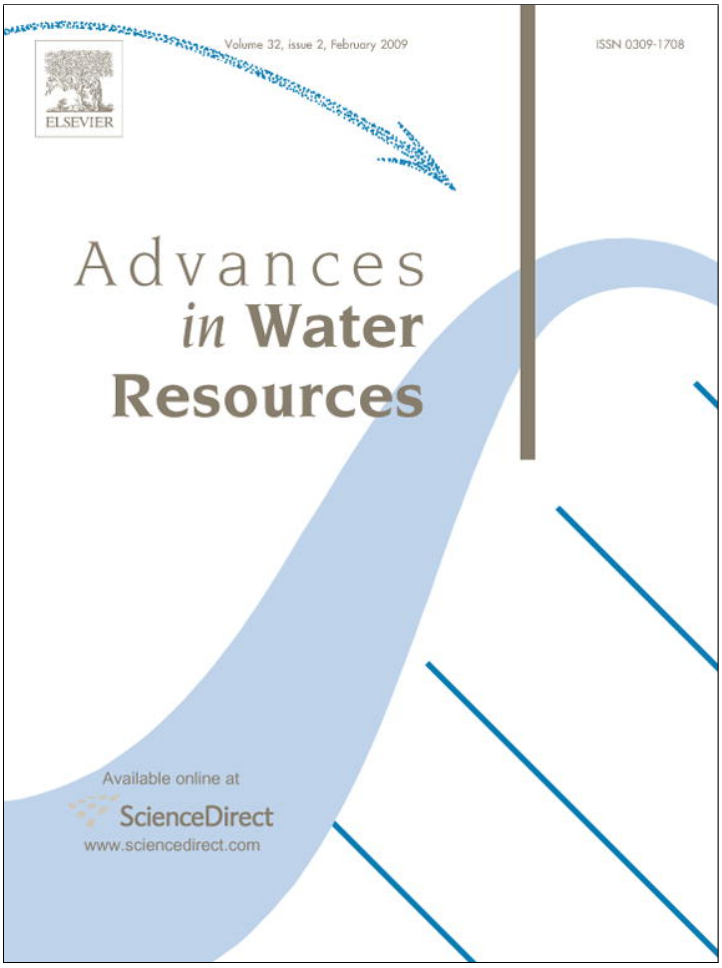
Volume 32, issue 2, February 2009
ISSN 0309-1708
ELSEVIER
Advances in Water Resources
Available online at
ScienceDirect
www.sciencedirect.com

Advances in Water Resources 32 (2009) 129–146

Contents lists available at ScienceDirect

Advances in Water Resources

journal homepage: www.elsevier.com/locate/advwatres

Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM). I: Model intercomparison with current land use

L. Breuer [a,*], J.A. Huisman [b], P. Willems [c], H. Bormann [d], A. Bronstert [e], B.F.W. Croke [f], H.-G. Frede [a], T. Gräff [e], L. Hubrechts [g], A.J. Jakeman [f], G. Kite [h], J. Lanini [i], G. Leavesley [j], D.P. Lettenmaier [i], G. Lindström [k], J. Seibert [l], M. Sivapalan [m,1], N.R. Viney [n]

[a] Institute for Landscape Ecology and Resources Management, Justus-Liebig-University Giessen, Heinrich-Buff-Ring 26, 35392 Giessen, Germany
[b] ICG-4 Agrosphere, Forschungszenrum Jülich, Germany
[c] Hydraulics Laboratory, Katholieke Universiteit Leuven, Belgium
[d] Institute for Biology and Environmental Sciences, Carl von Ossietzky University, Oldenburg, Germany
[e] Institute for Geoecology, University of Potsdam, Germany
[f] The Fenner School of Environment and Society, The Australian National University, Australia
[g] Afdeling Ecologie en Water, Lisec NV, Genk, Belgium
[h] Hydrologic Solutions, Pantymwyn, Flintshire, United Kingdom
[i] Department of Civil and Environmental Engineering, University of Washington, USA
[j] Denver Federal Center, US Geological Survey, Denver, USA
[k] SMHI, Norrköping, Sweden
[l] Department of Physical Geography and Quarternary Geology, Stockholm University, Sweden
[m] Centre for Water Research, University of Western Australia, Australia
[n] CSIRO Land and Water, Canberra, Australia

ARTICLE INFO

Article history:
Received 18 July 2007
Received in revised form 1 October 2008
Accepted 3 October 2008
Available online 18 October 2008

Keywords:
Hydrological modeling
Ensemble modeling
Land use
Model structure
Catchment model performance

ABSTRACT

This paper introduces the project on 'Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM)' that aims at investigating the envelope of predictions on changes in hydrological fluxes due to land use change. As part of a series of four papers, this paper outlines the motivation and setup of LUCHEM, and presents a model intercomparison for the present-day simulation results. Such an intercomparison provides a valuable basis to investigate the effects of different model structures on model predictions and paves the ground for the analysis of the performance of multi-model ensembles and the reliability of the scenario predictions in companion papers. In this study, we applied a set of 10 lumped, semi-lumped and fully distributed hydrological models that have been previously used in land use change studies to the low mountainous Dill catchment, Germany. Substantial differences in model performance were observed with Nash–Sutcliffe efficiencies ranging from 0.53 to 0.92. Differences in model performance were attributed to (1) model input data, (2) model calibration and (3) the physical basis of the models. The models were applied with two sets of input data: an original and a homogenized data set. This homogenization of precipitation, temperature and leaf area index was performed to reduce the variation between the models. Homogenization improved the comparability of model simulations and resulted in a reduced average bias, although some variation in model data input remained. The effect of the physical differences between models on the long-term water balance was mainly attributed to differences in how models represent evapotranspiration. Semi-lumped and lumped conceptual models slightly outperformed the fully distributed and physically based models. This was attributed to the automatic model calibration typically used for this type of models. Overall, however, we conclude that there was no superior model if several measures of model performance are considered and that all models are suitable to participate in further multi-model ensemble set-ups and land use change scenario investigations.

* Corresponding author. Tel.: +49 641 9937395; fax: +49 641 9937389.
E-mail address: lutz.breuer@umwelt.uni-giessen.de (L. Breuer).
[1] Now at: Department of Geography and Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA.

0309-1708/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.advwatres.2008.10.003

1. Introduction

1.1. Hydrological modeling of land use changes

Evaluating the impact of land use change on water and matter fluxes is a major challenge in hydrological research. Changes in land surface properties ultimately modify the energy and water exchange of the soil–vegetation–atmosphere system. A large number of field studies exist where the influence of land use changes (mainly afforestation and deforestation) is investigated in either paired site studies or single catchment experiments (see the reviews of [11,19,79]). Hydrological modeling of such changes has been conducted since the 1970s [10,21,64,81]. One of the challenges in hydrological modeling is to account for the changes in land surface properties by altering (at least some of) the model parameters. Therefore, it is commonly argued that process-based fully distributed models are best suited to simulate land use change effects [8]. Since in most cases only parts of the land use within a catchment changes, it is further argued that spatially distributed models depict these changes more precisely as compared to lumped modeling approaches [9]. Following this philosophy, several complex models have been developed that are capable of simulating land use changes, such as DHSVM [75,88], MIKE-SHE [3,70], RHESSys [5,6], SHETRAN ([32], a modeling system based on the SHE model [1]), TAC-D [65], TOPLATS [33,67] and WASIM [61,73]. However, these fully distributed process-based modeling approaches are often criticized because an *a priori* estimation of model parameters is difficult [9,32]. As an alternative, physically based, semi-distributed models with less complex spatial representation have been proposed. This group of models simulates all hydrological processes within spatially non-explicit Hydrological Response Units (HRU). Results for each HRU are lumped within subcatchments and routed downstream. Models following this approach are SWAT [4,86], SWIM [50,51], PRMS [53], HYLUC [20] and SLURP [48,49]. HRUs can be defined based on soil units, land use or a combination of both. Although the impact of land use change is not simulated with the same spatial resolution as in a fully distributed approach, these semi-distributed models still require a considerable number of parameters that might be difficult to obtain. A further simplification is achieved if hydrological fluxes are simulated with the subcatchment scale as the smallest spatial unit. Models such as HBV [7,54] and LASCAM [76] follow this concept. Depending on the size of these subcatchments, the spatial resolution of the simulations is rather coarse. All model types described so far utilize parameters that can be estimated directly from land use data. At the lower end of complexity, conceptual lumped models such as the IHACRES [26,46] or the NAM [62] model can be found. These models are characterized by a simple model structure and a small number of conceptual model parameters. A change of land use in a particular catchment is simulated by deducing parameter values from other catchments that have similar land use compositions as the one for which predictions are being made. Most often, general catchment attributes, such as land use area and catchment area, are used to regionalize model parameters. It is argued that data limitations in many catchments limit the application of physically based models, and that conceptual models provide a more appropriate alternative [27].

No matter what model type is chosen (conceptual vs. physically based and lumped vs. spatially distributed), extrapolations to situations where no measured data exist (e.g. in ungauged basins), where basin characteristics are not stationary (e.g. land use change scenarios), or where boundary conditions are changing (e.g. climate change) are difficult to undertake [9,71]. This research topic has attracted worldwide attention in the hydrological community and led to the ongoing Predictions in Ungauged Basin (PUB) initiative by the International Association of Hydrological Sciences (IAHS) [75]. Climate change and its impact on hydrological fluxes are of growing concern and have also resulted in a set of research initiatives. These include the Global Energy and Water Cycle Experiment (GEWEX) and the Program for Climate Model Diagnosis and Intercomparison (PCMDI). No such research initiatives have been developed to approach the problem of land use change and its impact on hydrology.

Despite this growing knowledge, a high degree of uncertainty remains in all model approaches. This global model uncertainty stems from stochastic and structural model uncertainty. Stochastic model uncertainty is introduced by measurement errors of model input and output data, problems in the observation of physically based model parameters, parameter estimates, and spatial as well as temporal parameter heterogeneity. Numerous techniques to investigate stochastic model uncertainty have been published in the field of hydrology (e.g. [25,57,84] amongst many others). These techniques are based on Monte Carlo simulations [31], Latin Hypercube sampling [25], General Likelihood Uncertainty Estimation GLUE [8], or the Shuffled Complex Evolution Metropolis (SCEM-UA) technique [84].

The validity of a given model depends on the scientifically acceptable explanation of the cause–effect relationships within the model [72]. The structural model uncertainty results from unknown, simplified, incomplete or incorrect process descriptions within the model. Models are always abstractions of real systems. Hence, not all processes can be included, may it be on purpose to keep the model structure simple or due to the lack of knowledge. Several approaches have been suggested recently to deal with this structural model uncertainty by using a set of different models. A simple method of combining several model structures (and further evaluation techniques) was suggested in the model protocol by Refsgaard et al. [71], Vrugt et al. [83] and Duan et al. [29] presented approaches for the investigation of structural and stochastic model uncertainty simultaneously. However, the basis for using such a set of different models is a detailed evaluation of the performance of individual models in a model intercomparison.

1.2. Model intercomparison and the concept of multi-model ensembles

A first step towards the evaluation of the effect of model structure on model output is a model intercomparison. Here, different model structures are applied to the same data set under identical boundary conditions. The simulation results are then interpreted in the light of differences in the description of processes and their spatial and temporal resolution. A range of model intercomparisons have been performed in the field of hydrology. Diekkrüger et al. [28] compared the performance of 19 agro-ecosystem plot-scale models for two sites in Germany. They concluded that the model results differed substantially and that no superior model type, either simple conceptual or complex process-based, could be defined. A conceptual lumped, a semi-distributed and a distributed physically based hydrological model were applied to three catchments in Zimbabwe by Refsgaard and Knudsen [69]. They showed that the models performed equally well in the calibration period. Under ungauged conditions (no data available for model calibration; parameter values transferred from adjacent calibrated catchments), the semi-distributed and fully distributed models slightly outperformed the conceptual model. In land surface hydrology, the Project for Intercomparison of Land surface Parameterization Schemes (PILPS) is an on-going model intercomparison project in its fifth stage [39,40,63, and references therein]. The project was established in 1992 and is designed to improve the parameterization of the continental surface, especially with respect to water, energy, momentum and carbon exchanges with the atmosphere. The Model Parameter Estimation Experiment (MOPEX) is aiming at developing techniques for a priori estimation of parameters in large scale hydrologic models and land surface parameterization schemes [30]. The MOPEX strategy is based on a three-step approach to relate model parameters to land surface characteristics, including the demonstration of parameter transferability. A change towards more sophisticated and comprehensive comparisons in distributed hydrological model predictions has recently been achieved with the Distributed Model Intercomparison Project (DMIP [77]) of the NOAA National Weather Service. DMIP sheds light on the evaluation of distributed models against both a calibrated lumped model and observed streamflow [68,78]. Carpenter and Georgakakos [22] built on the experience of DMIP and developed a probabilistic methodology to compare streamflow simulations using lumped as well as distributed realizations of the Sacramento soil moisture accounting model. They concluded that the distributed approach outperformed the lumped approach with superior performance for peak flow magnitude as well as timing. Using the same Sacramento soil moisture accounting model in a semi-distributed and lumped set-up, Boyle et al. [14] showed that the main improvements in model performance of the semi-distributed approach were attributable to the spatial representation of precipitation input and routing. This led to a better representation of flood peaks and quick recession, but did not improve baseflow simulations.

