Hydrol. Earth Syst. Sci., 18, 343–352, 2014
www.hydrol-earth-syst-sci.net/18/343/2014/
doi:10.5194/hess-18-343-2014

Hydrology and Earth System Sciences
Open Access

Data-driven scale extrapolation: estimating yearly discharge for a large region by small sub-basins

L. Gong

Department of Earth Sciences, Uppsala University, Uppsala, Sweden

Correspondence to: L. Gong (lebing@gmail.com)

Received: 29 March 2012 – Published in Hydrol. Earth Syst. Sci. Discuss.: 1 June 2012
Revised: 30 October 2013 – Accepted: 6 December 2013 – Published: 28 January 2014

Abstract. Large-scale hydrological models and land surface models are so far the only tools for assessing current and future water resources. Those models estimate discharge with large uncertainties, due to the complex interaction between climate and hydrology, the limited availability and quality of data, as well as model uncertainties. A new purely data-driven scale-extrapolation method to estimate discharge for a large region solely from selected small sub-basins, which are typically 1–2 orders of magnitude smaller than the large region, is proposed. Those small sub-basins contain sufficient information, not only on climate and land surface, but also on hydrological characteristics for the large basin. In the Baltic Sea drainage basin, best discharge estimation for the gauged area was achieved with sub-basins that cover 5 % of the gauged area. There exist multiple sets of sub-basins whose climate and hydrology resemble those of the gauged area equally well. Those multiple sets estimate annual discharge for the gauged area consistently well with 6 % average error. The scale-extrapolation method is completely data-driven; therefore it does not force any modelling error into the prediction. The multiple predictions are expected to bracket the inherent variations and uncertainties of the climate and hydrology of the basin.

1 Introduction

The interests in understanding current and future water resources have driven the rapid development of large-scale hydrological models (e.g. Arnell, 1999, 2003, 2004; Vörösmarty et al., 1989, 2000a, 2004). Water resource projections made by those models are an important basis for socio-economical analyses and decision-making processes (e.g. Vörösmarty et al., 2000a). Projections of water resources are believed to be associated with large uncertainty, especially in ungauged basins that cover around 50 % of the global land area. For instance, global runoff estimates from various models differ between 29 000 $km^3\ yr^{-1}$ and 43 000 $km^3\ yr^{-1}$ (i.e. around 30 %), and continental estimates differ up to 70 % (Widén-Nilsson et al., 2007). Besides climate and discharge data uncertainties, model uncertainties also significantly contribute to the uncertainties of the simulated discharge (Widén-Nilsson et al., 2009). A number of regionalisation methods have been developed to extend the prediction capability of hydrological models into ungauged areas. Commonly used regionalisation methods utilise spatial proximity and catchment similarity to transfer model parameters from gauged to ungauged basins (e.g. Kokkonen et al., 2003; Huang et al., 2003; Xu 1999, 2003; Kim and Kaluarachchi, 2008; McIntyre et al., 2005). Model averaging (i.e. using average of model outputs from different proximity or similarity approaches) was found to provide more robust results in regionalisation (e.g. McIntyre et al., 2005). Hydrological models inherently have limited parameter transferability over different spatial scales; therefore large-scale regionalisation methods use large gauged river basins as potential donors. However, averaged basin characteristics often cannot sufficiently summarise small-scale variability and nonlinearity, which might limit the prediction accuracy of the regionalisation methods.

Recent advance in the prediction in ungauged basins has identified that information such as timing of seasonal precipitation and potential evaporation, as well as higher frequency variations in rainfall-runoff process, may also contribute to the prediction of annual runoff in ungauged basins (Blöschl et al., 2013). In the meantime, annual water balance

and annual runoff variability are governed, to the first order, by the relative availability of water and energy, while topography, basin storage and biological processes modulate these effects (Blöschl et al., 2013). It has long been recognised that the interaction between climate and hydrology controls the nonlinear partitioning of precipitation (e.g. L'vovich, 1979; Budyko, 1974; Wagener et al., 2007). L'vovich (1979) and Budyko (1974) were among the first to characterise climate and hydrology using long-term average water and energy balance variables. The aridity index, as expressed by the ratio of long-term average potential evapotranspiration to that of precipitation, has long been used as a useful index describing the interaction between climate and hydrology of a region (e.g. Wagener et al., 2007). Interestingly, a number of similarity studies have shown that climate has a universal control over hydrology for basins over a wide range of spatial scales, i.e. from 10 to 10 000 km^2 (Troch et al., 2009; Voepel et al., 2011; Brooks et al., 2011). The scale independence of hydrological similarity indicates that small gauged basins can potentially be used as predictors for large-scale hydrological responses, provided that the small basins and the large region are similar in their essential climatic and hydrological parameters. In contrast to regionalisation methods, this paper uses the similarity of climate time series as the foundation for extrapolation, instead of using similarity index and regression-based methods. This paper aims at developing a systematic methodology that allows discharge data of small basins to be extrapolated to a much larger scale. The main purpose of the paper is to present the methodology of scale extrapolation; however, we also showed how the method worked in one test basin (the Baltic Sea basin) with the preliminary results.

2 Study area and data

3 The Baltic Sea drainage basin

The extrapolation method was tested in the Baltic Sea drainage basin (Fig. 1). The Baltic Sea is one of the largest brackish seas in the world; the Baltic Sea drainage basin lies between maritime temperate and continental subarctic climate zones. With a surface area of 415 000 km^2, the drainage basin spans 14 countries with 85 million inhabitants, a majority of them living in big cities. The Baltic Sea is semi-enclosed and therefore vulnerable to pollution, and its environmental status is one of the major concerns for the northern European countries. The Baltic Sea is affected by pollution from various sources including nutrient input from rivers, pollution from industries, and direct atmospheric depositions (Wulff et al., 2001). Many of these factors are dependent on the climate and hydrology in the basin.

