Streamflow Data from Small Basins: A Challenging Test to High-Resolution Regional Climate Modeling

KERSTIN STAHL
Department of Geosciences, University of Oslo, Oslo, Norway, and Institute of Hydrology, University of Freiburg, Freiburg, Germany
LENA M. TALLAKSEN AND LUKAS GUDMUNDSSON
Department of Geosciences, University of Oslo, Oslo, Norway
JENS H. CHRISTENSEN
Danish Meteorological Institute, Copenhagen, Denmark

(Manuscript received 23 August 2010, in final form 17 February 2011)

ABSTRACT
Land surface models and large-scale hydrological models provide the basis for studying impacts of climate and anthropogenic changes on continental- to regional-scale hydrology. Hence, there is a need for comparison and validation of simulated characteristics of spatial and temporal dynamics with independent observations. This study introduces a novel validation framework that relates to common hydrological design measures. The framework is tested by comparing anomalies of runoff from a high-resolution climate-model simulation for Europe with a large number of streamflow observations from small near-natural basins. The regional climate simulation was performed as a “poor man’s reanalysis,” involving a dynamical downscaling of the 40-yr European Centre for Medium-Range Weather Forecasts Re-Analysis (ERA-40) with the Danish “HIRHAM5” model. For 19 different anomaly levels, two indices evaluate the temporal agreement (i.e., the occurrence and frequency of dry and wet events based on daily anomalies), whereas two other indices compare the interannual variability and trends based on annual anomalies. Benchmarks on each index facilitated a comparison across indices, anomaly levels, and basins. The lowest agreement of observed and simulated anomalies was found for dry anomalies. Weak to moderately wet anomalies agreed best, but agreement dropped again for the wettest anomalies. The results could guide the decision on thresholds if this regional climate model were used for the assessment of climate change scenario impacts on flood and drought statistics. Indices vary across Europe, but a gradient with decreasing correspondence between observed and simulated runoff characteristics from west to east, from lower to higher elevations, and from fast to slowly responding basins can be distinguished. The suggested indices can easily be adapted to other study areas and model types to assist in assessing the reliability of predictions of hydrological change.

1. Introduction
Large-scale gridded models, including global (general circulation) and regional climate models and large-scale hydrological models, are employed for a variety of purposes in hydrology and related disciplines. They provide spatial simulations of hydrological variables such as soil moisture, runoff, and river discharge for historical records and can be used to simulate the response of the hydrological cycle to future global change, that is, climate scenarios and human impacts. Examples of global or continental-scale mapping of recent and predicted hydrological change include changes in mean streamflow globally (e.g., Milly et al. 2005; Dai et al. 2009) and extremes (floods and low flows) in Europe (Lehner et al. 2006) and globally (Hirabayashi et al. 2008).

To assess the limits of interpretability of model simulations, evaluation against observational data is crucial. Such comparisons are usually carried out in the model

Corresponding author address: Kerstin Stahl, Dept. of Geosciences, University of Oslo, P.O. Box 1047, Blindern, NO-0316 Oslo, Norway.
E-mail: kerstin.stahl@geo.uio.no
DOI: 10.1175/2011JHM1356.1

© 2011 American Meteorological Society
development phase—for example, for parameter tuning in hydrological models (e.g., Hunger and Döll 2008), as perturbed physics experiments that aim to isolate the cause of a specific climate change signal (e.g., Gedney et al. 2006) or as part of a model intercomparison experiment (e.g., Haddeland et al. 2011). The models are commonly evaluated against long-term annual or monthly mean values of discharge in large rivers, which is important in terms of future changes in water resources and freshwater input to the oceans. Changes in the frequency of extreme events are also a major concern, however, and changes in extreme events rather than long-term mean values may be more relevant to hydrological design. So far, such changes have been assessed primarily with catchment-scale hydrological models calibrated to past discharge and projected into the future by simple delta-change scenarios or statistically downscaled climate variables. The delta change in particular does not take into account potentially important propagation of changes to climate dynamics, such as the length of wet and dry spells, and both approaches are restricted to gauged catchments. As the spatial refinement of regional climate models continues to improve, spatial patterns of the hydrological response (particularly of extremes) to climate forcing are used directly for regional planning and adaptation. This requires, however, that we have sufficient trust in the models—that is, the models need to be validated at the spatial and temporal scale of interest.