These intercomparison projects have shown that different model types with varying degrees of complexity have assets and drawbacks when used for prediction. To increase the confidence in any future model prediction, different model structures that were previously evaluated in a model intercomparison study can be combined in an ensemble model. These so called multi-model ensembles are commonly used in meteorology and climatology to quantify the predictive uncertainty of weather forecasts and climate change simulations [59]. Recently, the envelope of such multi-model climate and weather predictions has been utilized to drive hydrological models (e.g. [24,38,80]). However, this approach focuses on the proper representation of input data uncertainty rather than on the structural uncertainty of the hydrological models. Applications of multi-model ensembles in hydrological research remain rather scarce. One of the first papers on this topic was published by Shamseldim et al. [74], who used five black-box-type rainfall–runoff models and two averaging schemes as well as a neural network to combine model outputs. Within the DMIP project, Georgakakos et al. [37] set up a multi-model ensemble to improve flood forecasting. As a contribution to the PUB effort, McIntyre et al. [56] applied a model ensemble to a set of 127 ungauged or poorly gauged UK catchments. New tools have recently been proposed to improve the combination of models into a multi-model ensemble, including deterministic (model average and median, trimmed means, linear regression, artificial neural networks; see for example [37,74]) as well as probabilistic techniques (Bayesian model averaging, Kalman filtering, Brier skill score, see for example [2,55,66,85]).

1.3. Motivation for the LUCHEM project

In the framework of the Collaborative Research Centre 299 "Land use concepts for peripheral regions" spatially differentiated land use scenarios have been developed for the Dill catchment, Germany. The land use scenarios have been calculated by the use of the *Pro*gnosis of *Land*use (ProLand) model. ProLand is an agro-economic model that predicts feasible land use systems based on political, economic, social and ecological conditions [52,87]. Several land use scenarios have been evaluated with respect to ecological and economic landscape services [12,16,18,34–36,58,86]. The land use scenarios analyzed in the present evaluation are based on an investigation of the economic and ecological effects of land consolidation (see Huisman et al. [43], this issue, for a detailed description).

As part of these evaluations, the SWAT model [4] was used to calculate the changes in the hydrological cycle, also considering the stochastic uncertainty in the model predictions [17,31,44]. For all of the land use scenarios investigated, it was concluded that the predicted changes in discharge, groundwater recharge or surface runoff were small. Obviously, the question remained whether the minor changes in the hydrological cycle predicted by SWAT are realistic, or a result of an incomplete or even wrong model structure. To address this question, the LUCHEM ('Assessing the impact of land use change on hydrology by ensemble modeling') project was initiated in which a multi-model ensemble was applied to the current land use and a set of land use scenarios. The results are presented in four companion papers.

This first paper outlines the motivation and setup of the LUCHEM project. We first present the experimental design of LUCHEM, followed by a summary of the results of the model intercomparison for the current land use. To address the effect of model input data, we investigate two different model set ups. We conclude with determining the differences in model performance depending on model complexity, and finally, highlight the impact of model structure on model output. In the second paper by Viney et al. [82] (this issue), different deterministic and conditional ensemble techniques are investigated with respect to model output and performance. Typically, the predictions of these multi-model ensembles are more accurate than any single model, probably because model structural error is partly removed by combining models. In the third paper by Huisman et al. [43] (this issue), the predictions for the land use scenarios are compared for the individual ensemble members. In addition, they applied the most promising deterministic ensemble methods of Viney et al. [82] (this issue) to the land use scenarios and quantified the uncertainty in the scenario predictions using the model ensemble. Finally, Bormann et al. [13] (this issue) analyzed the effect of spatial resolution and distribution of model input data on model output for the current land use and the land use change scenarios. For this, spatial input data were systematically aggregated and redistributed (randomly and based on topography). Sensitive grid size aggregation levels were identified.

2. Study area and available data

The Dill catchment is located in the Lahn-Dill low mountainous range approximately 70 km NNW of Frankfurt, Germany. Its catchment area is 693 km^2 with altitudes ranging from 155 to 674 m above sea level (Figs. 1 and 3). The eastern part of the Westerwald, extending into the western area of the Lahn-Dill low mountainous region is mainly characterized by tablelands of tertiary basalts at heights around 500–600 m above sea level. The average slope is approximately 14%. A 25 m digital elevation model (DEM) is available for watershed delineation [41].

During the Pleistocene, Hessian low mountain ranges were subject to periglacial processes. Solifluction mixed weathering debris with loess. Therefore, periglacial layers strongly influenced by the underlying geologic substrate are the main soil parent material within the area. Due to the heterogeneous nature of these periglacial layers, the pattern of soil types is complex. However, more than a third of all soil types in the Dill catchment belong to the general class of cambisols [45], equivalent to inceptisols in the USDA soil taxonomy. These soils cover more than 64% of the total catchment area (Table 1). Soil digital data is provided on the scale of 1:50,000 for the catchment [42]. Soil physical properties for 149 soil types include information on soil depth, available water con-

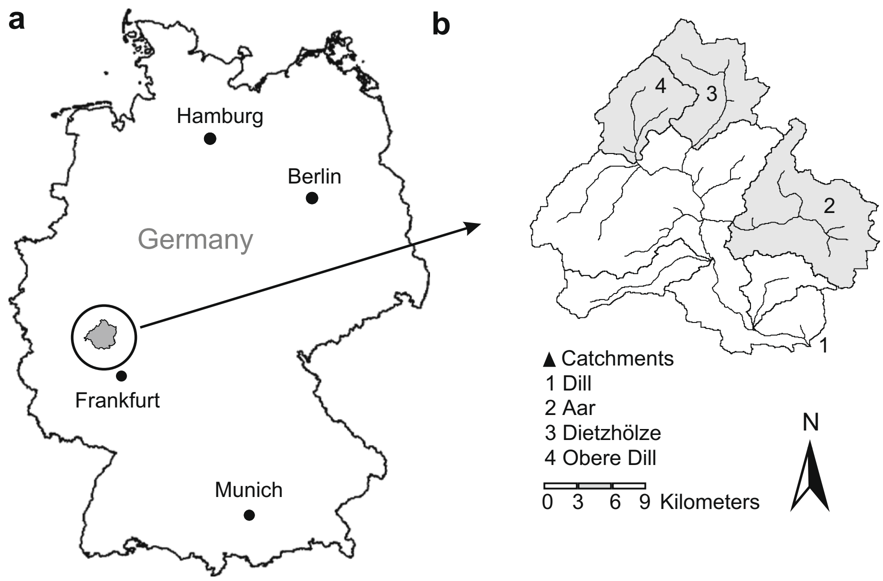

Fig. 1. Location (a) and gauging stations (b) of the Dill river catchment.

tent (AWC), saturated conductivity (K_{sat}), soil albedo for visible light, rock fragments and clay/silt/sand content. Soil digital data are available on the same 25 m grid as the DEM.

Land use information was derived from a multi-temporal set of Landsat TM 5 images taken in 1994 and 1995. Similar to the other spatial data, land use information was scaled to a 25 m grid size. The catchment is characterized by a heterogeneous small-scale land use pattern with more than 50% of the area covered by forests and 20% by pasture (Table 1 and Fig. 2). Cropland is mainly located in the eastern part of the catchment. A crop rotation composed of winter rape, winter barley, and oats is typical for the research area. Fallow land makes up almost 10% of the area, with up to 40% in certain regions, and is hence more abundant than anywhere else in Germany. Grassland use is equally distributed between dairy and extensive livestock farming (suckler cows, hay production, recreational horse husbandry). Deciduous forests are mainly composed of *Quercus robur*, *Quercus petrea* (oak) and *Fagus sylvatica* (beech), whereas *Picea abies* (spruce) and *Pinus sylvestris* (Scots pine) dominate coniferous forests. Measured plant parameters were not available and were, therefore, obtained from literature [15].

Table 1
Characteristics of the Dill catchment.

Annual rainfall[a] [mm]	907.3
Annual discharge[b] [mm]	437.7
Drainage area [km^2]	693.0
Land use [% of watershed area]	
Cropland [%]	6.5
Pasture [%]	20.6
Fallow [%]	9.1
Urban [%]	9.2
Deciduous forest [%]	29.5
Coniferous forest [%]	24.8
Water [%]	0.3
Soil type[c] [% of watershed area]	
Cambisol [%]	64.2
Planosol [%]	13.9
Gleysol [%]	5.3
Fluvisol [%]	4.1
Regosol [%]	1.4
Luvisol [%]	0.5
Others [%]	10.6

[a] Arithmetic annual mean (1983–1998) calculated on the basis of the 16 rain gauging stations shown in Fig. 3.
[b] Arithmetic mean calculated for the overall model period 1983–1998.
[c] Soil type classification according to World Reference Base for Soil Resources [45].

Mean annual precipitation ranges between 700 and 1300 mm depending on elevation and longitude. Precipitation is highest in the western and northern part of the catchment, whereas substantially less precipitation falls in the eastern part. Areas with low precipitation have a summer-dominated precipitation pattern, whereas areas with high precipitation are characterized by winter-dominated precipitation patterns of westerly low-pressure systems. The contribution of snow to total precipitation is estimated to be less than 5%. Average annual mean temperature is 8 °C. Climatic daily data, provided by the German weather service (DWD), are available for the period of 1980–1998. Data include precipitation [mm], wind speed [m s^{-1}], global radiation [MJ m^{-2} d^{-1}], air temperature [°C] and relative humidity [%]. Whereas precipitation was recorded at 16 locations within or close to the catchment, spatial information for the remaining climatic variables is limited to two locations (Fig. 3). Mean annual precipitation for the gauges is given in Table 2.

Besides daily discharge data at the Dill catchment outlet (gauging station Asslar), data from three additional gauging stations were available: gauging station Herbornseelbach for the Aar catchment (133.4 km^2), gauging station Dillenburg for the Dietzhölze (Dtz) catchment (80.0 km^2) and gauging station Haiger for the Obere Dill (Obd) catchment (62.1 km^2) (Fig. 1).

3. Discharge characteristics in the Dill river catchment

The mean annual discharge of the Dill river catchment for the period of investigation is 437 mm with a mean annual precipitation of approximately 900 mm. Under normal climatic conditions, discharge peaks in the winter months (Fig. 4). Temporal variation of discharge for the subcatchments is similar to the Dill. Based on hydrogeological investigations and baseflow separation published elsewhere [47], the portion of baseflow contribution to discharge is minor in the order of 10–20%. The contribution of surface runoff is even less and assumed to be <10%. Most of the discharge of the Dill is delivered by interflow through the shallow soils.

For a comparison of measured and simulated flow series a statistical separation into three discharge (Q) components (slow,

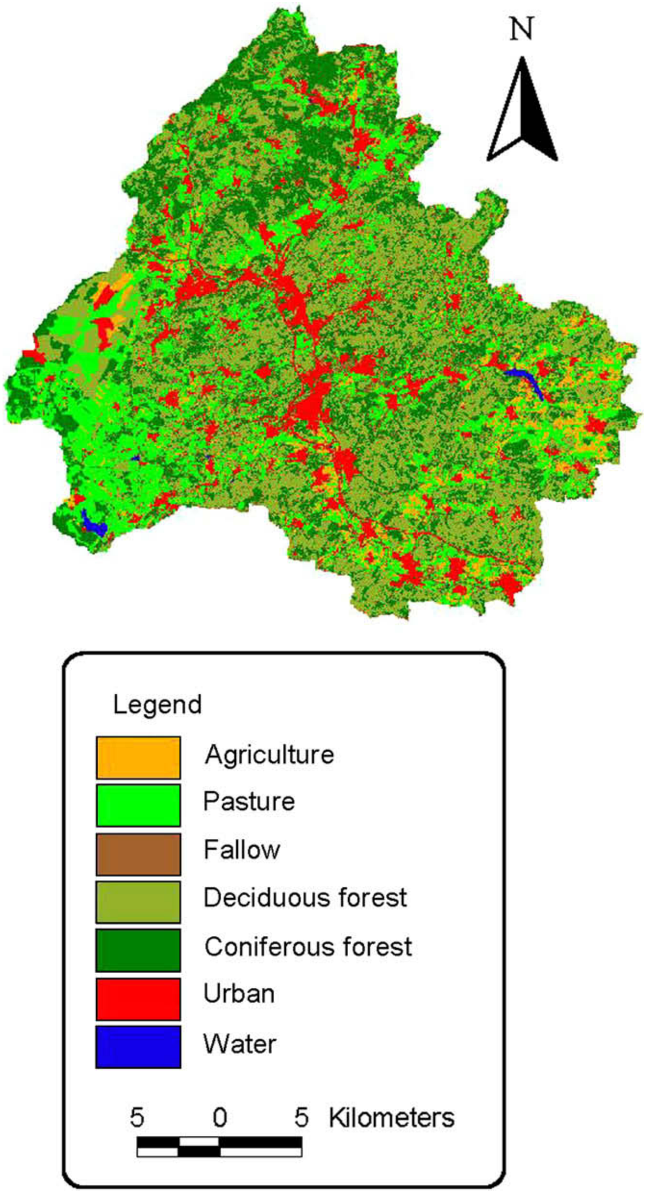

Fig. 2. Land use distribution in the Dill catchment as derived from a multi-temporal composite of three Landsat TM 5 images from 1994 to 1995 (pixel size 25 m). A three-step procedure has been applied using a semi-automized image classification scheme that consisted of raw data pre-processing, hierarchical classification of large landscape units and a sub-pixel classification for small heterogeneous land surfaces. Sub-pixel classification was based on a spectral mixture analysis, and was necessary to gain land use information on the very small structured agricultural landscape and its high portion of fallow land with secondary vegetation.

intermediate and quick flow) was done by the WETSPRO procedure presented by Willems [89] that is built on a generalized version of the numerical digital filter developed by Chapman [23]. The filter is based on the assumption of exponential recessions for the discharge components with constant recessions for different slow flow periods. The recession constants were derived by analysis of the slope of the linear recession of the flow during long dry periods in a $\log Q$ vs. time plot. The slope was estimated for different dry periods in the flow series. The average value for the Dill river for slow flow equals ~30 d and is similar for the subcatchments. For the intermediate flow recession constant, a value of ~3 d was estimated.