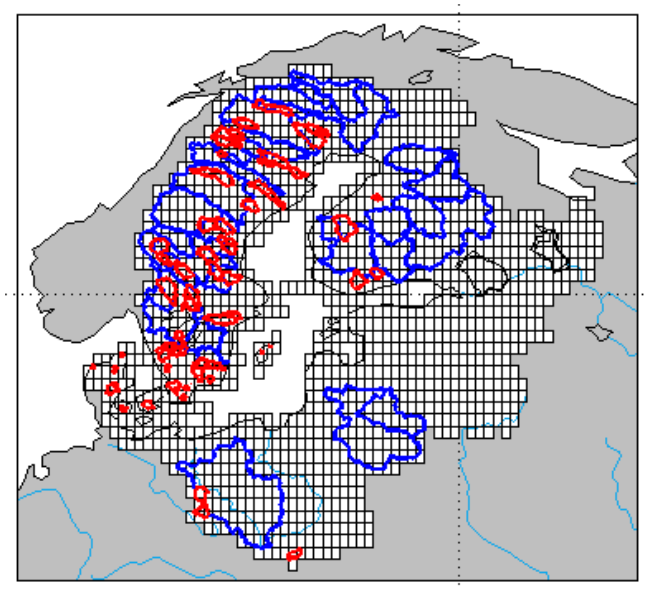

Fig. 1. Map of the Baltic Sea drainage basin as shown by 0.5 degree STN-30p global grid cells, with boundaries of 100 gauged sub-basins shown by lines. Source sub-basins are marked with red and the rest marked with blue.

4 Data sets

Monthly precipitation for the period of 1975–2001 was taken from the 30-minute monthly Climatic Research Unit Time-Series (CRU TS) 2.1 database (Mitchell and Jones, 2005). The number of stations used by the CRU TS 2.1 data set has significant temporal variations (Mitchell and Jones, 2005). Spatial density of CRU precipitation stations in the Baltic Sea drainage basin decreased after 1990. Monthly precipitation data from 1984 SMHI (Swedish Meteorological and Hydrological Institute) precipitation stations for the period of 1961–2002 were interpolated to a regular 30 min grid, and the quality of the CRU precipitation data within Sweden was validated against the SMHI data prior to the analysis. The results (figure not shown) showed that the spatial differences between CRU and SMHI annual average precipitation were similar for the period of 1961–1990 and 1991–2002. Differences between 1991–2002 and 1961–1990 mean annual precipitation as calculated by CRU data and SMHI data also agreed well in their general spatial pattern, although those calculated with SMHI data showed much higher spatial variability at smaller scales.

WATCH (WATer and global CHange) forcing data (WFD, Weedon et al., 2010) for the period between 1975 and 2001 at 30 min spatial resolution were used to derive potential evaporation. The WFD provides bias-corrected variables based on the ERA-40 reanalysis product of the European Centre for Medium-Range Weather Forecasting (ECMWF) as described by Uppala et al. (2005). Specific humidity, atmospheric pressure, 2 m air temperature, 10 m wind speed,

downward short-wave radiation and net long-wave radiation were used to calculate reference evaporation using the Penman–Monteith FAO-56 equation (Allen et al., 1998). Specific humidity was first converted to relative humidity using a mixing-ratio method, and 10 m wind speed was converted to 2 m wind speed using a logarithmic relationship (Allen et al., 1998). Prior to the calculation of reference evaporation, the quality of the WFD air temperature, wind speed, and WFD-derived relative humidity was tested in a comparison with daily weather data (Global Surface Summary of the Day, or GSOD) from the National Climatic Data Center (NCDC, 2011). In the Penman–Monteith FAO-56 equation, surface albedo is fixed at 0.23; however we found this value is too high for the Baltic Sea basin. Therefore, the albedo values were taken directly from the ERA-Interim data set (Simmons et al., 2007). The daily WATCH forcing data were aggregated to obtain yearly values (calendar year) for each 30 min grid cell. The monthly CRU precipitation data were also aggregated to yearly values.

STN-30P data set (Vörösmarty et al., 2000b) was used to identify 1386 cells on a regular 30 min global grid that belong to the Baltic Sea drainage basin. HYDRO1k (USGS, 1996) was used to delineate the upstream area of the discharge stations. The discharge data were taken from the Global Runoff Data Centre database (GRDC, 2012) and the SMHI Vatten Web (http://vattenweb.smhi.se/). Among 425 available sub-basins, 100 sub-basins were selected under the following criteria: (1) they do not contain nested sub-basins; (2) when registered in the Hydro1k river network, the register area does not differ by more than 20 % with the reported area from GRDC or SMHI; and (3) they have complete daily data coverage from 1975 to 2001. Figure 1a shows the location of the 100 sub-basins. The sizes of the sub-basins vary between 5 and 109 564 km^2. The area covered with the 100 gauged sub-basins, denoted as "gauged basin area", was used to validate the scale-extrapolation method (Fig. 1). The successfulness of the scale extrapolation depends on the abundance of discharge data from small river basins. For the extrapolation to perform well, it is critical to select river basins within a suitable size range. The resolution of the available global or regional climate data set defines the lower limit for the size of the small river basins that can be used for extrapolation (i.e. the size of a river basin should be comparable with the climate grid), so reliable climate data can be obtained for the basin. Preliminary results showed that river basins between 500 and 5000 km^2 are most useful for discharge extrapolation at the global scale, considering that the resolution of most global climate data sets is 0.5 degree. Therefore, only 51 sub-basins between 500 and 5000 km^2, denoted as "source sub-basins" (Fig. 1), were selected for discharge extrapolation.

5 Self-similarity in hydrological response

In this paper, hydrological similarity refers to two or more basins that share similar factors controlling the discharge dynamics. The controlling factors may include basin size, topography, soil, vegetation, climate, geology, as well as factors that can be derived directly from data, for instance runoff coefficients, and factors that can be derived with the help of modelling or data analysis techniques, for instance topographic index, aridity index and Horton index.

What can be more similar to a basin than the basin itself? If an ungauged basin A (Fig. 4a) is identical in every hydrological controlling factor with a gauged basin B, then A shall have the same discharge as B. But it is virtually impossible to find such a identical gauged basin, especially if A is a large-scale basin.

Topography and river channel networks have long been known to be self-similar. Inside a river basin one can always find a sub-region with similar topographic and channel network features. However, if discharge is to be extrapolated from a sub-region to the whole basin, all first-order controlling factors of the sub-region should be self-similar to the basin. Therefore, there is a need to extend the self-similarity measures to include all important factors that control the hydrological response.

Each hydrological controlling factor, be it climate forcing or land surface parameters, exhibits spatial auto-correlation. Part of the spatial information is repetitive or redundant; there is only a limited number of unique patterns that define the hydrological dynamics of a basin. Those patterns can be time series of climate forcing, or spatial statistics of a land surface parameter. For instance, when a number of cells within the gauged area of the Baltic Sea drainage basin was selected by the criterion that they must well resemble the temporal variation of yearly precipitation of the gauged area, the average correlation among those cells dropped significantly if less than 5 % of the cells were selected (Fig. 2). If more cells were selected, there would be significant correlation among the cells so that the addition of new cells may be redundant (Fig. 2).