Hydrological observations that can be used for model validation are commonly point measurements (soil moisture) or they integrate over basin areas (streamflow). The spatial and temporal scale, coverage, and density of available observations vary strongly. Nevertheless, modeled soil moisture anomalies have been compared with point measurements (e.g., Balsamo et al. 2009), whereas runoff mostly has been routed along river channels and then compared with gauged discharge for large river basins (for Europe; e.g., Hagemann et al. 2009). A commonly used dataset for comparison with observed streamflow from large continental rivers is the archive assembled by the Global Runoff Data Centre (GRDC). This archive consists of records with different length, however, with many ending in the 1980s or 1990s (Widén-Niellson 2007; Hunger and Döll 2008), and there is a need for updated observations. In addition, withdrawals and regulations influence streamflow in large rivers (Döll et al. 2009). Therefore, only annual values or long-term averages can be compared unless naturalized series are employed. In the case of soil moisture anomalies, the high spatial variability of soil properties restricts conclusions on model errors. Notable differences between modeled and observed streamflow dynamics, on the other hand, may be difficult to attribute to separate hydrological processes in the model—for example, either runoff generation or channel routing (Balsamo et al. 2009). Hence, additional high-quality high-resolution data sources are critical to improve the identification and validation of model theories related to runoff generation, event characteristics, and spatial variability in large-scale models. A limitedly used resource is records from small- and medium-sized near-natural basins. One reason is that few such river archives exist at the large scale. Those that do are often subject to restriction on license and usage (Hannah et al. 2011). Examples such as the U.N. Educational, Scientific and Cultural Organization (UNESCO) Flow Regimes from International Experimental and Network Data (FRIEND) European Water Archive (EWA) and the U.S. Hydro-Climatic Data Network of stream gauges, however, have contributed to a better understanding of large-scale processes at the land surface, including changes across a range of scales (e.g., Krakauer and Fung 2008; Stahl et al. 2010). In the United States, streamflow data from small basins have been used to calibrate and evaluate land surface models at the grid scale (Troy et al. 2008; Lohmann et al. 2004).

In catchment hydrology, the way models and observations are compared has recently been questioned. Gupta et al. (2008) argue for a “signature”-based evaluation of models because common metrics express model performance relative to weak benchmarks. The widely used Nash–Sutcliffe efficiency, for example, uses the mean of the observations as a benchmark. The correlation coefficient compares only relative fluctuations. Considering that land surface models coupled to climate models usually suffer from a strong bias in the atmospheric and ultimately in the hydrological variables, these metrics appear to have limited diagnostic value. Signatures, however, are indices reflecting the dynamic system response behavior and patterns (Gupta et al. 2008). While their use has been promoted mainly for the detection of model structure deficiencies and the improvement of model parameterization (e.g., Yilmaz et al. 2008), the concept may also motivate model diagnostics useful for hydrological change assessment. In particular, as models are increasingly expected to provide estimates about changes in the frequency and duration of extreme events, signatures related to these metrics should be developed.

The main objective of this study is to introduce a set of exceedance-based indices for benchmarking simulations of hydrological event dynamics. The chosen approach is tailored to hydrological design questions, where events below (low flow/drought) or above (high flow/flood) a particular anomaly level (threshold) are considered for
water resources management and planning. As models at all scales often fail to simulate extremes well, it is envisaged that this approach would provide insight as to at which anomaly level a particular model can be trusted. Whereas the approach and indices chosen would be suitable to other similar hydrological model validations, this study tested them in a comparison of the simulations from a high-resolution climate model over Europe. A second objective was to test to which degree this dynamical downscaling product captures the complexity of hydrological dynamics across Europe.

The paper begins with a description of the observations dataset and the regional climate simulations and then explains in detail the approach for comparison of observed and simulated anomalies, the anomaly transformation, and the indices derived. In addition, benchmarks are introduced that facilitate comparison across indices, anomaly levels, and basins. The discussion critically evaluates the results and approach chosen and concludes with an assessment of its suitability for model validation.

2. Data

a. Observations

The study was based on observed streamflow records for northern, western, and central Europe from the EWA and additional streamflow records collected for the European Union (EU) Integrated Project Water and Global Change (WATCH) (http://www.eu-watch.org). The EWA is the streamflow archive of the European FRIEND project, which is part of the FRIEND initiative of the UNESCO International Hydrological Programme. It includes records from relatively small catchments that represent local conditions and relatively natural flows. The database was recently updated to 2004 (Stahl et al. 2008) and is now held at the GRDC (http://grdc.bafg.de), which also manages data requests. The selection for each country was done by the national agencies, which were also responsible for data quality control. In addition, all series were visually checked for detectable inhomogeneities or quality problems during low flows (Stahl et al. 2010). For this study, only records covering the period from 1961 to 2004 without gaps of more than a few days were chosen.