To avoid serial dependence in the model residuals to be used in the statistical analysis of slow flows and quick flows (i.e. surface runoff and interflow), the daily discharge series were divided into independent quick (discharge maxima) and slow flow (discharge minima) periods. For this, peak discharges (also referred to as peak over threshold values), were extracted from the discharge time series using independency criteria based on the recession constant. Consecutive discharge peaks were considered to be independent when (1) the inter-event time exceeded the recession constants of the quick flow components and (2) the minimum inter-event discharge approached the slow flow value (Fig. 5). The slow flow component was derived when quick flow components decreased to zero during the dry periods. In a second step, independent slow flow periods were selected as well. The slow flow periods were separated in a similar way as the quick flow periods. Slow flow recession periods were considered independent when the slow flow was reduced to a very low value at the end of a given period and was followed by a significant increase in the slow flow before the next recession period started (Fig. 5). This procedure was also applied to define discharge minima and maxima for the simulated discharges by the 10 models that participated in the model intercomparison.

From a detailed recession analysis (data not shown here), it became clear that the exponential recession stops below a specific slow flow value. This indicates that the lowest discharge values most probably do not originate from rainfall–runoff processes. For this reason, an external inflow was assumed that originates from the effluent of wastewater treatment plants and industrial discharges. The estimate of the external inflow for the entire Dill catchment is 0.02 mm d^{-1} (0.16 $m^3 s^{-1}$).

4. Project design

4.1. Participants of the LUCHEM project

The selection of models which were invited to participate in the model ensemble was based on a search in Web of Science® [access date 31.08.2004]. We only selected models that had been applied in studies dealing with land use change effects on hydrological processes in at least two different catchments. Out of 12 groups invited, 10 groups decided to participate in the model ensemble workshop in March 2005 (Table 3). All necessary data for the model setup were provided to the model groups before the workshop. General characteristics of the models and details of their application in the Dill catchment are given in Table 4. The smallest unit considered within a model ranged from 0.01 km^2 to the size of the gauged subcatchments.

A ranking of the complexity of the models was based on the number of model input parameters (i.e. time invariant characteristics such as soil and vegetation properties), calibrated parameters and the spatial resolution given as spatial computational areas (Table 5). Spatial computational areas are equivalent to pixels of the distributed models and to HRUs in the case of the semi-distributed models. The fully distributed models formed the group of the most complex models. Within this group, TOPLATS and WASIM were ranked lower than MIKE-SHE and DHSVM because these models were used without an explicit groundwater modeling scheme. The second and third group contained the semi-distributed models and the lumped models. Within the group of semi-distributed models, the models were further ranked by the number of HRUs and the number of parameters calibrated for the Dill catchment. Here the SWAT model is the most complex one as it has an internal dynamic plant growth module that is very parameter intensive.

4.2. Modeling instructions

Meteorological forcing data and observed streamflow data covered the period from 01.01.1980 to 31.12.1998. The period from

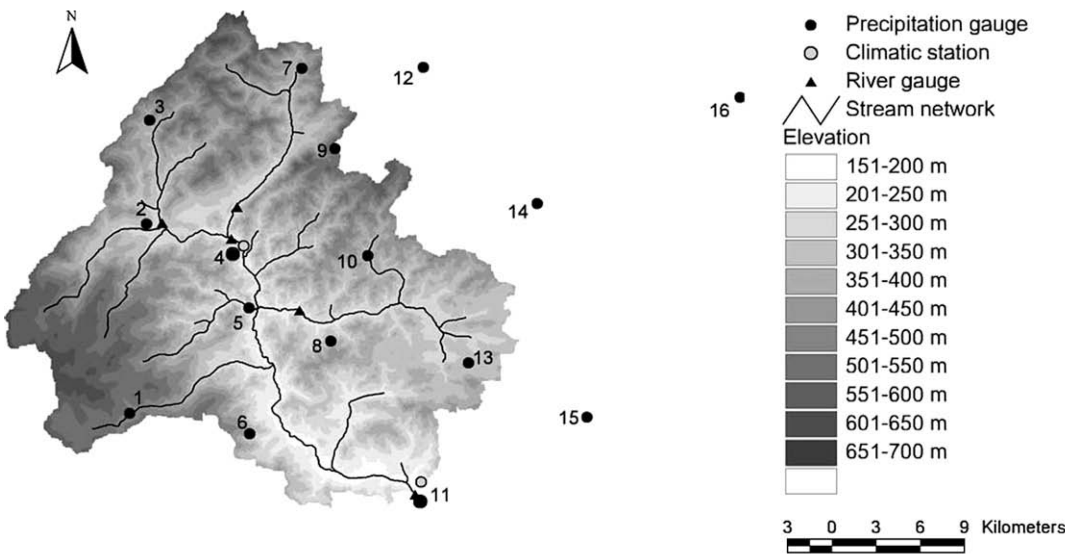

Fig. 3. Digital elevation model and location of gauging and meteorological stations of the Dill catchment (pixel size 25 m).

01.01.1980 to 31.12.1982 was defined as model warm-up period and not used in the evaluation of model results. All modeling groups were instructed to use the period from 01.01.1983 to 31.12.1989 for model calibration and the period from 01.01.1990 to 31.12.1998 for model validation. It should be noted that the mean annual precipitation in the calibration period is lower than the precipitation in the validation period for almost all precipitation stations (Table 2).

Streamflow data were provided for the Dill catchment and three of its subcatchments, the Aar, the Dietzhölze and the Obere Dill. Participants were asked to provide results for these four gauged catchments. Two sets of simulations were performed for the actual land use distribution. The first set was based on the simulations all participants provided prior to the modeling workshop (further referred to as the *original* set of simulations). During the workshop, it was decided that differences in model results were difficult to interpret because (1) the provided input data were interpreted differently (e.g. different annual courses of LAI from provided maximum LAI) and (2) interpolation of precipitation and other climate variables varied considerably between the models. Consequently, it was decided that all models should also provide results for a *homogenized set* of data.

In an attempt to homogenize the precipitation input data, each pixel of the DEM (25×25 m) was allocated to a precipitation station based on nearest neighborhood. Apart from the fully distributed models, all models had to further process this precipitation data. Because of this, differences in precipitation input remained even in the homogenized data set. To homogenize the simulation of plant development, monthly average values of leaf area index (LAI) were calculated by a simplified EPIC approach [90] using average daily weather data for each land use type shown in Fig. 2. To homogenize the remaining climatic variables, it was decided to only use the climate station data of Dillenburg.

We argue that the type and amount of model calibration is part of the model philosophy. Hence, no calibration procedure was prescribed for the different models. Table 4 shows that the process-based fully distributed models were generally calibrated manually (e.g. DHSVM, MIKE-SHE). Here, visual hydrograph evaluation was preferred over the use of statistical evaluation measures. Most of the more conceptual models were automatically calibrated using an objective function based on squared residuals (e.g. root mean squared error [RMSE] or Nash–Sutcliffe Efficiency [NSE] [60]). The HBV, TOPLATS and IHACRES model were calibrated by applying two or more different objective functions (Table 4).

Table 2
Mean annual precipitation for the calibration and validation period. Station numbers refer to Fig. 3.

Station #	Precipitation gauging station	Height a.s.l. [m]	Precipitation calibration period 1983–1989 [mm a^{-1}]	Precipitation validation period 1989–1998 [mm a^{-1}]
1	Driedorf	482	1221.6	1307.1
2	Haiger	290	890.6	912.3
3	Haiger-Dillbrecht	340	1050.8	1069.2
4	Dillenburg	277	814.6	814.6
5	Herborn	237	828.8	872.9
6	Greifenstein	430	963.1	969.0
7	Dietzhölze-Mandeln	376	1170.5	1206.0
8	Mittenaar-Bicken	240	755.8	782.8
9	Eschenburg-Hirzenhain	530	1076.8	1046.3
10	Siegbach-Eisemroth	330	860.2	860.6
11	Aßlar	200	810.0	848.9
12	Steffenberg-Quotshausen	338	907.6	902.1
13	Hohenaar-Erda	306	772.9	803.7
14	Gladenbach	270	853.2	859.3
15	Wettenberg-Krofdorf	235	694.2	737.7
16	Coelbe	187	704.6	692.2

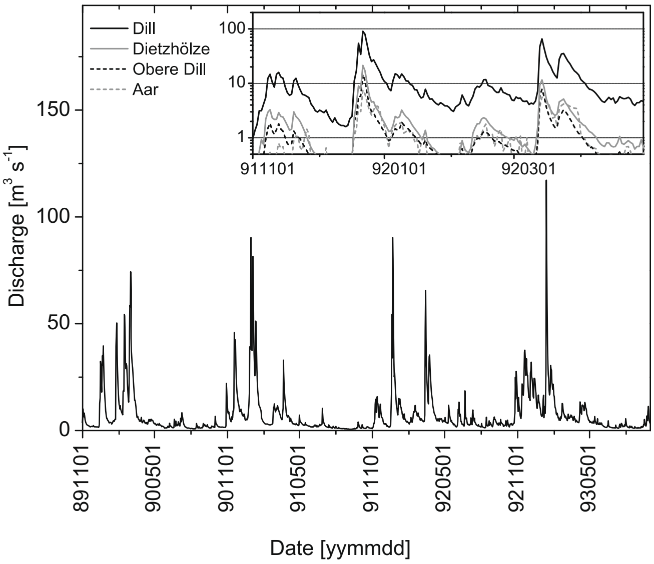

Fig. 4. Daily discharge [$m^3\ s^{-1}$] of the Dill river and its tributaries Aar, Dietzhölze and the Obere Dill, period 01.11.1991–30.04.1992. The insert depicts log-scaled discharge to better represent low flow conditions.

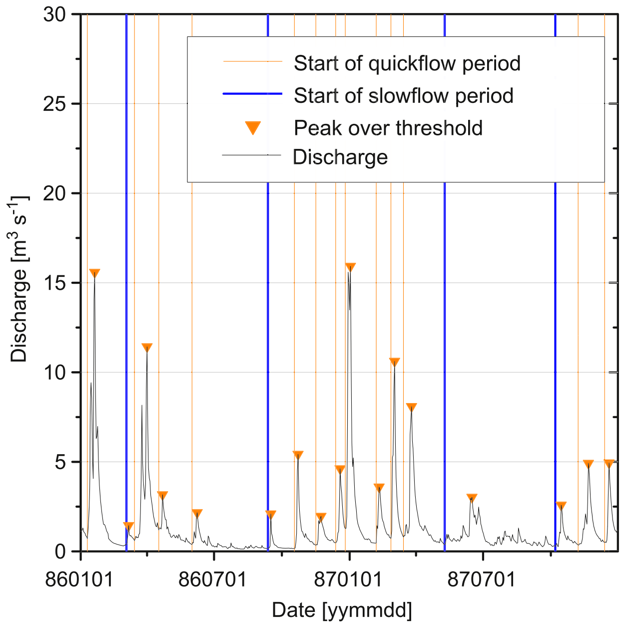

Fig. 5. Peak separation of the Dill river daily discharge series in nearly independent quick flow and slow flow periods (period 01.01.1986–31.12.1987). Orange lines indicate the start of a quick flow period, where rain storms lead to nearly independent periods if separated by a time span longer than the recession constant of the quick runoff process. Blue lines indicate slow flow periods that are characterized by longer independence times. They originate from different storms when baseflow is reduced to low values at the end of a given period, followed by a significant increase in baseflow before the next recession period starts. Note that blue lines mask additional orange lines. (For interpretation of the references in color in this figure legend, the reader is referred to the web version of this article.)