An important step towards finding a hydrologically self-similar sub-set of a basin is not to restrict self-similar sub-set to be one single enclosed area, but instead to be a collection of several spatially independent sub-regions, each representing a unique pattern of the climate–hydrology interaction. The process of finding a hydrologically self-similar sub-set is denoted as "factor matching", i.e. finding a number of grid cells (denoted as "source cells") inside a basin that share similar hydrological controlling factors as the basin itself. Once the matching is done and source cells found, the area-weighted discharge of the source cells can be extrapolated to the entire basin. Two strategies were used in this paper to maximise the chance of finding the source cells:

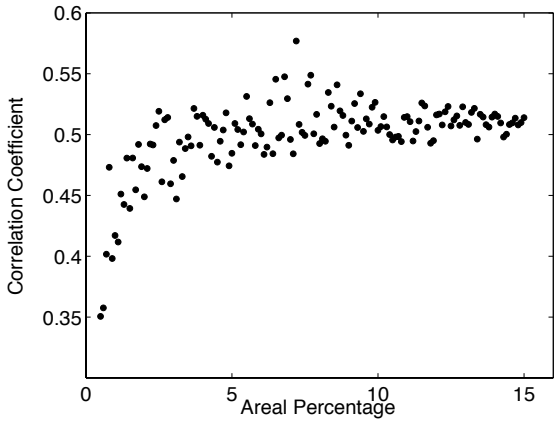

Fig. 2. Correlation coefficients (y axis) between annual precipitation time series of selected cells from the source sub-basins, as a function of the areal ratio of selected cells to the gauged area of Baltic Sea drainage basin (x axis).

1. Use only small (in the context of global hydrology) sub-basins. Both climate and hydrology exhibit larger spatial and temporal variability at smaller scales. A large spectrum of climate and land surface patterns can be obtained by combining several small basins.

2. Allow partial selections of cells within each source sub-basin. Therefore, a source sub-basin can contribute any number of cells (from zero to its total number of cells) to the final selected source cells. This strategy not only greatly increases the chance of a good "factor matching", but also opens up the possibility of having a vast number of equally good realisations of source cells (i.e. different groups of cells that are hydrologically similar to each other and to the large basin).

A simple two-step test is made to illustrate the importance to allow the source cells to be spatially discrete. In step one, each source sub-basin alone was selected as a candidate to represent the yearly discharge time series of a gauged area. A number of hydrological controlling factors, including yearly, monthly and average monthly (climatology) precipitation and potential evaporation, and the frequency distribution of topographic index were calculated for both the source sub-basins and the gauged area. The degree of similarity of those factors was calculated by the standardised RMSE (root-mean-square error) values (SRMSE) as follows:

$$\mathrm{SRMSE} = \frac{1}{\bar{x}} \cdot \sqrt{\frac{\sum_{i=1}^{n} \left(x_i^{'} - x_i\right)^2}{n}}, \quad (1)$$

where x_i stands for the time series of a controlling factor (e.g. precipitation) of the entire gauged area, and $x_i^{'}$ stands for the time series of the same controlling factor for a single source sub-basin. A smaller SRMSE value indicates more similarity. Similarly, the SRMSE values of yearly discharge were also calculated, and plotted against the SRMSE of each controlling factor in Fig. 3a–g as black circles. The number of black circles corresponds to the number of source sub-basins. In step two, the source sub-basins were allowed to be randomly combined, and the combined area was used instead of a single sub-basin to represent the yearly discharge of the gauged area. A total of 10 000 such randomly combined areas were obtained. Their similarities in terms of hydrological controlling factors and discharge with the entire gauged area were also calculated by SRMSE values and are plotted by grey dots in Fig. 3a–g.

Figure 3 shows that all controlling factors have significant control over the similarity of the discharge dynamics. For instance, if a source sub-basin or a combined area has similar precipitation dynamics as the gauged area, its discharge is more likely to well resemble the discharge of the gauged area. On the other hand, a large deviation in any of the controlling factors is likely to mean poor discharge resemblance. Figure 3 also shows the limited ability of individual source sub-basins to capture the variation of any controlling factor of the gauged area. Combined source sub-basins can achieve much better resemblance for all controlling factors, and as they do so, they also better resemble the discharge dynamics of the gauged area.

Many controlling factors in Fig. 3 are correlated with each other; therefore, a multiple regression analysis was performed in order to identify the first-order controlling factors. Firstly a regression was made with the SRMSE of discharge as an independent variable and SRMSE of all controlling factors as dependent variables. The result showed that the best linear combination of the dependent variables was able to explain 84 % of the variations of the independent variable. If only SRMSE values of yearly precipitation and potential evaporation were used as dependent variables, they would be able to explain 82 % of the variations of the independent variable. Although the addition of topographic index as a dependent variable can increase the degree of explanation (i.e. 1 % more of the variation of the dependent variable), in this paper only yearly precipitation and potential evaporation were used as first-order controlling factors for yearly discharge.

6 Data-driven scale extrapolation

Scale extrapolation is defined as the extrapolation of hydrological parameters (e.g. discharge) from small to large scale. We denote the collection of hydrological controlling factors as X, so that

$$X = \{x_1, x_2, \ldots x_n, \}, \quad (2)$$

where $x_1, x_2, \ldots, x_n$ are individual controlling factors. The discharge, if not measured, can be estimated by X, such as