Basin boundary estimates were available from CCM2, the second version of the Catchment Characterisation and Modelling (CCM) River and Catchment database for Europe (Vogt et al. 2007). While catchment area was provided by the national agencies that contributed to the dataset, basin properties such as elevation, elevation range, or slope were derived from CCM2. Figure 1a shows the geographic location of the basins with the catchment outlines upstream of the gauging stations. The distribution of basin areas depends on the national gauging networks, with generally larger basin areas found in France and Scandinavia and smaller basin areas in countries such as Germany and the Alpine countries. The majority of the basins have an area of less than 500 km² (Fig. 1b). The mean elevations of the basins range from close to sea level to 400 m MSL in the flat areas of the southern United Kingdom, northern France and Germany, Denmark, and eastern Scandinavia, whereas in the western and northern United Kingdom, Norway, and central Europe, higher elevations are found. Some basins from the Alpine rim with mean elevations above 1800 m MSL are also included. The distribution of basin elevations reflects the hypsography of the model domain for elevations of >400 m MSL; lowland areas below 400 m MSL are underrepresented (Fig. 1b), however.

b. Regional-climate-model simulations

A poor-man’s reanalysis (PMR) with a regional climate model (RCM) provided the simulations of the land surface hydrology used in this study. The term “poor man” refers to the method of dynamical downscaling from a larger-scale reanalysis. Berg and Christensen (2008) used the Danish RCM HIRHAM5 [the model name HIRHAM (version 5) comes from combining the High-Resolution Limited-Area Model (HIRLAM) and the German “ECHAM” Model], which was nested into the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40) and ERA Interim reanalysis data (Uppala et al. 2005) for a domain covering most of Europe at a high spatial resolution of 0.12° (~200 km²) (Fig. 1). The output consists of atmospheric and surface variables. In contrast to the typical RCM setup nested in a reanalysis, the PMR keeps the atmospheric circulation inside the domain close to the reanalysis by a daily adjustment toward the larger-scale reanalysis but without an actual data assimilation step—hence the term poor man (for details, see Berg and Christensen 2008). The approach may be seen as a variation of the so-called spectral nudging technique (von Storch et al. 2000; Denis et al. 2002).

HIRHAM5 uses a land surface scheme similar to that of the ECHAM5 GCM (Roeckner et al. 2003). Soil water changes are simulated as a single bucket with geographically varying maximum field capacity. The variables “runoff” and “drainage” were used in this study. HIRHAM5 calculates these variables according to the saturation excess runoff generation scheme by Dümenil and Todini (1992). It uses a storage capacity distribution curve to mimic subgrid-scale partial saturation of the soil and allows for a fraction of water from the soil water budget to become runoff even if the cell’s bucket is not entirely filled. Drainage occurs independent of the

Unauthenticated | Downloaded 08/24/22 08:06 AM UTC
water input and depends on soil water content. Conceptually, the processes modeled at the grid cell level are hence very similar to many lumped catchment-scale hydrological models that are used in small basins that do not require channel routing. Applications of the ECHAM5/HIRHAM5 land surface scheme that aimed to compare simulations with river flow in large basins commonly subject the two fluxes (runoff and drainage) that leave a grid cell to a channel routing scheme (e.g., Hagemann and Dumenil Gates 2003).

3. Methods

a. Corresponding observed and simulated runoff time series

For each observed streamflow series $Q_{\text{obs}}(t)$, a corresponding daily time series of simulated basin runoff $Q_{\text{sim}}(t)$ was obtained as

$$Q_{\text{sim}}(t) = \sum_{i=1}^{n} \frac{[\text{Runoff}_i(t) + \text{Drainage}_i(t)]A_i}{A},$$

where $i$ are the individual grid cells of the total number of $n$ cells that contribute to a particular basin. As some grid cells are only partly within the basin boundary, each cell’s value was weighted by the fraction of its area $A_i$ that contributes to the entire basin area $A$.

The small basins included in the EWA (Fig. 1) encompass only a few model grid cells. With the weighted overlay, however, a 500-km$^2$ basin may use information from up to nine grid cells. In basins of this size, runoff response is fast and the concentration time can be expected to be on the order of hours to a few days. Therefore, routing is not considered to be necessary. To account for potential surface runoff storage and attenuation (which the model does not consider), however, daily observed streamflow and modeled basin outflow

![Fig. 1. (a) Spatial distribution of basins with streamflow observations and basin elevation, (b) distribution of basin areas, and (c) distribution of mean basin elevations and grid cell elevations of model domain.](image-url)
(from here on referred to as “simulated streamflow”) were smoothed by a 7-day backward averaging.

As land surface properties in regional climate models are usually derived from generalized maps, there is some uncertainty as to what degree surface and subsurface properties—such as topography, soil type, and waterholding capacity—are resolved by the model. To minimize the error due to resolution or spatial bias, the mean elevation of the basins was compared with the mean elevation of the model grid cells composing each basin. The difference was used as an indicator of whether the basin properties were sufficiently resolved by the model, and basins with a deviation of >300 m were not included in the study. Such large differences were found for only a few basins at the edge of the domain in northern Norway and along the northern rim of the Alps, where there is a strong relief. For 75% of the basins, the elevation difference is less than 50 m. The influence of a potential bias in local climate from simulated regional atmospheric variables is minimized by the comparison of anomalies rather than absolute values (see next section).