Most models were only calibrated to observed streamflow at the outlet of the Dill catchment (gauging station Asslar). Streamflow data for the remaining interior gauged subcatchments of the Obere Dill, Aar and Dietzhölze were only used in the calibration process of the IHACRES and LASCAM model, whereby the first model was calibrated independently to the single subcatchments and the latter one followed a multi-criteria calibration scheme for all subcatchments concurrently to yield a single catchment-wide parameter set. Except from the MIKE-SHE model, all participants provided results for the original and homogenized data set, for all catchments and land use scenarios.

4.3. Evaluation measures

The following statistical measures were used for model calibration and model evaluation. In the following equations, N is the total number of measurements, S is simulated discharge, O is observed discharge and i is a counter.

(a) Root mean squared error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(S_i - O_i)^2}{N}} \tag{1}$$

The RMSE provides information on the agreement between measured and simulated data, whereby negative and positive values do not cancel out each other. Large deviations are emphasized.

(b) Bias

$$\mathrm{Bias} = \sum_{i=1}^{N}(S_i - O_i) \tag{2}$$

(c) Absolute percent bias (PB%)

$$\mathrm{PB\%} = \frac{\left|\sum_{i=1}^{N}(S_i - O_i)\right|}{\sum_{i=1}^{N} O_i} \times 100 \tag{3}$$

Absolute percent bias is a measure for total volume differences between measured and observed data. It was used to evaluate the long-term performance of the model simulations.

Table 3
Participating modeling groups in the LUCHEM model ensemble.

Model	Reference	Represented by	Institution
DHSVM	[88]	Jordan Lanini, Dennis Lettenmaier	Department of Civil and Environmental Engineering, University of Washington, USA
HBV	[7]	Jan Seibert	Environmental Assessment, SLU Uppsala, Sweden
		Göran Lindstrom	SMHI, Norrköping, Sweden
IHACRES	[26]	Barry Croke, Tony Jakeman	The Fenner School of Environment and Society, The Australian National University, Australia
LASCAM	[76]	Neil Viney	CSIRO Land and Water, Wembley, Australia
		Murugesu Sivapalan	Center for Water Research, University of Western Australia, Australia
MIKE-SHE	[70]	Lode Hubrechts	Afdeling Ecologie en Water, Lisec NV, Genk
		Patrick Willems	Hydraulics Laboratory, Katholieke Universiteit Leuven, Belgium
PRMS	[53]	George Leavesley	Denver Federal Center, USGS Denver, USA
SLURP	[49]	Geoffrey Kite	Hydrologic Solutions, Pantymwyn, Flintshire, UK
SWAT	[4]	Lutz Breuer, Johan A. Huisman, Hans-Georg Frede	Institute for Landscape Ecology and Resources Management, Justus-Liebig-Universität, Gießen, Germany
TOPLATS	[33]	Helge Bormann	Institute for Biology and Environmental Sciences, Carl von Ossietzky University Oldenburg, Germany
WASIM	[73]	Thomas Gräff, Axel Bronstert	Institute for Geoecology, University of Potsdam, Germany

(d) Nash–Sutcliffe Efficiency (NSE)

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}(S_i - O_i)^2}{\sum_{i=1}^{N}(S_i - \bar{S})^2} \quad (4)$$

Values for Nash–Sutcliffe Efficiency [60] vary from negative infinity to 1.0. Values close to 1.0 indicate good model performance, whereas negative values indicate that the average of the observed data is a better predictor than the model.

These four measures together with the mean and standard deviation of measured and simulated streamflow were calculated for the calibration, validation and the entire modeling period for all models.

5. Model results for the current land use distribution

In the following, we will present and compare the results for the simulations of water fluxes for the Dill catchment and its tributaries in detail. Results will be shown for the suite of 10 hydrological models given in Table 4. The variation of model results can generally be attributed to three major effects: (1) the effect of model input data, (2) the effect of model calibration, and (3) the physical basis of the models.

5.1. Effects of model input data

In order to better compare the effects of model structure on model output, two sets of model input data were provided as described above. A homogenization of model input data was conducted to reduce variation between models with respect to LAI, precipitation and further climatic data. Homogenization of the model input data resulted in a reduction of precipitation input for almost all models except for WASIM and LASCAM (Fig. 6). The span of annual precipitation across all models was reduced from 96 to 37 mm, but considerable differences between precipitation inputs still remained. These differences can be attributed to the spatial resolution of the different model units and the allocation of precipitation to these units. HBV and DHSVM calculated a reduction of mean annual precipitation by ~50 mm. Both models

Table 4
General model characteristics.

Model	Spatial resolution	Soil horizons	Precipitation interpolation	ET_{pot}	Dynamic plant growth/crop rotation	Calibration/objective function	Other info
DHSVM	100 m	3	Inverse distance	Penman–Monteith	No/no	Manual/no	Minimum channel source area 10 ha, saturation access
MIKE-SHE	200 m	Depending on soil type	Nearest neighbour	Penman–Monteith	No/no	Manual/R^2 $\sum\|S_i - O_i\|$	LAI and root development calculated with internal database, aggregation to 25 different soil types covering 90% of the area
TOPLATS	100 m	2	Nearest neighbor	Penman–Monteith	No/no	Manual/NSE, bias	
WASIM	200 m	Conceptual	Inverse distance	Penman–Monteith	No/no	Manual/NSE	
SWAT	52 subcatchments	2–5 (depending on soil type)	Nearest neighbour	Penman–Monteith	Simplified EPIC crop growth model/yes	Automatic/NSE	
PRMS	25 subcatchments	2	Nearest neighbour	Modified Jensen–Haise parameters *f*	No/no	Automatic/$\sum\|S_i - O_i\|$	1 non-linear subsurface reservoir, 1 linear groundwater reservoir
SLURP	39 subcatchments	2 (based on field capacity)	Nearest neighbor for each grid cell, aggregated to subcatchment/ land cover unit	Penman–Monteith	No/yes (with multi-year LAI)	Manual/NSE	Used estimated LAI_{max} distributed over the year instead of remote sensed LAI; used dominant soil type for each land cover; pedotransfer functions applied to calculate soil moisture and field capacity; separate vertical water balance for each land cover
HBV	10 subcatchments	1 (no real soil depth approach)	Geometrically weighted for each subcatchment	Penman–Monteith, temperature driven monthly factors	No/no	Automatic/NSE, bias	100 m elevation intervals, percentage of elevation land use zones in each subcatchment, 5 land use classes, spatial distributed soil water storage by using AWC for each elevation and land use zone in each subcatchment; used additional interception routine
LASCAM	29 subcatchments	No	Inverse distance	Radiation based estimate assuming linear trend with latitude	No/ dynamic vegetation change	Automatic/NSE	Mean annual and monthly scaling factor for LAI, average LAI for each subcatchment
IHACRES	4 subcatchments	No	Weight Thiessen polygon	Conceptual, based on daily T_{max}	No/no	Linear module: automatic/simple refined instrumental variable approach non-linear module: manual/NSE, NSE of square root of flow, bias, arithmetic relative parameter error, lag I correlation coefficients between model error and observed stream flow and estimated effective rainfall	No distinction of grass and crops, only between open and closed vegetation, set linear module parameters using baseflow filters

HRU = Hydrologic Response Unit; NSE = Nash–Sutcliffe Efficiency [60].

Table 5
Ranking of model complexity of the LUCHEM model ensemble.

Model	Available model parameters	# calibrated parameters	# spatial computational areas	Ranking[a]
Distributed			*# of pixels*	
DHSVM	>100	3[b]	70.000	1
MIKE-SHE	>100	7	17.500	1
TOPLATS	>100	3	70.000	2[c]
WASIM	>100	30	17.500	2[c]
Semi-distributed			*# of HRU*	
SWAT	>100	6[b]	795	3
PRMS	50	5	312	4
SLURP	36 (6 per land cover)	8	252	4
HBV	10–20	10	100	5
LASCAM	30	22	29	6
Lumped			*# of subcatchments*	
IHACRES	6	5	4	7

[a] Ranking was agreed upon by all participants of the model ensemble based on subjective perception of each model.
[b] Soil parameters, ratio defined for all soils.
[c] No explicit groundwater modeling structure applied.

compensated this change mainly by a decrease in actual evapotranspiration (ET) (Fig. 6). Actual ET was also reduced for all but the WASIM model which showed a slight increase. The reaction to changing model input data can also be seen in the mean annual discharge (Fig. 7). The highest increase in discharge was found for the DHSVM model, followed by LASCAM. In contrast to this increase, discharge in the WASIM and TOPLATS models decreased by 20 and 10 mm/year, respectively. Overall, the homogenization procedure changed precipitation inputs, calculated as the mean of all models, from 929 mm for the original data set to 920 mm for the homogenized data set during the validation period. After homogenization, the standard deviation of mean annual precipitation was distinctly lower (31 mm vs. 11 mm). Runoff coefficients remained stable for both input data schemes across all models and varied by up to 0.03, except for the DHSVM model that showed an increase of the runoff coefficient from 0.42 (original data) to 0.51 (homogenized data). The latter value is more similar to all other models and we conclude that this also reflects the value of homogenizing the input data.

Although the differences for the original and homogenized data sets are in some cases substantial, modeling results with respect to the NSE are similar for both approaches (Fig. 8). NSE for the Dill catchment for the calibration period varied between 0.59 and 0.92, with the highest efficiencies simulated by the more conceptual models such as LASCAM or HBV. Absolute percent bias was reasonably low, except for the DHSVM model. A similar pattern of NSE between 0.63 and 0.91 was found for the Obere Dill, although the percent bias for SWAT, DHSVM and WASIM were higher at values around 10–20%. Simulations for the Dietzhölze subcatchment were characterized by somewhat lower NSEs between 0.53 and 0.87. The reason for this is that precipitation gauging stations in the study area are located in the lower valleys

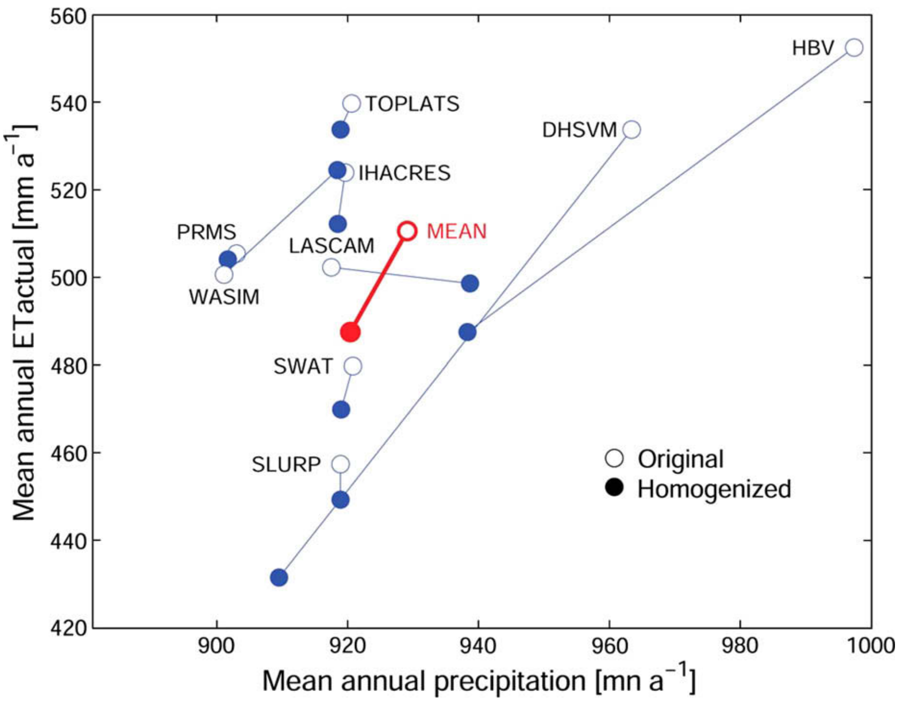

Fig. 6. Effect of homogenization of model input data on mean annual precipitation vs. actual evapotranspiration, calibration period 1983–1989.