$$\hat{D} = f(X). \quad (3)$$

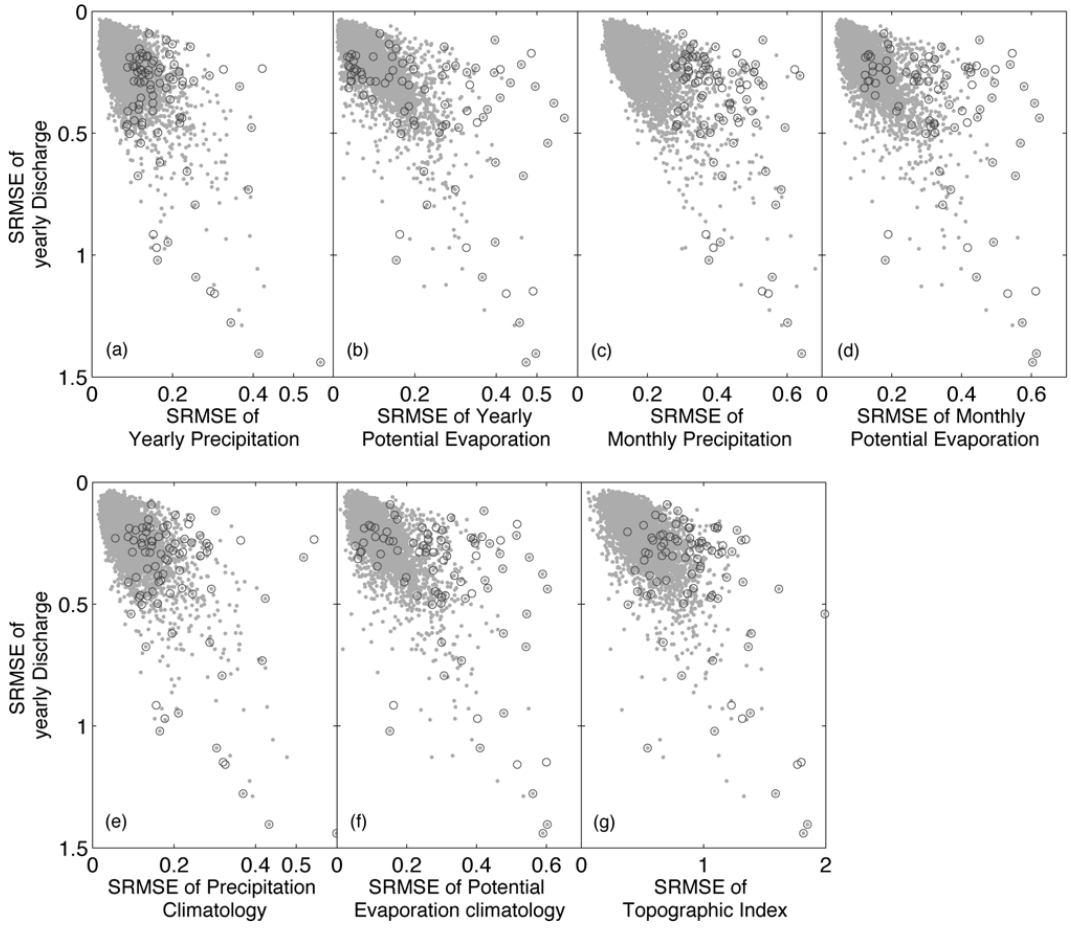

Fig. 3. The standardised root-mean-square error (SRMSE) of yearly discharge (1975–2001) calculated between sub-sets of the gauged area and the gauged area itself (y axis), plotted against SRMSE of seven hydrological controlling factors also calculated between subsets of the gauged area and the gauged area itself. Sub-sets were selected in two different ways: (1) by using individual source sub-basin alone (black circles) and (2) by randomly combing source sub-basins (grey dots). The seven hydrological controlling factors are yearly precipitation **(a)** and potential evaporation **(b)**, monthly precipitation **(c)** and potential evaporation **(d)**, precipitation **(e)**, and potential evaporation climatology **(f)** and the frequency distribution of topographic index **(g)**.

Figure 4a illustrates an ungauged basin A and three of its source sub-basins, S^1, S^2 and S^3, which have discharge data D^1, D^2 and D^3, respectively. Inside each sub-basin, a group of cells (C^1, C^2 and C^3) is selected according to the following two criteria:

1. Inside each source sub-basin, a group of cells is selected so that it can resemble the yearly precipitation and potential evaporation of the sub-basin, such that

$$X_{C^i} \approx X_{S^i},\ i = 1, 2, 3, \tag{4}$$

where X_C and X_S are the hydrological controlling factors for a cell group and for a source sub-basin, respectively. Therefore, it can be assumed that the cell group has the same discharge dynamics as the sub-basin (Fig. 4b), such that

$$\hat{D}_{C^i} \approx D_{S^i},\ i = 1, 2, 3. \tag{5}$$

2. The combination of all cell groups (i.e. the source cells) shall resemble the yearly precipitation and potential evaporation of the whole basin, such that

$$X_{(C^1+C^2+C^3)} \approx X_{\mathrm{A}}. \tag{6}$$

Therefore, the area-weighted average discharge from the source cells can be used to estimate the discharge of the whole basin (Fig. 4c), such that

$$\hat{D}_{\mathrm{A}} \approx \left[\, \hat{D}_{C^1}\ \hat{D}_{C^2}\ \hat{D}_{C^3} \,\right] \times \frac{\left[\, a_{C^1}\ a_{C^2}\ a_{C^3} \,\right]'}{\sum_{i=1}^{3} a_{C^i}}$$

$$\approx \left[\, D_{S^1}\ D_{S^2}\ D_{S^3} \,\right] \times \frac{\left[\, a_{C^1}\ a_{C^2}\ a_{C^3} \,\right]'}{\sum_{i=1}^{3} a_{C^i}}, \tag{7}$$

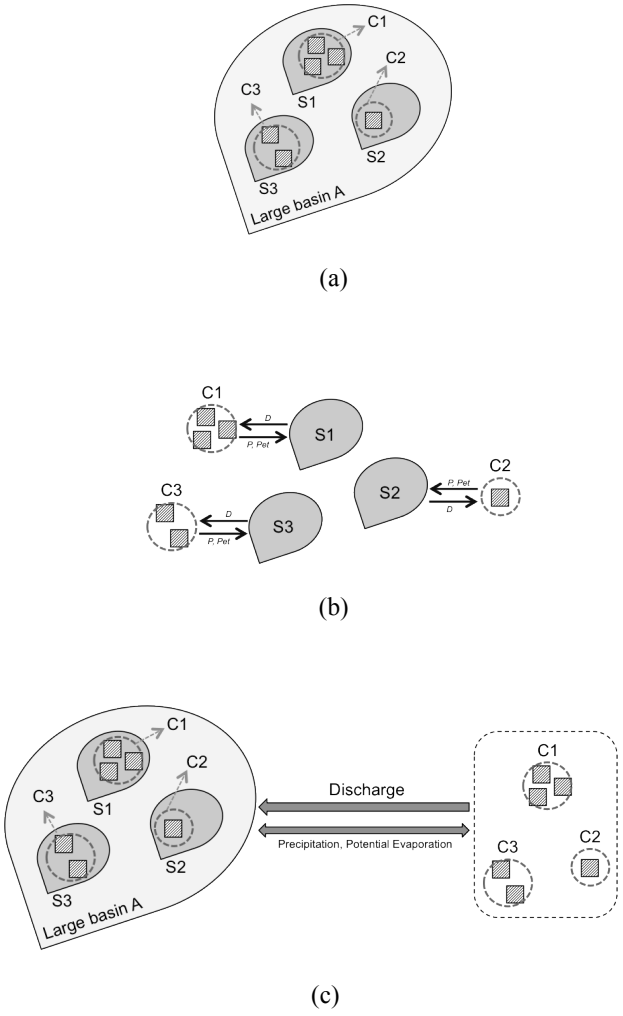

Fig. 4. Schematic illustration of the scale-extrapolation method. **(a)** An ungauged large basin A and its gauged sub-basins S^1, S^2 and S^3, each containing a group of cell C^1, C^2 and C^3, respectively. **(b)** Cell groups C^1, C^2 and C^3 can well resemble essential climate variables of their respective sub-basins; therefore, C^1, C^2 and C^3 are expected to have same discharge dynamics as their respective sub-basins. **(c)** The combination of all cell groups can well resemble essential climate variables of basin A; therefore, area-weighted discharge from all cell groups can be used to estimate the discharge of basin A.

