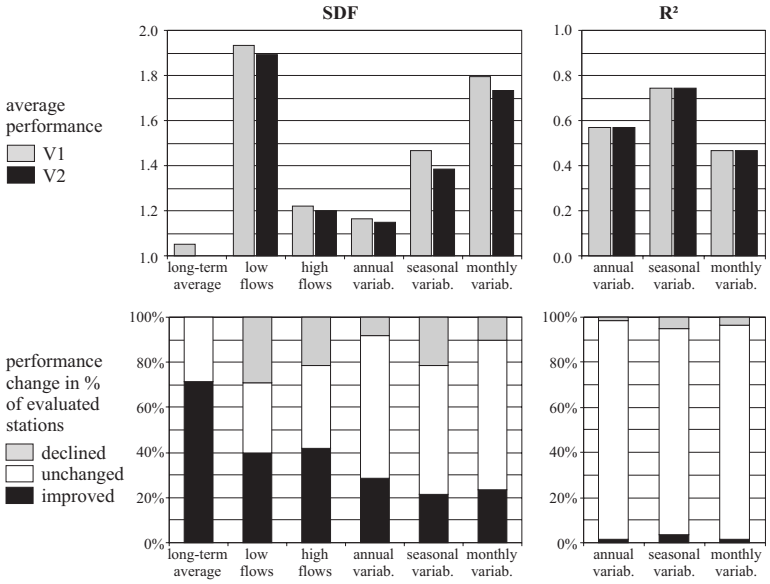


Fig. 5. Model performance of WGHM 2.1f at discharge tuning stations with extended or more recent time series in V2 as compared to V1 (subset A with 60 stations). Low SDF and high R^2 values indicate good model performance.

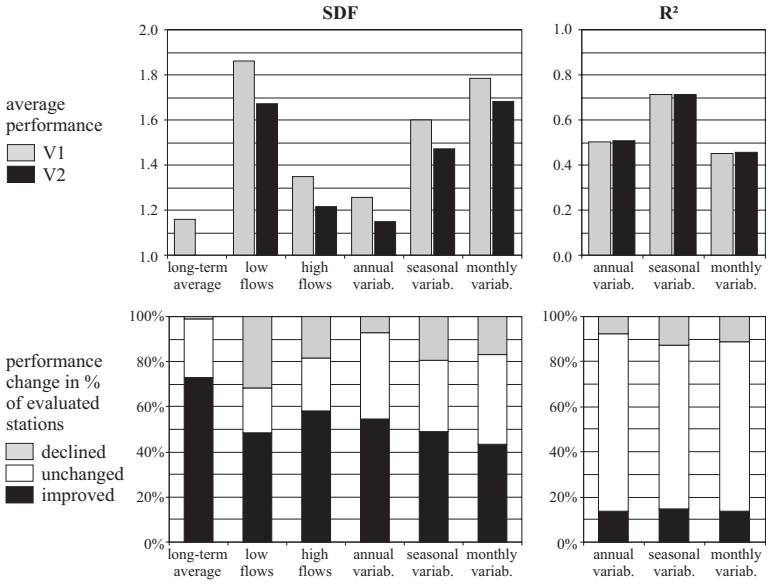


Fig. 6. Model performance of WGHM 2.1f at discharge tuning stations with altered V2 sub-basin structure within the V1 tuning area (subset B1 with 691 stations). Low SDF and high R^2 values indicate good model performance.

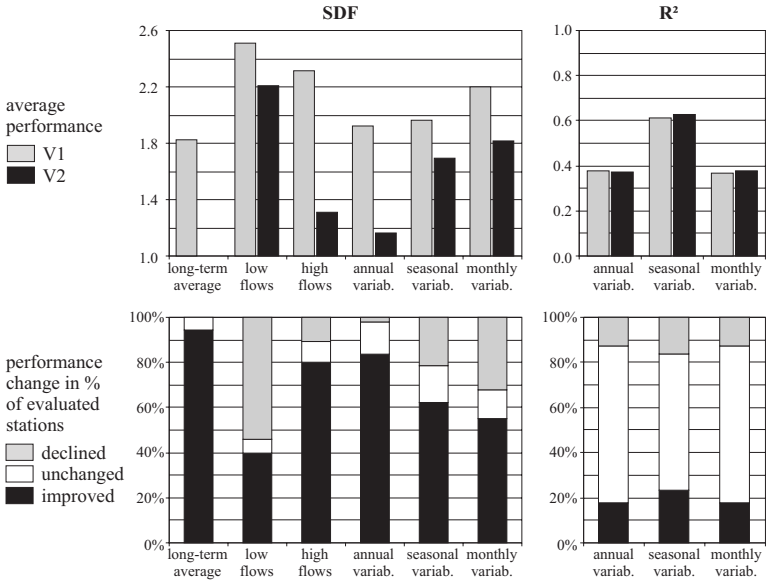


Fig. 7. Model performance of WGHM 2.1f at V2 discharge tuning stations outside the V1 tuning area (Subset B2 with 56 stations). Low SDF and high R^2 values indicate good model performance.

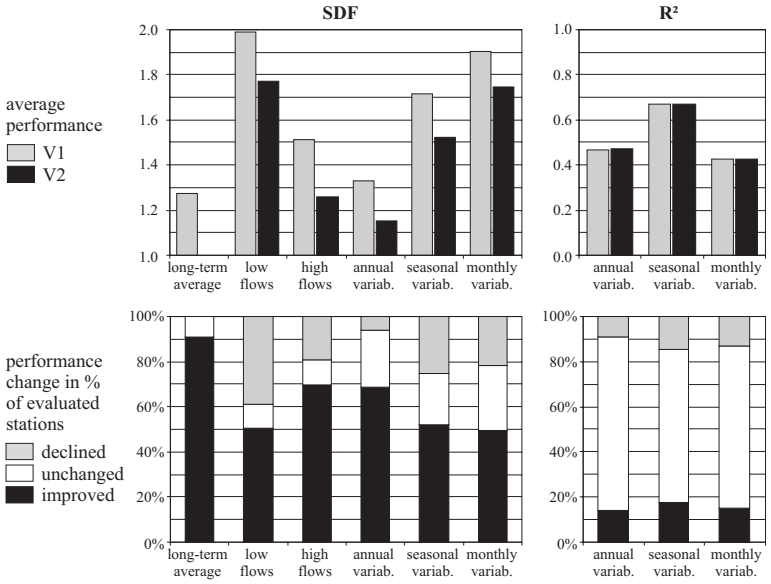


Fig. 8. Model performance of WGHM 2.1f at discharge tuning stations inside river basins where average V2 sub-basin size has been decreased by at least 50% compared to V1 (subset D with 387 stations). Low SDF and high R^2 values indicate good model performance.