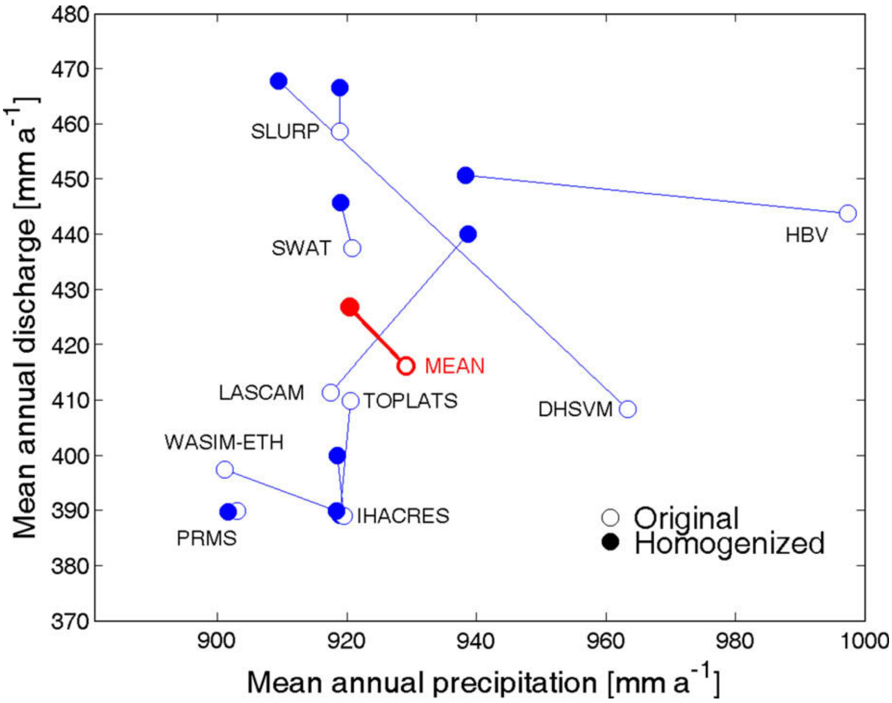

Fig. 7. Effect of homogenization of model input data on mean annual precipitation vs. discharge, calibration period 1983–1989.

(see also Table 2 and Fig. 3), which introduces a bias in the mean annual precipitation because of the underrepresented higher precipitation sums on the hills. A distinctly different pattern of the percent bias between the homogenized and original data approach was found for the Aar subcatchment (Fig. 8). Here, the percent bias for several models such as TOPLATS and WASIM considerably decreased during the homogenization of the input data, whereas the percent bias of DHSVM and SLURP substantially increased from 5% and 11% to 19%, respectively. NSEs were comparable to the Obere Dill. DHSVM was the only model where the percent bias deteriorated for all subcatchments during the homogenization.

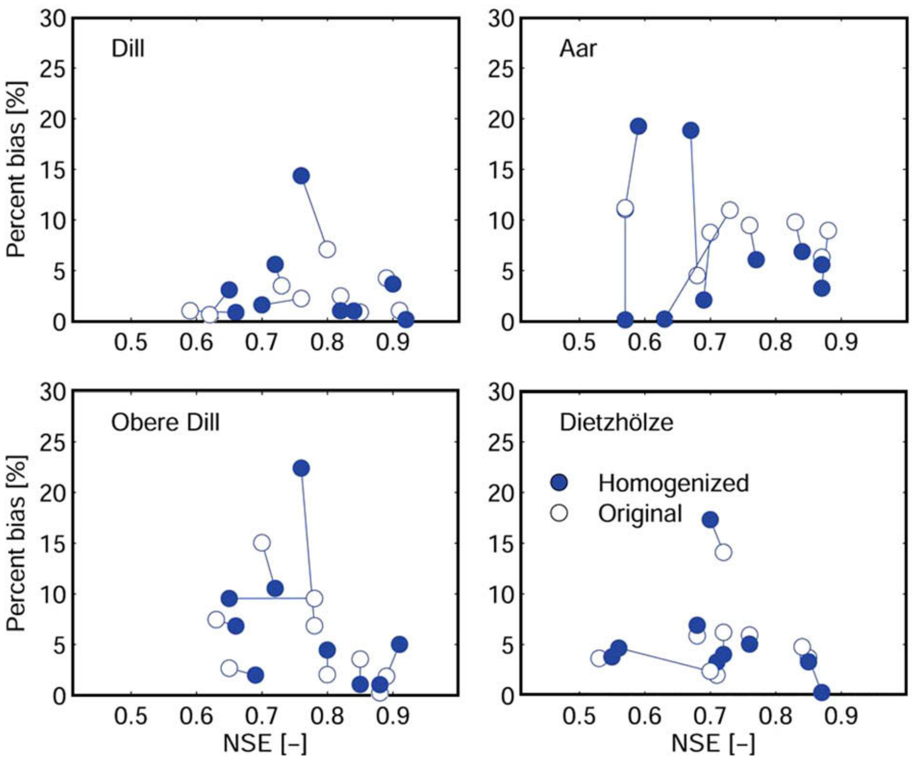

Fig. 8. Comparison of the Nash–Sutcliffe Efficiency and the change in absolute bias of the model ensemble members for the original and homogenized model approach for the Dill catchment and the three subcatchments Aar, Dietzhölze and Obere Dill, calibration period 1983–1989.

Table 6
General statistics for the Dill catchment (calibration period 01.01.1983–31.12.1989 and validation period 01.01.1990–31.12.1998) of the model ensemble members for the homogenized model approach of the Dill catchment.

Model	Mean [mm]		Sd [mm]		RMSE [mm]		Bias [mm]		Percent bias [%]		NSE [–]	
	Calibration	Validation	Calibration	Validation	Calibration	Validation	Calibration	Validation	Calibration	Validation	Calibration	Validation
DHSVM	1.52	1.33	1.64	1.66	0.922	0.629	0.191	0.233	14.4	21.3	0.76	0.85
MIKE-SHE	1.37	1.25	1.71	1.66	1.118	0.839	0.043	0.152	3.2	13.9	0.65	0.73
TOPLATS	1.32	1.07	2.04	1.80	1.105	1.011	−0.012	−0.030	0.9	2.7	0.66	0.61
WASIM	1.35	1.07	1.97	1.92	1.026	0.821	0.021	−0.028	1.6	2.5	0.70	0.74
SWAT	1.41	1.20	1.85	1.71	0.994	0.845	0.075	0.106	5.6	9.7	0.72	0.73
PRMS	1.32	1.12	1.75	1.74	0.742	0.561	−0.014	0.028	1.0	2.6	0.84	0.88
SLURP	1.37	1.27	1.33	1.31	1.106	0.876	0.041	0.172	3.1	15.7	0.65	0.71
HBV	1.33	1.16	1.87	1.67	0.540	0.444	−0.002	0.065	0.2	6.0	0.92	0.92
LASCAM	1.28	1.14	1.86	1.73	0.606	0.538	−0.049	0.050	3.7	4.5	0.90	0.89
IHACRES	1.32	1.16	1.58	1.54	0.806	0.595	−0.014	0.062	1.0	5.7	0.82	0.87
Observed	1.33	1.1	1.88	1.62	–	–	–	–	–	–	–	–

Overall, we conclude that homogenization of model input data improved the comparability of model structures and reduced the differences in model results due to different ways of interpolating precipitation and simulating plant development. Therefore, the homogenized data set is used in the remaining part of this paper, as well as in the land use scenario analyses presented by Huisman et al. [43].

5.2. Effect of calibration

Results for the calibration and validation period for all models using the homogenized data set are given in Table 6. NSE for discharge varied between 0.65 and 0.92 for the calibration period and between 0.61 and 0.92 for the validation period. Some models

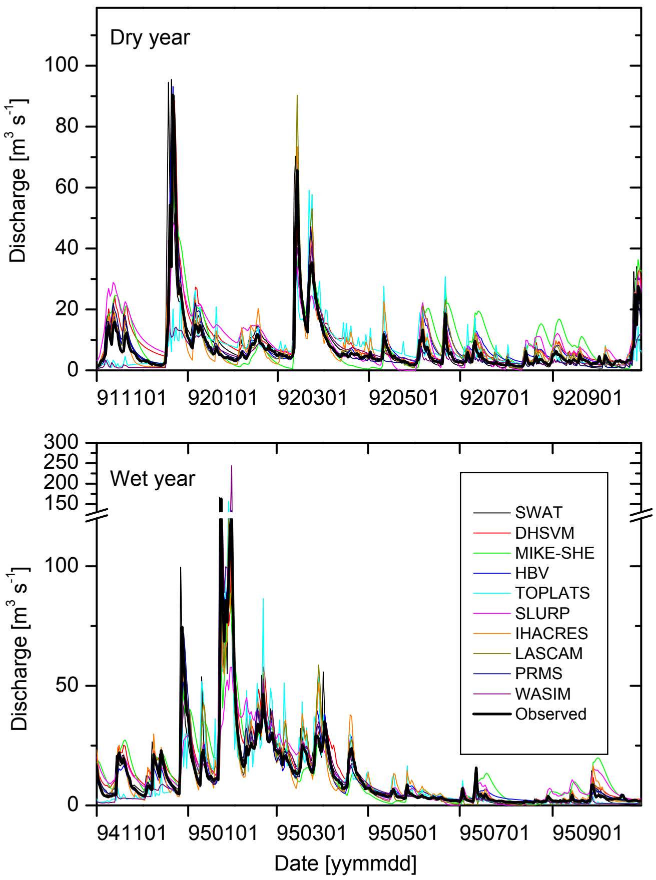

Fig. 9. Prediction of daily discharge for the Dill catchment of the model ensemble members using homogenized input data for a dry year (period 01.11.1991–30.10.1992) and a wet year (period 01.11.1994–30.10.1995).

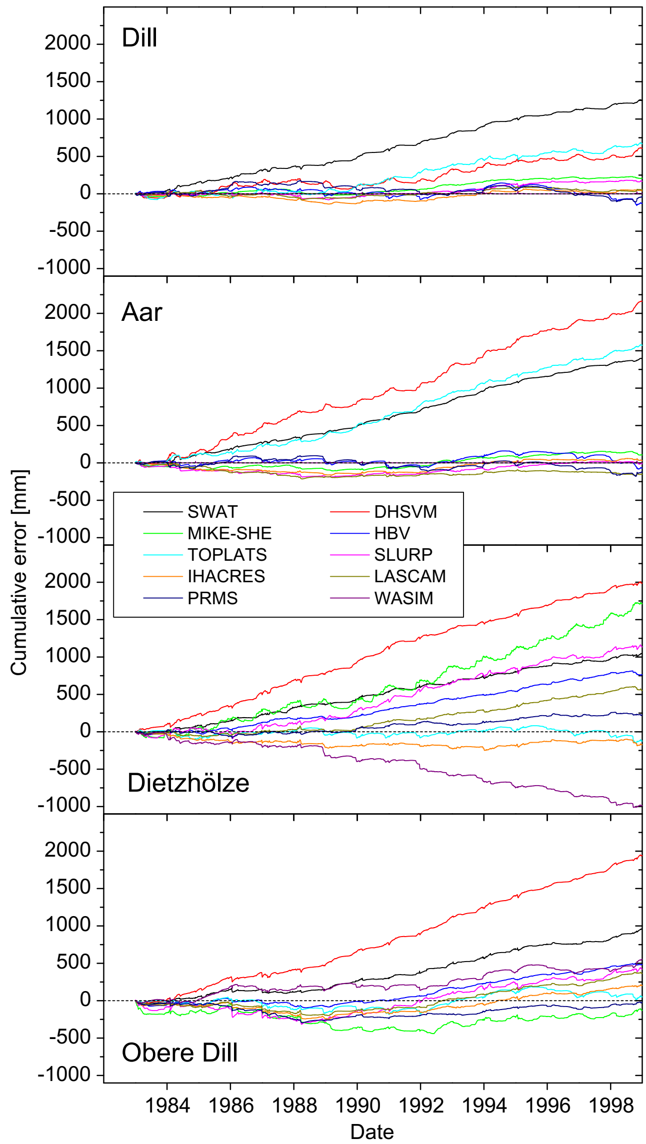

Fig. 10. Cumulative difference between observed and simulated daily discharge for the overall model period (1983–1998) for the Dill catchment and the three subcatchments Aar, Dietzhölze and Obere Dill (homogenized input data).

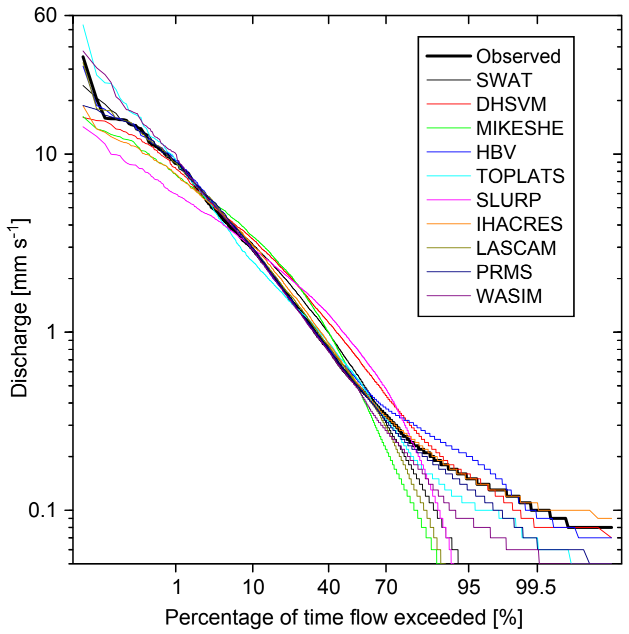

Fig. 11. Observed and predicted flow duration curves for the Dill catchment, overall model period 1983–1989, homogenized input data.

such as DHSVM, SLURP or MIKE-SHE showed an improvement for NSE going from the calibration to the validation period. This can be partly attributed to a drastic flooding event in the calibration period in February 1984 that made it difficult for these models to obtain better efficiencies during calibration.