where a_{C^i} is the area of the cell group C^i. It is important to note that basin A can have more than three source sub-basins, and it is not necessary that all source sub-basins should contribute to source cells. If a source cell is on the border of a sub-basin, only the overlapping area is used in the area weighting. In this paper, we tested the scale-extrapolation method in the gauged basin area, formed by the 100 gauged sub-basins of the Baltic Sea drainage basin. 51 gauged sub-basins between 500 and 5000 km^2 were selected as source sub-basins. Monte Carlo method was used to select

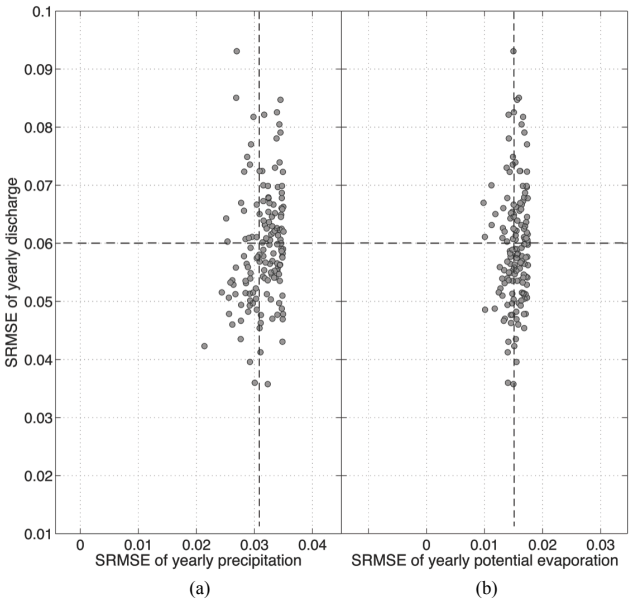

Fig. 5. SRMSE values of yearly precipitation **(a)** and potential evaporation **(b)** plotted against SRMSE values of yearly discharge, for 196 realisations of selected source cells with discharge SRMSE values less than 0.1. Dashed lines indicate mean values.

200 realisations of source cells (i.e. 200 different groups of cells that fulfil the above criteria). Each realisation of source cells was selected so that the SRMSE values for precipitation and potential evaporation do not exceed threshold values 3.5 % and 1.75 % respectively. The threshold value was selected to ensure that (1) there is a good resemblance of climate time series between selected cells and the gauged basin area and (2) a sufficient number of different cell groups can be found to equally well resemble the gauged basin area. The threshold value can be region-dependent, and it should be a function of data quality and the spatial variability of regional climate. A total of 200 area-weighted discharge time series from the 200 realisations of source cells were then derived, and their similarity with the discharge of the entire gauged area was examined by calculating the SRMSE value.

7 Result

All 200 realisations of source cells closely resemble the yearly dynamics of precipitation and potential evaporation of the gauged area with very small SRMSE values (Fig. 5). The average SRMSE for precipitation is 3.1 % with a standard deviation of 0.2 %. The average SRMSE for potential evaporation is 1.5 % with a standard deviation of 0.1 %. The average SRMSE for the 200 extrapolated yearly discharge time series is 6 % with a standard deviation of 1 %.

Figure 6 shows the quality of discharge extrapolations measured by SRMSE of yearly discharge between gauged basin area and source cells, plotted against the area ratio

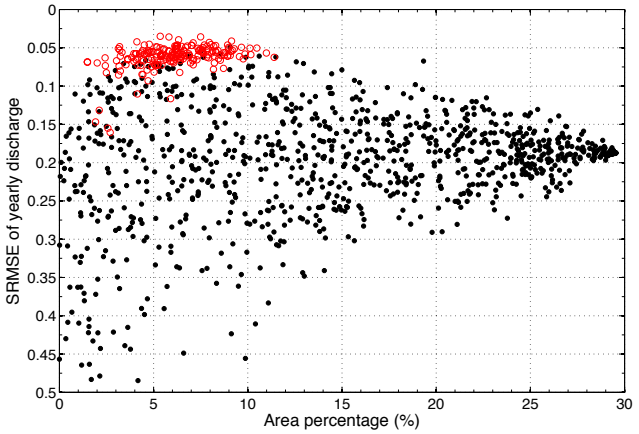

Fig. 6. Quality of discharge extrapolation measured by SRMSE of yearly discharge between gauged basin area and selected source cells, plotted against areal ratio of selected source cells to the entire gauged basin area. Two different source-cell selection methods are plotted: (1) randomly combining source sub-basins (black) and (2) using the scale-extrapolation method, i.e. allowing a sub-set of a sub-basin to be selected (red).

of source cells to the entire gauged area. For the purpose of comparison, two different source-cell selection methods are plotted: (1) randomly selected source sub-basins were combined and all cells within each sub-basin were used to form the source cells; 1000 such combinations were used, and their SRMSE values were plotted as black dots; (2) the scale-extrapolation method was used (i.e. allowing a sub-set of a source sub-basin to be selected). The SRMSE values of 200 realisations of selected source cells were plotted as red circles. Figure 6 shows that most realisations of source cells selected by the scale-extrapolation method have the area ratio between 3 % and 10 %. With such area percentages it is most probable to find a good match of climate dynamics with the gauged area. Figure 6 also shows that the area ratio of the source cells to the whole gauged area plays an important control over the extrapolation quality. It seems that when the source cells are around 5 % of the entire gauged area, there is the best chance for a good extrapolation. The largest extrapolation error occurred when the area ratio was too small.

Figure 7a and b show two examples of totally different realisations of selected source sub-basins (blue boundaries) and source cells (red). In the first example, the selected sub-basins represent precipitation and potential evaporation of the whole basin area with SRMSE values of 3.5 % and 2 % respectively (Fig. 8a and b); the extrapolated discharge well resembles discharge of the gauged area with SRMSE of 6 % (Fig. 9a). In the second example, precipitation and potential evaporation of the gauged area are represented with SRMSE values of 3.3 % and 2.6 % respectively (Fig. 8c and d), and the extrapolated discharge resembles discharge for the whole basins also with an SRMSE of 6 % (Fig. 9b).