b. Nonparametric anomaly time series

Observed and simulated streamflow series, \( Q_{\text{obs}} \) and \( Q_{\text{sim}} \), were transformed into nonparametric anomalies (i.e., the percentiles of their empirical distribution on each calendar day). The reference time series for the transformation were the daily empirical exceedance frequencies from 1961 to 2004. The transformed time series of observed streamflow anomalies \( Q_{\text{obs}}' \) and simulated anomalies \( Q_{\text{sim}}' \) are expressed as percentiles (ranging from 0% to 100%). They are comparable to commonly used hydrological indices, such the Q10 (\( Q = 10\% \) that describes low flows, Q50 for the median flow, and Q90 for high flows with respect to a particular day of the year within the reference period. This transformation addresses two issues. First, it minimizes the effect of a potential bias in the magnitude of the water fluxes. Second, it relates directly to exceedance-based concepts of operational hydrology and hydrological design. The same approach is used in the U.S. Geological Survey (USGS) WaterWatch and real-time streamflow and drought monitoring systems (USGS 2010). Note that this study follows the nonexceedence notation common in North America, whereas in Europe the reverse exceedence notation, with Q90 being low flows and Q10 being high flows, is commonly used. Figure 2 shows an example of observed and simulated anomalies for the river Skjern at Alergaard in Denmark.

c. Similarity indices

Four indices were chosen for the comparison of simulated and observed anomalies. Two are derived from daily and two from annual time series (Table 1). The indices are based on the agreement of the two series \( Q_{\text{obs}}' \) and \( Q_{\text{sim}}' \) below or above a particular anomaly level \( z \). They thus allow the evaluation of different properties related to dry and wet spells (events). The two indices based on daily agreement compare directly the event characteristics, whereas the two indices based on the annual summary of event characteristics compare interannual variability and changes with time.

All indices were calculated for 19 anomaly levels—that is, for levels from \( z = 5 \) to \( z = 95 \) at intervals of 5. The indices are based on the similarity of the binary indicator series \( I_{z,\text{obs}} \) and \( I_{z,\text{sim}} \). For dry anomalies (i.e., anomaly levels below the 50th percentile: \( 0 < z \leq 50 \)), \( I \) is defined as

\[
I_z(t) = \begin{cases} 
1 & \text{for } Q_z'(t) \leq z \\
0 & \text{for } Q_z'(t) > z
\end{cases}
\]  

(2a)

For wet anomalies (i.e., anomaly levels above the 50th percentile: \( 50 < z \leq 100 \)), \( I \) is defined as

\[
I_z(t) = \begin{cases} 
1 & \text{for } Q_z'(t) > z \\
0 & \text{for } Q_z'(t) \leq z
\end{cases}
\]  

(2b)

Figure 2 schematically illustrates the derivation of indicator series for \( z = 10 \) and \( z = 30 \) (dry anomalies below these levels) and for \( z = 70 \) (wet anomalies above this level).

The first two indices relate directly to the daily agreement of the two indicator time series \( I_{z,\text{obs}} \) and \( I_{z,\text{sim}} \). The agreement index \( A \) describes the overall agreement for each anomaly level \( z \), with

\[
A_z = \left( \frac{d}{m} \right) - p_z^2
\]

(3)

where \( d \) is the observed agreement (i.e., the number of days with \( I_{z,\text{obs}} = 1 \) and \( I_{z,\text{sim}} = 1 \)), \( m \) is the total number of days in the series, and \( p_z = \Sigma I_z/m \) is the probability of \( I_z = 1 \) for each anomaly level (e.g., \( p_z = 0.3 \) for \( z = 30 \) and \( z = 70 \)). This index was motivated by the kappa statistic (Cohen 1960), which corrects the similarity counts by the theoretically expected number of counts of two random binary series \( p_z^2 \). The kappa statistic also considers the agreement of zeros, however. This is not necessary here because, by definition, the overall number of ones and zeros in \( I_{\text{obs}} \) and \( I_{\text{sim}} \) for a particular anomaly level \( z \) are equal. In addition, kappa has been found to be sensitive to the number of ones and zeros, which would hamper comparison between different anomaly levels. The agreement index allows a comparison between the observed and simulated timing of events.
The ratio of frequencies of events $F$ relates to the frequency of occurrence (i.e., number of events) of the anomalies and is defined as

$$F_z = \frac{r_{z,\text{sim}}}{r_{z,\text{obs}}},$$

where $r_{z,\text{sim}}$ is the number of events for a given anomaly level (where an event is defined as a run with one or more consecutive days with $I_{z,\text{sim}} = 1$). Similarly, $r_{z,\text{obs}}$ is the number of runs with one or more consecutive days with $I_{z,\text{obs}} = 1$. A higher number of events indicates a more variable hydrological system, whereas a lower number indicates a more persistent hydrological system (with respect to a given threshold $z$). The number of runs and run lengths over and under a given threshold are commonly used for flood and drought statistics (e.g., Zelenhasic and Salvai 1987; Tallaksen et al. 1997; Madsen and Rosbjerg 1997).