For a comparison of model behavior for dry and wet conditions, daily discharge was plotted in Fig. 9 for the hydrologic years (01.11–31.10) 1991/1992 and 1994/1995, respectively. Simulations for dry conditions scatter more as compared to simulations for wet conditions. Especially, the SLURP and the MIKE-SHE models overestimated the low flows during November–December 1991 and June–October 1992. The WASIM and TOPLATS models both had problems in simulating discharge in late 1991 and 1994, where they underestimated discharge. Overall, TOPLATS simulated lowest and SLURP and DHSVM predicted highest discharge for the current land use using the homogenized data set.

The evaluation of the bias for all models for the entire simulation period showed that some models overestimated discharge substantially (e.g. DHSVM, TOPLATS and SWAT, see Fig. 10). Bias was lowest for the Dill catchment, which is due to the fact that models were calibrated to the discharge at the gauging station Asslar. In many cases, the more complex and fully distributed models showed a higher bias when applied to the subcatchments, such as DHSVM, SWAT and TOPLATS in the Aar, or DHSVM, SLURP and MIKE-SHE in the Dietzhölze and DHSVM and SWAT in the Obere Dill.

A comparison of the observed and predicted flow duration curve is shown in Fig. 11. Most models performed well in the range of discharges between 0.5 and 8 mm^{-1} (4–64 m^3 s^{-1}), corresponding approximately to the 5 and 70 percentiles of flows. Observed high flows are bounded by the envelope of model predictions. The SLURP model revealed a substantial underestimation of high flows. However, even larger discrepancies of observed and predicted flows were found for low flow conditions, a fact that is attributable to the focus on calibration of peak events by using, for example, NSE or RSME as objective functions (Table 4). Almost all models underestimated flow minima. Only the DHSVM, HBV and IHACRES models seemed to provide acceptable low flow predictions. Nevertheless, a closer look to discharge maxima and minima is necessary to test the behavior of the models under extreme conditions.

Fig. 12 shows observed and simulated discharge maxima for the 10 models. In general, the fully distributed models such as TOPLATS, MIKE-SHE and WASIM were less accurate than the more conceptual models such as HBV or LASCAM (Table 7). The better fitting of peaks for the conceptual models can partly be explained by the fact that they were more effectively calibrated using automatic calibration techniques, whereas the distributed physically based models were typically manually calibrated (Table 4). The models SLURP, WASIM and MIKE-SHE underestimated discharge maxima in many situations, whereas substantial under as well as overestimation was obvious for the TOPLATS model. The IHACRES model had a tendency to overestimate low discharge maxima and to underestimate high peaks. An underestimation of high peaks was also found for the DHSVM model and to a lesser extent for the PRMS model. SWAT and LASCAM, shortly followed by HBV, were the only models that did not show this underestimation of extreme flow conditions.

Observed and simulated low flow discharges were also analyzed (Fig. 13 and Table 7). Several models such as SWAT, LASCAM, SLURP and MIKE-SHE had considerable problems in simulating discharge minima, in that they all underestimated low flows. This can partially be explained by the 0.02 mm d^{-1} external inflow that some of the LUCHEM models considered whilst others such as SWAT and LASCAM did not. Most rainfall–runoff models asymptotically reduced towards zero discharge during the long dry periods. Hence, they were not able to produce a constant low flow value. Apart from an underestimation of the lowest discharges, the models HBV and DHSVM overestimated low flow discharge in many situations, which could not be observed when analyzing the flow duration curves depicted in Fig. 11 alone. From visual inspection, IHACRES and PRMS performed best during low flows. We acknowledge that some models perform better to some parts of the hydrograph than others do and have reasons to believe that better predictions could be made by combining different models for certain parts of the hydrograph. This was further investigated in the companion paper by Viney et al. [82], where different conditional multi-model ensembles that explicitly considered low flows and high flows, seasonality as well as rising and falling flows, were composed.

From the above, it is clear that we attribute part of the differences between models to differences in calibration practice. In the next section, we will revisit the results presented here and discuss them in the light of the physical differences between the models. One could argue that the way of model calibration is not an intrinsic part of the structure of a model, and that if all models had used the same calibration routines, the effects of model structure, i.e. the physical description of how water fluxes are simulated, could have been analyzed more straightforwardly. However, we feel that the way of model calibration is part of the

Table 7
Root mean squared error (RMSE) for peak flow maxima and low flow minima, Dill catchment, homogenized data approach, period 01.01.1983–31.12.1998.

	Minima [mm]	Maxima [mm]
DHSVM	0.32	27.32
MIKE-SHE	0.75	27.99
TOPLATS	0.77	41.79
WASIM	0.44	30.42
SWAT	0.40	18.31
PRMS	0.26	25.41
SLURP	0.76	34.81
HBV	0.36	14.48
LASCAM	0.37	16.39
IHACRES	0.66	25.49

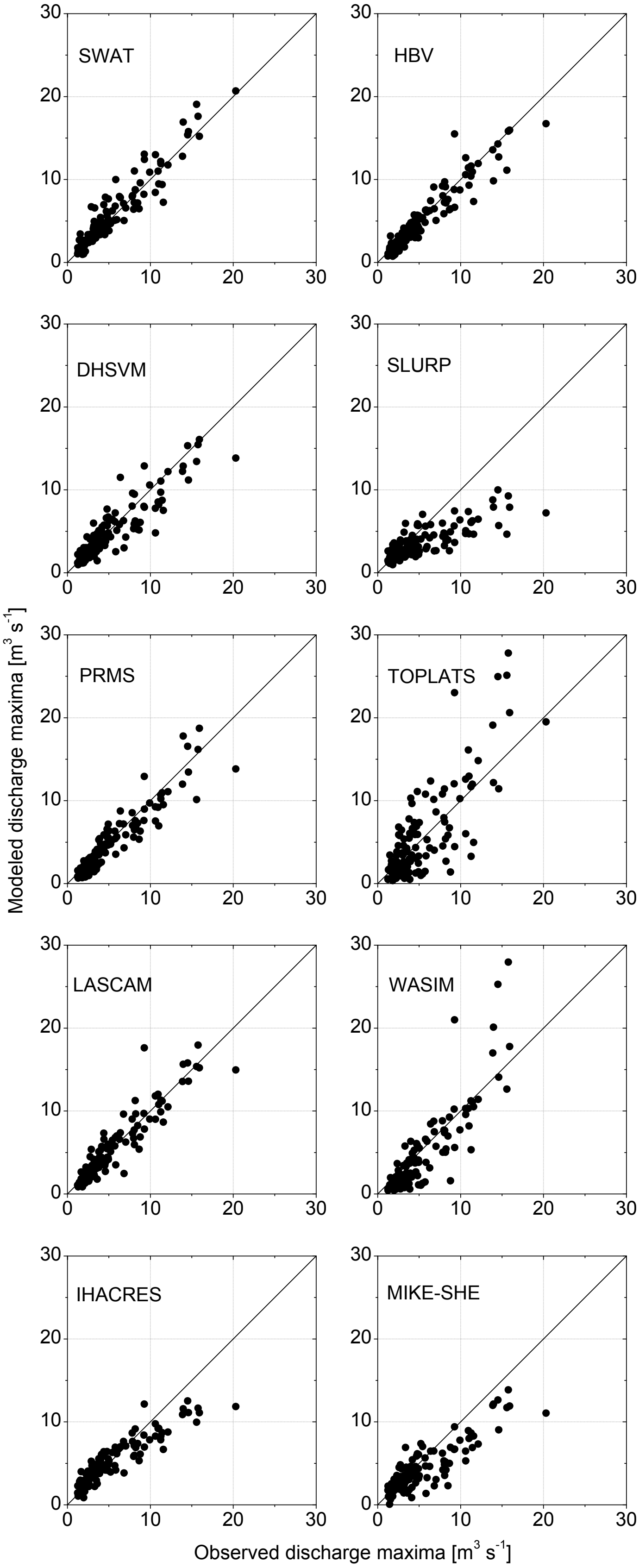

Fig. 12. Scatterplot of modeled vs. observed discharge maxima (nearly independent peak flows) for the Dill river using the homogenized data approach, overall model period 1983–1998.

model philosophy and hence, is inevitably part of the overall model structure. In addition, computation time still prohibits automatic calibration with global optimization methods for many fully distributed models, such as DHSVM or WASIM.

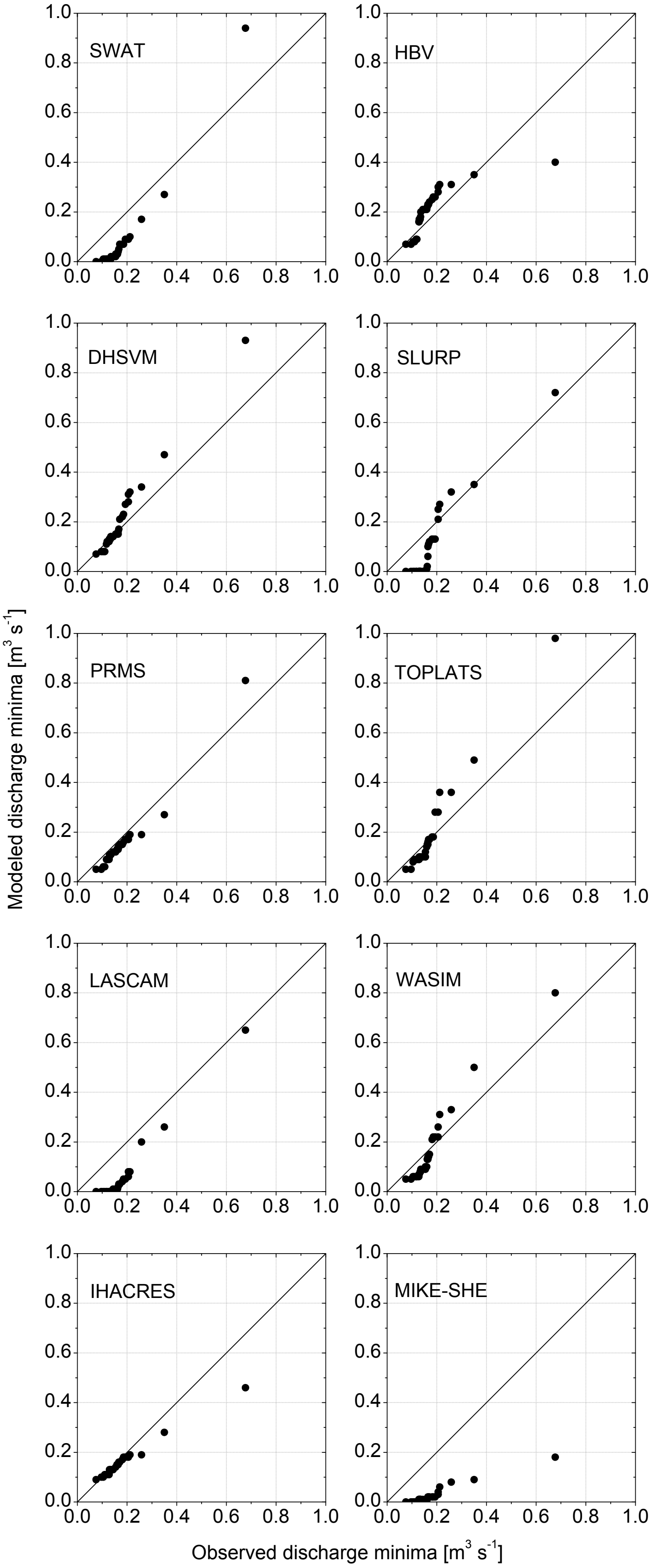

Fig. 13. Modeled vs. observed slow flow minima (nearly independent low flows), for the Dill catchment using the homogenized data approach, overall model period 1983–1998.

5.3. Physical basis of LUCHEM models

Land use and land cover impact on several components of the water balance. For example, plant physiology regulates transpiration; canopy structure determines interception storage and throughfall; rooting depth, density and structure affect plant water uptake and infiltration capacity. All these processes feed back to surface runoff and leaching, and can, therefore, potentially influence interflow and groundwater recharge. As summarized in Tables 5 and 6, the models in the LUCHEM project include a broad range of concepts to predict these processes.

For the Dill catchment, several publications indicate that the sensitivity of the long-term water balance components to belowground drivers are of secondary importance [17,44] (i.e. an increase in surface runoff at the cost of interflow will not strongly affect actual ET, which results in a low sensitivity of the long-term water balance). Therefore, we first analyze the differences in ET between the models. The simulated multi-model average was 490 mm for actual ET as compared to an average discharge of 430 mm. No field measurements were available for the evaluation of predicted actual ET rates. Therefore, model performance was assessed qualitatively. Predicted ET rates differed substantially for the 10 models, which might be attributed to the varying spatial resolution of the models as well as the different concepts of modeling potential and actual ET.