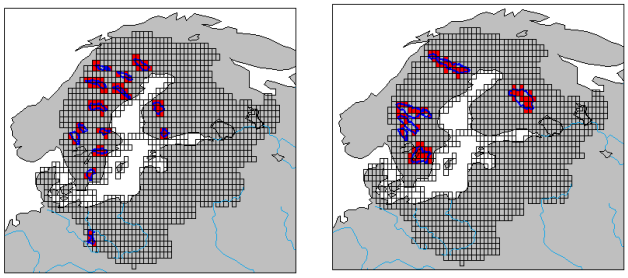

Fig. 7. (a, b) Map of the Baltic Sea drainage basin as shown by 0.5 degree STN-30p global grid cells, with two examples of the selected source basins and selected source cells for the discharge extrapolation in the Baltic Sea drainage basin.

8 Discussions and conclusion

Small-scale dynamics can have a crucial impact on large-scale hydrological responses. A fundamental problem for large-scale hydrology is the difficulty in preserving the non-linearity at small scales. The superposition principle, applicable only for linear systems, states that the response caused by two or more inputs equals the sum of the responses, which would have been caused by each input individually. In terms of hydrology, this would imply that the hydrological response of a basin (or a grid cell), under distributed inputs, could be perfectly reproduced with spatially averaged inputs. This is not valid because hydrological systems are nonlinear, so that

$$\frac{f(X_1)+f(X_2)+\ldots+f(X_n)}{n} \neq f\left(\frac{X_1+X_2+\ldots+X_n}{n}\right), \quad (8)$$

where n is the number of response units (e.g. number of cells in a basin), and $f(X_i)$ is the distributed hydrological response under distributed hydrological controlling factor (X_i) as defined in Sect. 4. Equation (8) has an interesting implication: if two basins differ significantly in size, even if they share similar average hydrological controlling factors, they may have different discharge dynamics.

The result from this paper illustrates that it is impossible to use a single sub-basin or a single cell to represent the average dynamics of the gauged area of the Baltic Sea drainage basin. It is always necessary to use spatially discrete and scattered sub-regions to represent unique patterns of the hydrological controlling factors, even though the area of the sub-regions can be as small as 1.5 % of the gauged area. For the Baltic Sea drainage basin, a fairly accurate approximation can be achieved by relaxing Eq. (8), so that

$$\frac{f(X_1)+f(X_2)+\ldots+f(X_n)}{n} \approx f\left(\frac{X_1+X_2+\ldots+X_{m1}}{m1}\right)$$

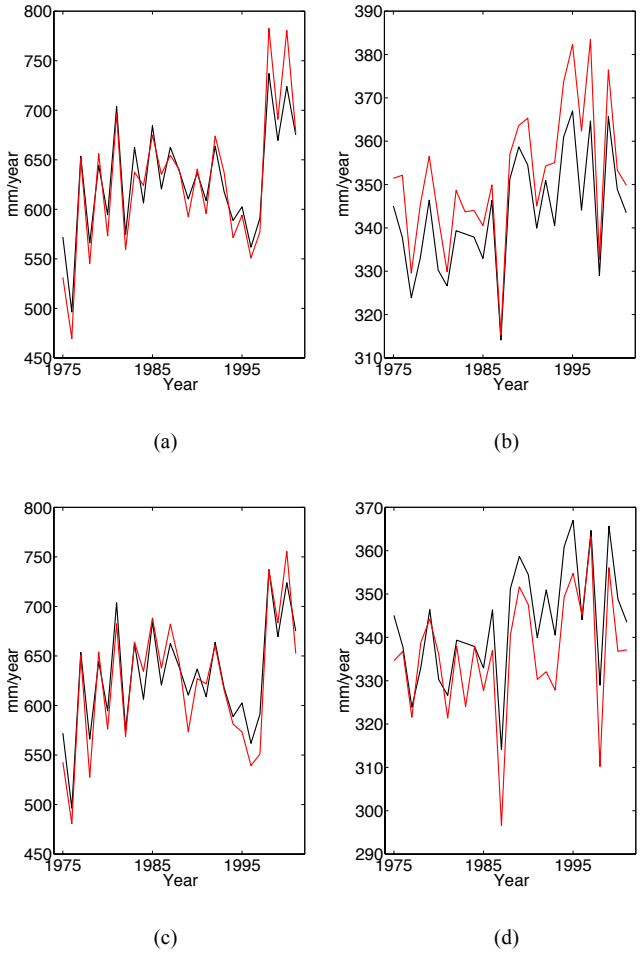

Fig. 8. Yearly precipitation **(a)** and potential evaporation **(b)** of the selected source cells (red) and of the entire gauged area (black) for example 1 (Fig. 7a). Yearly precipitation **(c)** and potential evaporation **(d)** of the selected source cells (red) and of the entire gauged area (black) for example 2 (Fig. 7b).

$$+ f\left(\frac{X_{m1+1} + X_{m1+2} + \ldots + X_{m2}}{m2}\right) + \ldots$$
$$+ f\left(\frac{X_{mk+1} + X_{mk+2} + \ldots + X_{n}}{mk}\right). \quad (9)$$

Equation (9) categorises the n cells of a basin into k groups. Cells within each group have correlated controlling factors, and therefore may be considered quasi-linear, such that the hydrological response from a cell group can be well approximated by using an average input. Cell groups are, however, mutually independent of each other, and a minimal number of cell groups are needed to capture the variability of the whole basin. Equation (9) lends theoretical support to the scale-extrapolation method. Precipitation time series among selected source cells are mutually independent (Fig. 2), and each selected source sub-basin represents the average hydrological controlling factors for a certain region

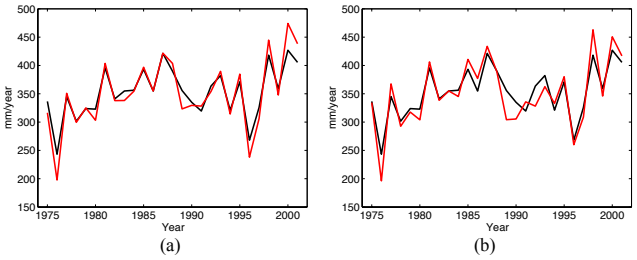

Fig. 9. Yearly discharge from the gauged basin area (black) and extrapolated discharge (red) using selected source cells from example 1 **(a)** and example 2 **(b)**.