The last two indices focus on the agreement of the interannual and mid- to long-term variability. They are based on annual time series $L_z$, which are defined as the total (cumulative) number of days per year with $I_z = 1$. The correlation index $R$ is the Pearson correlation coefficient between two time series $L_{z,\text{obs}}$ and $L_{z,\text{sim}}$. It describes the similarity in the year-to-year fluctuations.

As the anomalies are based on the respective percentiles of the observed and modeled time series, the index compares the ranking of the years during the study period, that is, it compares the order of wet and dry years. The trend index $T$ compares trends in the annual series $L_{z,\text{obs}}$ and $L_{z,\text{sim}}$ over the study period. The trend index is based on the direction (different, same positive, or same negative) and magnitude. Magnitude is defined by the absolute difference between the slopes of the Kendall–Theil robust lines $S_{z,\text{obs}}$ and $S_{z,\text{sim}}$ (Theil 1950) derived from each standardized annual time series $L_{z,\text{obs}}$ and $L_{z,\text{sim}}$. The categorical index values are given in Table 2. This index hence compares any transient changes that may be present in the time series.

d. Benchmarks

To facilitate a comparison across the four indices, 19 anomaly levels, and 318 basins, pass/fail benchmarks were introduced. The benchmark for each index was chosen such that one-half of the values of a given index pass/fail (for all 318 basins and all 19 anomaly levels pooled together). For indices $A$ and $R$, the benchmark is hence the median of the respective index. For the ratio of the frequency of events (runs ratio) $F$, the benchmark had to be defined as a range (i.e., as an upper and lower

<table>
<thead>
<tr>
<th>Index name:</th>
<th>$A$</th>
<th>$F$</th>
<th>$R$</th>
<th>$T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Agreement of events</td>
<td>Ratio of frequency of events</td>
<td>Correlation coefficient</td>
<td>Trend direction and difference</td>
</tr>
<tr>
<td>Possible range</td>
<td>$-1.0$ to $1.0$</td>
<td>$0.0$–$m/2$</td>
<td>$-1.0$ to $1.0$</td>
<td>$0$–$5$</td>
</tr>
<tr>
<td>Optimum value</td>
<td>$1$</td>
<td>$1$</td>
<td>$1$</td>
<td>$3$–$5$</td>
</tr>
</tbody>
</table>
limit above and below the optimum of 1) in which 50% of all $F$ values are found. For the trend index, “pass” is defined by $S_{\text{diff}}$, the median absolute difference between $S_{z,\text{obs}}$ and $S_{z,\text{sim}}$, before separating into categories of trend direction (Table 2).

4. Results

a. Distribution of index values

The similarity indices were found to vary strongly among the basins and with anomaly levels (Table 3). Overall, indices $A$ and $F$ show the largest deviations. For example, the agreement index $A$ varies from 0 to 0.63, with a maximum far from the optimum value of 1. The ratio of the frequency of events $F$ shows values above and below 1. The two indices based on annual aggregated information also varied among the basins. The correlation index $R$ ranges from slightly negative values to as high as $R = 0.94$. For the trend index $T$, all categories were found, but with different frequencies.

The agreement of simulated and observed anomalies differs systematically with the anomaly level $z$. Figure 3 shows the distribution of index values of all basins for the first two indices. The agreement index $A$ shows the lowest agreement for the driest anomaly ($z = 5$), then gradually increases up to $z = 60$, where it remains stable until $z = 80$. A reduction is again found for the wettest anomalies, although not as low as for the driest anomaly (Fig. 3a). The box plots for the ratio of the frequency of events $F$ (Fig. 3b) show only values above 1.0, as a few outliers that are less than 1.0 are not shown. A value of $F$ of $>1$ means that the simulated number of events is greater than the observed. Overall $F$ decreases with increasing anomaly level. The dry extremes show the largest ratios of up to 5 times the frequency of events in the simulations as compared with the observations.

The results for the second set of indices representing the year-to-year and long-term variability are shown in Fig. 4. The correlation coefficient $R$ shows relatively low correlations of dry anomalies $5 \leq z \leq 20$ with the lowest values found for $z = 5$ (Fig. 4a). In the range $25 \leq z \leq 85$, the distribution is rather stable with values of $0.4 \leq R \leq 0.8$. Only for $z = 95$ does $R$ drop again to lower values. The result for the trend index $T$ (Fig. 4b) shows that direction and magnitude of the trends in observations and simulations differ for a large percentage of the basins ($T = 0$; about 30%). For the driest ($z = 25$) and wettest anomaly levels ($z = 95$), trend direction also differs for a large number of basins, but differences in magnitude are small ($T = 3$). Between $z = 30$ and $z = 90$, the distribution of index values is relatively constant. Among the basins for which the direction of observed and simulated trends are similar, trends toward drier (less wet) conditions ($T = 1, 4$) make up a larger percentage than trends toward wetter (less dry) conditions ($T = 2, 5$). Within the group of basins with corresponding trends to drier conditions, slightly more than one-half of the trend slopes agree well in observations and models ($T = 4$); whereas within the group with corresponding trends to wetter conditions, at least two-thirds agree well ($T = 5$). This means that trends to wetter conditions, although less frequent among the basins and anomaly levels, appear better represented by the PMR simulation.