The spatial resolution of the models varied from 1 ha to several km^2 (Table 4). As is shown in a companion paper by Bormann et al. [13], the impact of different spatial resolution of model input data on simulated actual ET rates for the SWAT, TOPLATS and WASIM models is rather small, except when the aggregation level is increased beyond 1 km^2. Hence, we conclude that differences in ET cannot be explained by differences in the spatial resolution of the models. Besides the spatial resolution, the aggregation of different land use classes to more general land use classes might have also impacted simulated ET rates. For example, HBV aggregates cropland and pasture to an 'open vegetation' land use class. However, we feel that potential differences in actual ET are removed rather than enhanced by such an aggregation.

Since spatial resolution and aggregation were found to be of secondary importance, we argue that the variation in simulated actual ET is mainly introduced by the use of different methods to calculate potential and actual ET (Table 4). For example, IHACRES estimates actual ET rates based on calibrated parameters [26], catchment moisture and temperature or potential evaporation data. In contrast, detailed meteorological data and plant specific parameters are needed to calculate ET according to Penman–Monteith, an approach that is implemented in several LUCHEM model members. Even though we homogenized LAI input for the different land use classes as described above, differences in the Penman–Monteith methods still remained, as there is no official agreement on how to implement it.

At this point, it is justified to ask which models provide the most plausible simulations of ET. Unfortunately, it is not possible to answer this directly from the model intercomparison presented here. This would require a detailed comparison of measured and simulated ET for individual land covers. This analysis is hampered by the lack of measured ET data and the inability of some models to provide simulated ET for individual land covers without substantial recoding. However, the analysis of the land use change scenarios by Huisman et al. [43] provided additional insights. They argue that the possibility to compare various scenario predictions in addition to the current land use helped to explain the diversity in model performance. For example, the scenario analysis showed that the actual ET simulations for pasture were high for TOPLATS compared to the other models and land covers. Since simulated discharge was not obviously different for TOPLATS, this indicates that different mean actual ET rates for different land covers can compensate each other and lead to similar discharge predictions. Only in the case of the varying land cover proportions in the scenario analysis, the high ET simulations for pasture became apparent. For the actual land use distribution, it would have been advantageous if the subcatchments had varied more in terms of land cover, because this would also have allowed a better plausibility check on the simulated ET rates for each land use. This is a point of attention for future model intercomparison studies.

As argued above, runoff generation does not strongly affect the long-term water balance. However, the timing and magnitude of particular runoff events obviously does depend on the runoff generation mechanism. The range of concepts to simulate runoff generation is large. For example, MIKE-SHE, WASIM and TOPLATS use process-based descriptions of water infiltration into the soil, whereas other models rely on more conceptual representations based on storage capacity or soil wetness. In addition, some models such as LASCAM separate between infiltration excess and saturation excess runoff, whereas other models only consider saturation excess runoff (e.g. HBV). During the workshop, we collected information on the flow components simulated by each model (i.e. surface runoff, interflow, baseflow). However, this information was not useful because of the varying time-scales associated with the reported flow components (i.e. interflow in SWAT might be baseflow in another model). In addition, some of the models did not explicitly consider interflow processes. To overcome this, we applied the WETSPRO filter [89] to all simulated discharge time series. Using this filter, three different flow components were distinguished (slow, intermediate and quick flow). Fig. 14 shows that variation in flow components is only moderate, which seems reasonable in the light of the good performance of most models. The simulated slow flow component ranges from 30% to 50%, whereas the slow flow fraction of the measured time series was ~35%. The two models with highest slow flow component and the lowest intermediate flow component (DHSVM and TOPLATS)

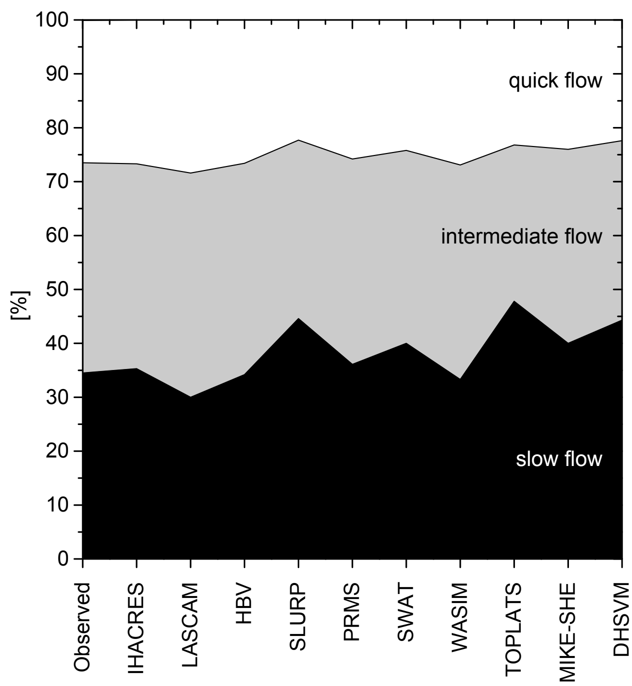

Fig. 14. Flow components for observed and predicted discharge in order of model complexity (see Table 5). Flow components were calculated by the use of the WETSPRO tool [89].

only consider two possible pathways of water (i.e. surface runoff and 'ground water' flow).

The variation in the amount of quick flow is low. As argued before, this partly reflects the emphasis on high flows during both manual and automatic model calibration. The results for discharge maxima presented in Fig. 12 should also be viewed in the light of physical differences between the models. The underestimation of discharge maxima is reflected in the lowest fast flow contribution for SLURP (Fig. 14), perhaps indicating an inadequate parameterization of the non-linear storage module for the top soil. Furthermore, it seems no coincidence that the models with the most physically based description of infiltration (TOPLATS, WASIM and MIKE-SHE) perform less well for discharge maxima. The use of physically based concepts using soil properties estimated from pedotransfer functions or laboratory measurements remains notoriously problematic at the scale of a landscape.

6. Conclusions

A detailed model intercomparison is a prerequisite for the set-up of a multi-model ensemble. In this paper, we paved the ground for such a multi-model ensemble to investigate the effects of land use change on hydrology in the frame of the LUCHEM project. Overall, the more conceptual models, especially HBV, LASCAM and PRMS, outperformed the physically based, fully distributed models such as DHSVM, TOPLATS, WASIM and MIKE-SHE. However, we conclude that the presented model intercomparison gave no reason to exclude any model, and that the entire set of models is valuable to build further multi-model ensembles and to analyze land use scenarios, topics that are presented in subsequent papers of this special section.

It is possible that some of the simulated differences between the models are insignificant given the typical uncertainty in model parameterization. However, a further quantification of stochastic uncertainty of individual models and a comparison of the differences that evolve from structural uncertainty was beyond the scope of this study, but a very interesting topic for future investigations.

We conclude that differences in model performance in this model intercomparison were attributable to the effects of model input data, the effect of model calibration, and the physical basis of the models:

– Depending on the spatial support and the associated data aggregation and interpolation, models deal differently with model input data, which leads to differences in model output. We achieved a better comparison of model results by homogenizing important model input data such as precipitation, climatic data and LAI. Nevertheless, homogenization is limited in its effects as the various models inherently treat spatial data differently. However, forcing models to use comparable model inputs can help avoid the drawing of wrong conclusions with respect to model structural effects.
– We argue that model calibration is an inevitable component of the model structure and its underlying philosophy. We therefore neither considered prescribing calibration routines nor objective functions for the different models. Following this, the fully distributed models utilized manual calibration schemes, whereas the more conceptual models were calibrated automatically, some of them with various objective functions (HBV), or by the use of multi-criteria calibration (LASCAM, IHACRES). We recognize that this view is debatable. Future model intercomparison projects should try to quantify this effect by prescribing fixed calibration routines. However, this would force some of the distributed models to be applied in a frame that they were not intended for, and which takes their basic philosophy of *a priori* parameterization *ad absurdum*.
– Structural differences that were responsible for differences in the long-term water balance were attributed to differences in ET for the various land use types. However, the identification of single cause–effect relationships between soil and vegetation for each land use type was not possible in this experimental set-up. To resolve the intrinsic interaction between soil water status and actual ET for single land use types would require an alternative modeling approach. We could think of applying the models in a type of 'lysimeter experiment', where lateral flow is excluded, so only vertical flux components of the participating models could be compared. Structural differences in runoff generation mechanism also affected the timing and magnitude of events. Despite an intertwining with the calibration intensity, it seemed that the models using a physically based representation of infiltration performed less well for the discharge maxima.

One should bear in mind that the quality of model predictions presented in the LUCHEM project reflect not only the appropriateness of the model structure and parameter estimation schemes, but also the skill, experience, and the time the individual modelers were able to devote to this project. A crucial next step in land use change studies is to investigate catchments where real land use changes have been documented and monitored during the change. Catchments with different causes of land use change such as storm damages, forest fires, land abandonment or large scale reforestation or clearing are considered to provide valuable case studies.

Acknowledgement

This study has been supported by the Deutsche Forschungsgemeinschaft within the scope of the Collaborative Research Centre (SFB) 299.

References

[1] Abbott MB, Bathurst JC, Cunge JA, ÓConnell PE, Rasmussen J. An introduction to the European Hydrological System – Système Hydrologique Européen, 'SHE'. 1: History and philosophy of a physically based distributed modelling system. J Hydrol 1986;87:45–59.
[2] Ajami NK, Duan Q, Sorooshian S. An integrated hydrologic Bayesian multimodel combination framework: confronting input, parameter, and model structural uncertainty in hydrologic prediction. Water Resour Res 2007:43. doi:10.1029/2005WR004745.
[3] Al-Khudairy DHA, Thompson RR, Gavin H, Hamm NAS. Hydrological modelling of a drained grazing marsh under agricultural land use and the simulation of restoration management scenarios. Hydrol Sci J 1999;44:943–71.
[4] Arnold JG, Srinivasan R, Muttiah RS, Williams JR. Large area hydrologic modeling and assessment. Part I: Model development. J Am Water Res Assoc 1998;341:73–88.
[5] Band LE, Peterson DJ, Running SW, Coughlan JC, Lammers R, Dungan J, et al. Forest ecosystem processes at the watershed scale: basis for distributed simulation. Ecol Model 1991;56:171–96.
[6] Baron JS, Hartman MD, Kittel TGF, Band LE, Ojima DS, Lammers RB. Effects of land cover, water redistribution, and temperature on ecosystem processes in the South Platte Basin. Ecol Appl 1998;84:1037–51.
[7] Bergström S. The HBV model. In: Singh VP, editor. Computer models of watershed hydrology. Highland Ranch, CO, USA: Water Resources Publications; 1995. p. 443–76.
[8] Beven K, Binley A. The future of distributed models: model calibration and uncertainty prediction. Hydrol Proc 1992;6:279–98.
[9] Beven KJ. Rainfall–runoff modelling. The primer. Chichester: Wiley; 2001.
[10] Binley AM, Beven KJ, Calver A, Watts LG. Changing responses in hydrology: assessing the uncertainty in physically based model predictions. Water Resour Res 1991;27:1253–61.
[11] Bosch JM, Hewlett JD. A review of catchment experiments to determine the effect of vegetation changes on water yield and evapotranspiration. J Hydrol 1982;55:3–23.
[12] Bormann H. Impact of spatial data resolution on simulated catchment water balances and model performance of the multi-scale TOPLATS model. Hydrol Earth Syst Sci 2006;10:165–79.
[13] Bormann H, Breuer L, Croke B, Gräff T, Huisman JA. Assessing the impact of land use change on hydrology by ensemble modelling (LUCHEM). IV: Model