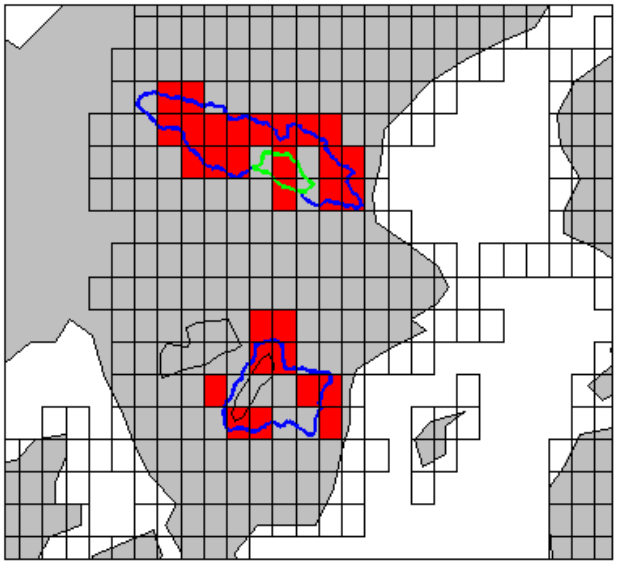

Fig. 10. Sub-basins selected for inverse extrapolation. The green-outlined small basin, with an area of 2250 km^2, is found to have similar climate conditions as the selected cells (red, 21 228 km^2 in total) of two larger sub-basins.

within the gauged area, which can be regarded as quasi-linear in its hydrological dynamics.

Figure 6 showed that out of 200 realisations of the extrapolated discharge, only 6 had SRMSE values of discharge of more than 10 %. Four of those relatively large extrapolation errors occurred when the area ratio of source cells to gauged area was too small (i.e. between 2 % and 3 %), while another two occurred when the area ratio was around 4 % and 6 %, respectively. This result further lends support to the fact that nonlinearity exists at large scale and even at yearly timescale. Although those small source-cell areas can perfectly resemble the average climate of the gauged area, they are unable to resemble the discharge dynamics in a good way, because they do not cover the minimum number of unique patterns required in order to preserve the nonlinearity, as shown in Eq. (9).

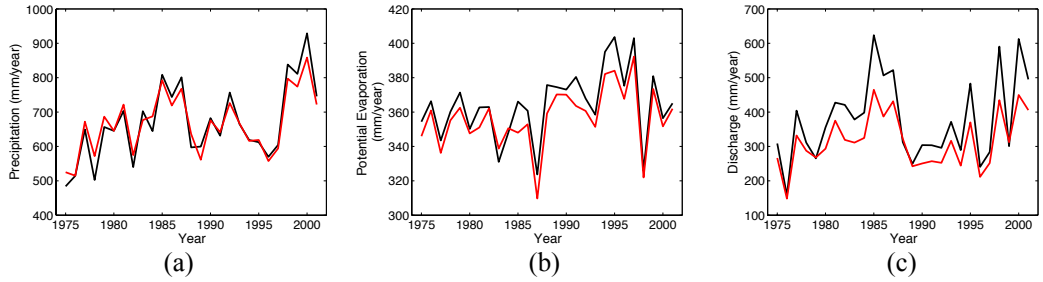

Fig. 11. Values of yearly precipitation **(a)**, potential evaporation **(b)** and discharge (**c)** of the small sub-basin (outlined in green in Fig. 10), plotted by black lines, and a larger area (red cells in Fig. 10), plotted by red lines.

An inverse extrapolation was made to illustrate the existence of nonlinearity on a yearly timescale further. The inverse extrapolation, similar to the scale-extrapolation method, tries to match the hydrological controlling factors of a number of sub-basins to a single small sub-basin, instead of the large basin. A small sub-basin with a size of 2250 km^2 (Fig. 10, outlined in green) was selected, and source cells (Fig. 10, red cells) from two other sub-basins (Fig. 10, blue outlined) with a total size of 21 228 (or 10 times bigger) were found to resemble the yearly precipitation and potential evaporation of the small sub-basin well, with SRMSEs of 4.7 % and 2.6 % respectively (Fig. 11a and b). However, the discharge differs between the small basin and the larger region by 21 % (Fig. 11c). Of course, this is only one example, and more thorough tests should be made, preferably with climate data sets of higher resolution. The results of this paper showed that a minimum of 5 % of the basin area is needed to be able to account for the nonlinearity of the system; 5 %–10 % appears to be the area percentage for which the best extrapolation quality can be expected (Fig. 6). This percentage is expected to increase with finer timescales and to change with different climate and hydrological regimes.

A new data-driven scale-extrapolation method was proposed to estimate annual water resources for large river basins. The new method builds upon the fact that the dynamic interaction between climate and hydrology of a large river basin can be equally well resembled by multiple small regions, each characterized by a number of small river basins that typically give around 5 % areal percentage of the large basin. Therefore, those multiple small regions can provide an ensemble of water resource estimations for the large basin. The new method, being purely data-based, makes it possible for regional water resource estimations to benefit from a multitude of readily available measurements from small river basins.

The scale-extrapolation method provides both new methodology and new data into the field of large-scale hydrology. It allows regional water resources to be estimated directly from small river basins that are typically 1–2 orders of magnitude smaller and therefore better preserve the small-scale dynamics and nonlinearity, which are vital for credible predictions. The extrapolation is modelling-free, and therefore the estimation is free of modelling uncertainties that usually contribute significantly to large-scale estimation uncertainties. The method is not sensitive to the bias of the climate data set because the climate data set is only used for sub-basin selection and not directly for extrapolation.

The scale-extrapolation methods made it possible to study the interaction between climate and hydrology, and the climate change impact in ungauged or partially gauged large river basins from data alone. At the same time, the method offers ensemble predictions that have the potential of bracketing the estimation uncertainty. Because the scale extrapolation uses completely different data and method compared to the modelling approach, it provides a unique opportunity to be compared with modelling results.

Acknowledgements. This paper is funded by the Swedish Research Council grant 621-2012-3903 and grant 214-2012-373 from the Swedish Research Council for Environment, Agriculture Sciences and Spatial Planning. Discharge data were provided by the Global Runoff Data Centre, 56068 Koblenz, Germany, and the Swedish Meteorological and Hydrological Institute. The author had discussions with Fritjof Fagerlund and Anna Kauffeldt and received valuable insights.

Edited by: S. Thompson

References

Allen, R. G., Pereira, L. S., Raes, D., and Smith, M.: Crop evapotranspiration – guidelines for computing crop water requirements, FAO Irrigation and Drainge Paper, FAO, 1998.