b. Distribution of benchmarks

The number of benchmarks passed for each index quantifies the differences among $z$ levels and basins (Table 3). More than one-half of the basins pass the benchmark for $A \geq 0.31$ ($R \geq 0.61$) for anomalies in the range $40 \leq z \leq 90$ ($25 \leq z \leq 80$). For $F$, only for the wet anomalies (i.e., anomaly levels of $z > 50$), more than one-half of the basins pass the benchmark. As the three indices $A$, $F$, and $R$ all compare aspects of the dynamics of modeled and observed hydrological events and show related dependencies on the $z$ level, their benchmarks were grouped (added) and mapped for three selected $z$ levels (Figs. 5a–c). Overall, the maps demonstrate the clear increase in number of benchmarks passed with increasing anomaly level (i.e., in the wet range). The spatial patterns are not very strong, however. For $z = 20$ (Fig. 5a), contiguous areas with two or three benchmarks passed are found only in the northern United Kingdom and in the western part of Switzerland. For $z = 50$ and $z = 80$, the patterns are similar. Most passes are found in the

---

**Table 2. Definition of the trend index $T_z$.**

| Trend direction                                      | Trend difference $|S_{\text{sim},z} - S_{\text{obs},z}| > S_{\text{diff}}$ | Trend difference $|S_{\text{sim},z} - S_{\text{obs},z}| \leq S_{\text{diff}}$ |
|------------------------------------------------------|---------------------------------------------------------------|---------------------------------------------------------------|
| Trend direction different                             | 0                                                             | 3                                                             |
| Trend direction toward drier conditions (negative trend for $z > 50$; positive trend for $z \leq 50$) | 1                                                             | 4                                                             |
| Trend direction toward wetter conditions (positive trend for $z > 50$; negative trend for $z \leq 50$) | 2                                                             | 5                                                             |

* Median absolute difference between $S_{z,\text{obs}}$ and $S_{z,\text{sim}}$.
northern part of the United Kingdom, and also in Denmark, Finland, the west coast of Norway, and in the Alps. The lowest number of passes consistently appears in the east and southeast of the domain (Austria, Czech Republic, and Slovakia), in Sweden, and inland and northern areas of Norway.

Different from the other three indices, the extremes of $z < 25$ and $z > 90$ for the trend index $T$ have a pass rate of over 50% of the basins, whereas the middle range of anomalies has a pass rate below 50% (Table 3). Figures 5d–f show the spatial distribution of the index (and benchmarks) for three anomaly levels. For all selected levels, observed and modeled trend directions are positive in the northern part of the domain and predominantly negative in central Europe. The pattern is strongest for $z = 50$ and $z = 80$. An area where many modeled and observed trend directions are different is southern Germany and the Swiss and Austrian Alps. The southern United Kingdom and France show mixed results for all anomaly levels. The most consistent agreement between modeled and observed trends (pass) is found for the positive trends in the northern United Kingdom and the negative trends in central Germany.

Last, the model test on the four indices and different anomaly levels was summarized in terms of the cumulative frequency of benchmarks passed (Fig. 6). Confirming the individual results, the frequency of passed benchmarks increases with increasing anomaly level $z$. For the driest anomaly levels $15 \leq z \leq 30$, more than one-quarter of all basins do not reach any benchmark. Overall, very few basins pass all benchmarks; even for higher $z$ levels, the number remains below 20%. The distribution of benchmarks passed is mostly uniform for wet anomalies, with the exception of $z = 95$.

### 5. Discussion

#### a. Appraisal of the indices

The indices employed in this study characterize commonly studied properties of hydrological events at a decreasing level of temporal detail: from the actual event, to general event characteristics, to their interannual variability and long-term change. Hence, they provide a challenging test of to what degree the PMR model simulations are able to reproduce observed dry and wet spells in the runoff from small basins. In general, the first set of indices, which is based on daily time series and compares the simultaneous occurrence of the events above or below a threshold, shows the lowest degree of similarity. This agreement of the events was found to be generally lower than the agreement of the interannual variability of the cumulative events. Another primary result was the overall low agreement found for the dry range of anomalies, which was contrasted by a much better agreement for moderately wet anomalies. The consideration of dry events below a threshold is a common approach to define drought events for drought frequency analysis. Several studies have applied thresholds in the range $10 \leq z \leq 50$ to analyze streamflow drought

---

### Table 3. Index ranges, benchmarks, and catchments that passed benchmarks (%).

<table>
<thead>
<tr>
<th>Index</th>
<th>$z$</th>
<th>A</th>
<th>F</th>
<th>R</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observ values</td>
<td>All</td>
<td>0.0–0.63</td>
<td>0.45–18.08</td>
<td>−0.15 to 0.94</td>
<td>0–5</td>
</tr>
<tr>
<td>Benchmarks</td>
<td>All</td>
<td>≥0.31</td>
<td>0.46 ≤ F &lt; 2.17</td>
<td>≥0.61</td>
<td>≥3 ($S_{diff} = 0.44$)</td>
</tr>
<tr>
<td>Passes</td>
<td>5</td>
<td>3.1</td>
<td>13.5</td>
<td>18.9</td>
<td>63.5</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>10.1</td>
<td>18.2</td>
<td>29.6</td>
<td>58.5</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>16.0</td>
<td>20.4</td>
<td>37.7</td>
<td>56.0</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>28.9</td>
<td>25.5</td>
<td>45.6</td>
<td>50.3</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>36.5</td>
<td>29.2</td>
<td>51.9</td>
<td>46.9</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>44.0</td>
<td>34.3</td>
<td>57.2</td>
<td>46.9</td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>49.7</td>
<td>39.0</td>
<td>61.0</td>
<td>47.2</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>57.5</td>
<td>44.7</td>
<td>61.9</td>
<td>46.5</td>
</tr>
<tr>
<td></td>
<td>45</td>
<td>61.3</td>
<td>47.8</td>
<td>60.4</td>
<td>46.2</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>62.9</td>
<td>53.1</td>
<td>59.4</td>
<td>47.5</td>
</tr>
<tr>
<td></td>
<td>55</td>
<td>66.7</td>
<td>54.4</td>
<td>57.5</td>
<td>47.5</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>69.5</td>
<td>59.7</td>
<td>58.2</td>
<td>48.4</td>
</tr>
<tr>
<td></td>
<td>65</td>
<td>68.9</td>
<td>63.8</td>
<td>56.3</td>
<td>45.0</td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>71.1</td>
<td>66.4</td>
<td>56.9</td>
<td>45.6</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>72.6</td>
<td>67.9</td>
<td>56.6</td>
<td>47.2</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>72.0</td>
<td>71.7</td>
<td>53.1</td>
<td>46.5</td>
</tr>
<tr>
<td></td>
<td>85</td>
<td>69.2</td>
<td>74.5</td>
<td>49.1</td>
<td>47.8</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>59.4</td>
<td>79.9</td>
<td>46.5</td>
<td>51.3</td>
</tr>
<tr>
<td></td>
<td>95</td>
<td>30.5</td>
<td>86.2</td>
<td>32.1</td>
<td>61.6</td>
</tr>
</tbody>
</table>

---

OCTOBER 2011 STAHLE TAL. 907
simulated from large-scale models and to map future changes of drought characteristics (e.g., Lehner et al. 2006; Andreadis et al. 2005; Vidal et al. 2010). The results of the study presented here show that a critical validation with respect to the models’ ability to capture event dynamics should precede such experiments. Findings suggest that, in particular, streamflow anomalies under low thresholds should not be considered in climate change impact assessments with HIRHAM5. Medium- to high-range anomaly levels, on the other hand, will provide a more reliable assessment, particularly if annual characteristics such as the total number of days below a threshold are assessed.

While the application of the indices $A$, $F$, and $R$ appeared to give consistent and usable results, $T$ was more difficult to assess. The index reflects similar trends in Europe over the past decades found in other studies with increasing runoff in northern Europe and decreasing runoff in central to eastern Europe (e.g., Dai et al. 2009; Stahl et al. 2010). This overall pattern of trend direction is generally reproduced by the model (Fig. 5) and is most clearly visible for $z = 50$ and $z = 80$. As opposed to the other indices, the benchmark pass rate for $T$ was higher for the driest anomaly levels. Many observed and simulated trends differ in their direction, however, and the low absolute difference between the trend magnitudes indicates that these were in fact close to zero ($T = 3$). Hence, for many basins the index does not compare a strong signal. The comparison of trends may thus only be useful where strong trends are present.

b. Streamflow dynamics

While the main aim of the presented approach is to identify limits of regional-climate-model simulations rather than to explain causes of the model performance, the indices themselves hint toward potential reasons why the model reproduces hydrological dynamics better under some conditions than others. The distribution-based index $F$ is the only index that shows a consistently increasing performance from dry to wet anomalies—that is, the wetter the conditions are, the better is the correspondence between observed and simulated number of events. However, $F$ is in most cases well above the optimal value ($F = 1$), indicating a general tendency of the model to exhibit a higher event-scale variability than the observations. The overestimation increases for dryer conditions.

Figure 7 suggests a relation between the number of events in the observed streamflow series and the results for index $A$. Observations and model agree better in
Fig. 5. Spatial distribution of benchmarks passed for (top to bottom) three selected anomaly levels: (a)–(c) sum of benchmarks passed for $A$, $F$, and $R$; and (d)–(f) values for index $T$. 
basins, where the observed streamflow exhibits a larger number of events. More events occur either in regions where rainfall events are more frequent or where the runoff response to rainfall is more direct. Basins in regions with most benchmarks passed—for example, Scotland and some areas on the west coast of Norway—combine both properties. In these regions, weather is dominated by midlatitude frontal systems and catchment storage is low. Therefore, the simple hydrological conceptualization of the subsurface in HIRHAM5 may be more appropriate.