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

Arnell, N. W.: Effects of IPCC SRES* emissions scenarios on river runoff: a global perspective, Hydrol. Earth Syst. Sci., 7, 619–641, doi:10.5194/hess-7-619-2003, 2003.

Arnell, N. W.: Climate change and global water resources: SRES emissions and socio-economic scenarios, Global Environ. Change, 14, 31–52, 2004.

Blöschl, G., Sivapalan, M., Wagener, T., Viglione, A., and Savenije, H.: Runoff Prediction in Ungauged Basins: Synthesis Across Processes, Places and Scales, Cambridge University Press, 2013.

Brooks, P. D., Troch, P. A., Durcik, M., Gallo, E., and Schlegel, M.: Quantifying regional scale ecosystem response to changes in precipitation: Not all rain is created equal, Water Resour. Res., 47, W00J08, doi:10.1029/2010WR009762, 2011.

Budyko, M. I.: Climate and life, New York, Academic Press, 508 pp., 1974.

GRDC: Global Runoff Data Centre, Global Runoff Data Centre, available at: http://grdc.bafg.de (last access: 15 October 2013), 2012.

Huang, M., Liang, X., and Liang, Y.: A transferability study of model parameters for the variable infiltration capacity land surface scheme, J. Geophys. Res., 108, 8864, doi:10.1029/2003JD003676, 2003.

Kim, U. and Kaluarachchi, J. J.: Application of parameter estimation and regionalization methodologies to ungauged basins of the Upper Blue Nile River Basin, Ethiopia, J. Hydrol., 362, 39–56, 2008.

Kokkonen, T. S., Jakeman, A. J., Young, P. C., and Koivusalo, H. J.: Predicting daily flows in ungauged catchments: model regionalization from catchment descriptors at the Coweeta Hydrologic Laboratory, North Carolina, Hydrol. Process., 17, 2219–2238, 2003.

L'vovich, M. I.: World Water Resouces and Their Future, 1979.

McIntyre, N., Lee, H., Wheater, H., Young, A., and Wagener, T.: Ensemble predictions of runoff in ungauged catchments, Water Resour. Res., 41, W12434, doi:10.1029/2005WR004289, 2005.

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.

NCDC: Global Surface Summary of the Day, National Climatic Data Center (NCDC), Asheville, NC, available at: http://www.ncdc.noaa.gov/cgi-bin/res40.pl?page=_climvisgsod.html (last access: 1 February 2011), 2011.

Simmons, A., Uppala, S., Dee, D., and Kobayashi, S.: ERA-Interim: New ECMWF reanalysis products from 1989 onwards, ECMWF Newsletter No. 110, 25–35, 2007.

Troch, P. A., Martinez, G. F., Pauwels, V. R. N., Durcik, M., Sivapalan, M., Harman, C., Brooks, P. D., Gupta, H., and Huxman, T.: Climate and vegetation water use efficiency at catchment scales, Hydrol. Process., 23, 2409–2414, doi:10.1002/hyp.7358, 2009.

Uppala, S. M., Kållberg, P. W., Simmons, A. J., Andrae, U., Bechtold, V., Da Costa, Fiorino, M., Gibson, J. K., Haseler, J., Hernández, A., Kelly, G. A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R. P., Andersson, E., Arpe, K., Balmaseda, M. A., Beljaars, A. C. M., Berg, L. Van De, Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B. J., Isaksen, L., Janssen, P. A. E. M., Jenne, R., Mcnally, A. P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N. A., Saunders, R. W., Simon, P., Sterl, A., Trenberth, K. E., Untch, A., Vasiljevic, D., Viterbo, P., and Woollen, J.: The ERA-40 re-analysis, Q. J. Roy. Meteor. Soc., 131, 2961–3012, 2005.

USGS (US Geological Survey): HYDRO 1K Elevation Derivative Database, available at: http://edc.usgs.gov/products/elevation/gtopo30/hydro/index.html (last access: 30 April 2009) from the Earth Resources Observation and Science (EROS) Data Center (EDC), Sioux Falls, South Dakota, USA, 1996.

Voepel, H., Ruddell, B., Schumer, R., Troch, P. A., Brooks, P. D., Neal, A., Durcik, M., and Sivapalan, M.: Quantifying the role of climate and landscape characteristics on hydrologic partitioning and vegetation response, Water Resour. Res., 47, W00J09, doi:10.1029/2010WR009944, 2011.

Vörösmarty, C. J., Moore, B., Grace, A. L., Gildea, M. P., Melillo, J. M., Peterson, B. J., Rastetter, E. B., and Steudler, P. A.: Continental scale models of water balance and fluvial transport: An application to South America, Global Biogeochem. Cy., 3, 241–265, 1989.

Vörösmarty, C. J., Green, P., Salisbury, J., and Lammers, R. B.: Global water resources: Vulnerability from climate change acid population growth, Science, 289, 284–288, 2000a.

Vörösmarty, C. J., Fekete, B. M., Meybeck, M., and Lammers, R. B.: Globalsystem of rivers: Its role in organizing continental land mass and defining landto-ocean linkages, Global Biogeochem. Cy., 14, 599–621, 2000b.

Vörösmarty, C. J., Lettenmaier, D., Leveque, C., Meybeck, M., Pahl-Wostl, C., Alcamo, J., Cosgrove, W., Grassl, H., Hoff, H., Kabat, P., Lansigan, F., Lawford, R., and Naiman, R.: Humans transforming the global water system, AGU Eos Transactions, 85, 509–514, 2004.

Wagener, T., Sivapalan, M., Troch, P., and Woods, R.: Catchment classification and hydrologic similarity, Geography Compass, 1, 901–931, 2007.

Weedon, G. P., Gomes, S., Viterbo, P., Österle, H., Adam, J. C., Bellouin, N., Boucher, O. and Best, M.: The WATCH forcing data 1958–2001: a meteorological forcing dataset for land surface- and hydrological-models, WATCH Technical Report, 2010.

Widén-Nilsson, E., Halldin, S., and Xu, C.-Y.: Global water-balance modelling with WASMOD-M: Parameter estimation and regionalisation, J. Hydrol., 340, 105–118, 2007.

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

Wulff, F. V., Rahm, L. A., and Larsson, P.: A Systems Analysis of the Baltic Sea, Springer, 2001.

Xu, C.-Y.: Estimation of parameters of a conceptual water balance model for ungauged catchments, Water Resour. Manage., 13, 353–368, 1999.

Xu, C.: Testing the transferability of regression equations derived from small sub-catchments to a large area in central Sweden, Hydrol. Earth Syst. Sci., 7, 317–324, doi:10.5194/hess-7-317-2003, 2003.