In a comparative study of anomalies during the extreme drought years 1976 and 2003, Stahl and Tallaksen (2010) already found a lower degree of persistence in the streamflow simulated by the PMR than in the observations. While the overall spatial extent and occurrence of these two extreme dry events appeared to correspond well, the simulated droughts were interrupted more frequently; that is, the model responded too quickly with runoff to precipitation input following a dry period. An overestimation of surface runoff as a result of low water-holding capacity of soils in regional climate models has also been found in other studies (van den Hurk et al. 2005; Seneviratne et al. 2006). In addition, the simulation of low flows with large-scale models is in many cases hampered by the poor representation or lack of groundwater and its contribution to base flow. Simulating low flows and droughts correctly is difficult even with basin-scale hydrological models. Maurer et al. (2010) found that low flows (7-day low flow) were more difficult to capture than peak flows (3-day annual peaks) when reanalysis-downscaled climate forcing was used to drive a hydrological model for 11 basins in the western United States.

The spatial distribution of the number of benchmarks passed (Fig. 5) reveals a weak geographical pattern for the agreement of observed and simulated wet anomalies, with a west–east decrease. Such a geographical dependence may be due to more accurate atmospheric forcing close to the Atlantic Ocean or to generally wetter climate and moisture state of the land surface during most of the year. With respect to precipitation, however, climate models in general tend to perform better for oceanic climate types than for more continental types. This is connected with the more frequently dominated convective events over land—a process that is heavily parameterized in models (e.g., Jacob et al. 2007; Christensen et al. 2008).

There are some systematic relations of the indices with basin properties. The two indices $A$ and $R$ were correlated with the basin area and elevation of the 318 basins for each of the anomaly levels. No significant correlations were found for area, and thus a scaling effect can be ruled out. While only basins were included in the study for which model grid and real elevation agree well, significant correlations were found for $A$ and elevation above sea level. The correlations are low ($-0.1 > r > -0.3$), particularly for the wet anomalies. Scatterplots (not shown) reveal a systematic decrease in $A$ only for basins with elevations of $>1000$ m MSL. Hence, only a few very high (mountain) basins in the dataset (Fig. 1c) dominate this correlation. Results for those basins likely suffer from known challenges of the simulation of mountain hydrology such as strong gradients and snow accumulation and melt. There is no systematic relation with elevation for mid- to low-elevation basins. The lack of availability of other basin properties, such as land cover, soils, and geology, restricts further analyses.

6. Conclusions

This study presents a challenging test to high-resolution regional climate modeling by comparing modeled runoff with a set of indices from a large number of streamflow records from small basins across Europe. The approach...
differs from existing model validation studies in the use of exceedance-based indices and benchmarks against which simulations are compared and in the smaller scale but larger number of streamflow observations used.

Rather than adding an intermediate model for bias correction, observed and simulated streamflow were compared directly in terms of their anomalies with respect to the day of the year. This transformation has similarities to currently employed quantile–quantile bias correction of atmospheric variables, but the comparison of the individually transformed series directly avoids the added uncertainty of a model chain.

The novel exceedance-based indices introduced relate closely to different properties of hydrological event dynamics that are of interest in a climate change context, including changes in extremes. The concept of comparing events defined at different anomaly levels allowed for a systematic comparison of model simulations and observations for wet and dry spells. A clear conclusion from the study is that changes in dry anomalies below approximately the 30% threshold should not be assessed with this particular regional climate model. A more detailed analysis showed regions in Europe where the model represents the characteristics of dry and wet spells better; in particular, good results were obtained for regions with a high number of events (i.e., fast-responding basins in a humid climate). Reasons behind the pattern seen need further investigation; the relations found between the indices and the spatial distribution suggests that a combination of the atmospheric forcing and deficiencies in the representation of soil and groundwater storage are key factors, however.

The benchmark framework suggested allowed exploitation of the large number of basins employed in this study. Criteria of how many basins must pass a benchmark could guide the decision of whether to map future changes spatially for a particular anomaly level and characteristic. The concept can easily be adapted to other studies and indices, including comparison across models and benchmark criteria. The resolution of the climate-model simulations used is very high in comparison with common standards. Given the lack of dense homogenized networks providing near-surface observations of a number of variables, this study demonstrated that the usage of small-basin streamflow datasets may offer a new source of independent validation data with which to confront climate models once resolution is increased.

Acknowledgments. We gratefully acknowledge the contribution of all national data providers to the European Water Archive. Peter Berg and Philippe Lucas-Picher, formerly from the Danish Meteorological Institute, helped with valuable technical details about their reanalysis data.

The research was funded by the EU (FP6) Integrated Project WATCH under Contract 036946.

REFERENCES