sensitivity on data aggregation and spatial (re-)distribution. Adv Water Resour 2008 [this issue].
[14] Boyle DP, Gupta HV, Sorooshian S, Koren V, Zhang Z, Smith M. Toward improved streamflow forecasts: value of semidistributed modeling. Water Resour Res 2001;37:2749–59.
[15] Breuer L, Eckhardt K, Frede H-G. Plant parameter values for models in temperate climates. Ecol Model 2003;169:237–93.
[16] Breuer L, Huisman JA, Frede H-G. The European Common Agricultural Policy and feasible hydrological impacts for the Dill catchment, Germany. In: Proceedings of the 3rd SWAT conference, 2005, 8 pp.
[17] Breuer L, Huisman JA, Frede H-G. Monte Carlo assessment of uncertainty in the simulated hydrological response to land use change. Environ Model Assess 2006;11:209–18.
[18] Breuer L, Huisman JA, Steiner N, Weinmann B, Frede H-G. Eco-hydrologic and economic trade-off functions in watershed management. 2004. TWRI Technical Report 266. In: Proceedings of the 2nd SWAT conference, 2003, p. 255–61.
[19] Brown AE, Zhang L, McMahon TA, Western AW, Vertessy RA. A review of paired catchment studies for determining changes in water yield resulting from alterations in vegetation. J Hydrol 2005;310:28–61.
[20] Calder IR. Assessing the water use of short vegetation and forests: development of the Hydrological Land Use Change (HYLUC) model. Water Res 2003. doi:10.1029/2003WR002040.
[21] Calder IR, Hall RL, Bastable HG, Gunston HM, Shela O, Chirwa A, et al. The impact of land use change on water resources in sub-Saharan Africa: a modelling study of Lake Malawi. J Hydrol 1995;170:123–35.
[22] Carpenter TM, Georgakakos KP. Intercomparison of lumped versus distributed hydrologic model ensemble simulations on operational forecast scales. J Hydrol 2006;329:174–85.
[23] Chapman T. Comment on "Evaluation of automated techniques for base flow and recession analyses" by R.J. Nathan and T.A. McMahon. Water Resour Res 1991;277:1783–4.
[24] Christensen NS, Lettenmaier DPA. Multimodel ensemble approach to assessment of climate change impacts on the hydrology and water resources of the Colorado River Basin. Hydrol Earth Syst Sci 2007;11:1417–34.
[25] Christiaens K, Feyen J. Use of sensitivity and uncertainty measures in distributed hydrological modeling with an application to the MIKE SHE model. Water Resour Res 2002. doi:10.1029/2001WR000478.
[26] Croke BFW, Jakeman AJ. A catchment moisture deficit module for the IHACRES rainfall–runoff model. Environ Modell Softw 2004;19:1–5.
[27] Croke BFW, Merritt WS, Jakeman AJ. A dynamic model for predicting hydrologic response to land cover changes in gauged and ungauged catchments. J Hydrol 2004;291:115–31.
[28] Diekkrüger B, Söndgerath D, Kersebaum KC, McVoy CW. Validity of agroecosystem models. A comparison of results of different models applied to the same data set. Ecol Model 1995;81:3–29.
[29] Duan QY, Ajami NK, Gao XG, Sorooshian S. Multi-model ensemble hydrologic prediction using Bayesian model averaging. Adv Water Resour 2007;30:1371–86.
[30] Duan Q, Schaake J, Andréassian V, Franks S, Goteti G, Gupta GV, et al. Model Parameter Estimation Experiment (MOPEX): an overview of science strategy and major results from the second and third workshops. J Hydrol 2006;320:3–17.
[31] Eckhardt K, Breuer L, Frede H-G. Parameter uncertainty and the significance of simulated land use change effects. J Hydrol 2003. doi:10.1016/S0022-1694(02)00395-5.
[32] Ewen J, Parkin G. Validation of catchment models for prediction land use and climate change impact. 1: Method. J Hydrol 1996;175:583–94.
[33] Famiglietti JS, Wood EF. Multiscale modeling of spatially variable water and energy balance processes. Water Resour Res 1994;30:3061–78.
[34] Fohrer N, Haverkamp S, Eckhardt KS, Frede H-G. Hydrologic response to land use changes on the catchment scale. Phys Chem Earth B 2001. 267–8, 577–82.
[35] Fohrer N, Haverkamp S, Frede H-G. Assessment of long-term effects of land use patterns on hydrologic landscape functions – sustainable land use concepts for low mountain range areas. Hydrol Process 2005;193:659–72.
[36] Fohrer N, Möller D, Steiner N. An interdisciplinary modelling approach to evaluate the effects of land use change. Phys Chem Earth B 2002;27:655–62.
[37] Georgakakos KP, Seo D-J, Gupta H, Schaake J, Butt MB. Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. J Hydrol 2004;298:222–41.
[38] Gourley JJ, Vieux BE. A method for identifying sources of model uncertainty in rainfall–runoff simulations. J Hydrol 2006;327:68–80.
[39] Henderson-Sellers A, Fischer M, Aleinov I, McGuffie K, Riley WJ, Schmidt GA, et al. Stable water isotope simulation by current land-surface schemes: results of iPILPS Phase 1. Glob Planet Change 2006;51:34–58.
[40] Henderson-Sellers A, McGuffie K, Pitman AJ. The Project for Intercomparison of Land-surface Parametrization Schemes (PILPS): 1992–1995. Clim Dynam 1996;12:849–59.
[41] HLBG. Hessische Verwaltung für Bodenmanagement und Geoinformation. Digitales Geländemodell DGM25; 2000. <http://www.hkvv.hessen.de/>.
[42] HLUG (Hessisches Landesamt für Umwelt und Geologie). Digital soil map from Hessen, 1:50,000; 2000.
[43] Huisman JA, Breuer L, Bormann H, Bronstert A, Croke BFW, Frede H, et al. Assessing the impact of land use change on hydrology by ensemble modelling (LUCHEM). III: Scenario analysis. Adv Water Resour 2007 [this issue].
[44] Huisman JA, Breuer L, Frede H-G. Sensitivity of simulated hydrological fluxes towards changes in soil properties in response to land use change. Phys Chem Earth 2004;29:749–58.
[45] IUSS Working Group WRB. World reference base for soil resources 2006, 2nd ed. World Soil Resources Reports No. 103. FAO, Rome; 2006.
[46] Jakeman AJ, Littlewood IG, Whitehead PG. Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments. J Hydrol 1990;117:275–300.
[47] Kaviany E. Zur Hydrogeologie im Niederschlagsgebiet der Dill (Hessen). In: Knoblich K, Strübel G, editors. Gießener Geologische Schriften, vol. 1. Gießen: Lenz Verlag; 1978. p. 91–250.
[48] Kite G. Modelling the Mekong: hydrological simulation for environmental impact studies. J Hydrol 2001;253:1–13.
[49] Kite GW. Development of a hydrological model for a Canadian watershed. Can J Civil Eng 1978;51:126–34.
[50] Krysanova V, Hattermann F, Wechsung F. Development of the ecohydrological model SWIM for regional impact studies and vulnerability assessment. Hydrol Process 2005;19:763–83.
[51] Krysanova V, Muller-Wohlfeil DI, Becker A. Development and test of a spatially distributed hydrological water quality model for mesoscale watersheds. Ecol Model 1998;1062–1063:261–89.
[52] Kuhlmann F, Möller D, Weinmann B. Modellierung der Landnutzung – Regionshöfe oder Raster-Landschaft? Ber Landwirtsch 2002;803:351–92.
[53] Leavesley GH, Stannard LG. In: Singh VP, editor. The precipitation runoff modeling system – PRMS. Highlands Ranch, CO: Water Resources Publications; 1995. p. 281–310 [Chapter 9].
[54] Lindström G, Johansson B, Persson M, Gardelin M, Bergström S. Development and test of the distributed HBV-96 hydrological model. J Hydrol 1997;201:272–88.
[55] Luo LF, Wood EF, Pan M. Bayesian merging of multiple climate model forecasts for seasonal hydrological predictions. J Geophys Res 2007;112:D10102.
[56] McIntyre N, Lee H, Wheater H, Young A, Wagener T. Ensemble predictions of runoff in ungauged catchments. Water Resour Res 2005. doi:10.1029/2005WR004289.
[57] Melching CS. Reliability estimation. In: Singh VP, Frevert DK, editors. Mathematical models of large watershed hydrology. Highlands Ranch: Water Resources Publications; 2002. p. 69–118 [Chapter 3].
[58] Möller D, Fohrer N, Steiner N. Quantifizierung regionaler Multifunktionalität land-und forstwirtschaftlicher Nutzungssysteme. Ber Landwirtsch 2002;803:393–418.
[59] Murphy JM, Sexton DMH, Barnett DN, Jones GS, Webb MJ, Collins M, et al. Quantification of modeling uncertainties in a large ensemble of climate change simulations. Nature 2004;430:768–72.
[60] Nash JE, Sutcliffe JV. River flow forecasting through conceptual models. Part I: A discussion of principles. J Hydrol 1970;10:282–90.
[61] Niehoff D, Fritsch U, Bronstert A. Land-use impacts on storm-runoff generation: scenarios of land-use change and simulation of hydrological response in a meso-scale catchment in SW-Germany. J Hydrol 2002;267:80–93.
[62] Nielsen SA, Hansen E. Numerical simulation of the rainfall–runoff process on a daily basis. Nordic Hydrol 1973;4:171–90.
[63] Pitman AJ, Henderson-Sellers A. Recent progress and results from the project for the intercomparison of landsurface parameterization schemes. J Hydrol 1998;213:128–35.
[64] Onstad CA, Jamieson DG. Modelling the effect of land use modifications on runoff. Water Resour Res 1970;65:1287–95.
[65] Ott B, Uhlenbrook S. Quantifying the impact of land-use changes at the event and seasonal time scale using a process-oriented catchment model. Hydrol Earth Syst Sci 2004;81:62–78.
[66] Palmer TN. Special issue on the DEMETER project (Development of a Multi-model Ensemble System for Seasonal to Interannual Climate Prediction). Tellus A 2005;57:217–512.
[67] Peters-Lidard CD, Zion MS, Wood EF. A soil–vegetation–atmosphere transfer scheme for modeling spatially variable water and energy balance processes. J Geophys Res 1997;102:4303–24.
[68] Reed S, Koren V, Smith M, Zhang Z, Moreda F, Seo D-J, et al. Overall distributed model intercomparison project results. J Hydrol 2004;298:27–60.
[69] Refsgaard JC, Knudsen J. Operational validation and intercomparison of different types of hydrological models. Water Resour Res 1996;32:2189–202.
[70] Refsgaard JC, Storm B. MIKE SHE. In: Singh VP, editor. Computer models of watershed hydrology. Highlands Ranch, CO: Water Resources Publications; 1995. p. 809–46 [Chapter 23].
[71] Refsgaard JC, van der Sluijs J, Brown J, van der Keur P. A framework for dealing with uncertainty due to model structure error. Adv Water Resour 2006;29:1586–97.
[72] Rykiel Jr EJ. Testing ecological models: the meaning of validation. Ecol Model 1996;90:229–44.
[73] Schulla J. Hydrologische Modellierung von Flussgebieten zur Abschätzung der Folgen von Klimaänderungen. Zürich. Zürcher Geographische Schriften; 1997.
[74] Shamseldin AY, O'Connor KM, Liang GC. Methods for combining the outputs of different rainfall–runoff models. J Hydrol 1997;197:203–29.
[75] Sivapalan M. Prediction in ungauged basins: a grand challenge for theoretical hydrology. Hydrol Proc 2003;17:3163–70.
[76] Sivapalan M, Ruprecht JK, Viney NR. Water and salt balance modelling to predict the effects of land-use changes in forested catchments.1. Small catchment water balance model. Hydrol Process 1996;10:393–411.

[77] Smith MB, Georgakakos KP, Liang X. Special issue: the distributed model intercomparison project (DMIP). J Hydrol 2004;298:1–335.
[78] Smith MB, Seo D-J, Koren VI, Reed SM, Zhang Z, Duan Q, et al. The distributed model intercomparison project (DMIP): motivation and experiment design. J Hydrol 2004;298:4–26.
[79] Stednick JD. Monitoring the effects of timber harvest on annual water yield. J Hydrol 1996;176:79–95.
[80] Takle ES, Jha M, Anderson CJ. Hydrological cycle in the upper Mississippi River basin: 20th century simulations by multiple GCMs. Geophys Res Lett 2005;32:L18407.
[81] VanShaar JR, Haddeland I, Lettenmaier DP. Effects of land-cover changes on the hydrological response of interior Columbia River basin forested catchments. Hydrol Process 2002;16:2499–520.
[82] Viney NR, Bormann H, Breuer L, Bronstert A, Croke BFW, Frede H-G, et al. Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM). II: Ensemble combinations and predictions. Adv Water Resour 2007 [this issue].
[83] Vrugt JA, Diks CGH, Gupta HV, Bouten W, Verstraten JM. Improved treatment of uncertainty in hydrologic modeling: combining the strength of global optimization and data assimilation. Water Resour Res 2005. doi:10.1029/2004WR003059.
[84] Vrugt JA, Gupta HV, Bouten W, Sorooshian S. A shuffled complex evolution algorithm for optimization and uncertainty assessment of hydrological parameters. Water Resour Res 2003. doi:10.1029/2002WR001642.
[85] Vrugt JA, Robinson BA. Treatment of uncertainty using ensemble methods: comparison of sequential data assimilation and Bayesian model averaging. Water Resour Res 2007;43:W01411.
[86] Weber A, Fohrer N, Möller D. Long-term land use changes in a mesoscale watershed due to socio-economic factors – effects on landscape and functions. Ecol Model 2001;140:125–40.
[87] Weinmann B, Schroers JO, Sheridan P. Simulating the effects of decoupled transfer payments using the land use model ProLand. Agrarwirtschaft 2006;55:248–56.
[88] Wigmosta MS, Vail LW, Lettenmaier DP. A distributed hydrology vegetation model for complex terrain. Water Resour Res 1994;306:1665–79.
[89] Willems P. A time series tool to support the multi-criteria performance evaluation of rainfall–runoff models. Environ Modell Softw, 2008. doi:10.1016/j.envsoft.2008.09.005.
[90] Williams JR, Jones CA, Dyke PT. A modeling approach to determining the relationship between erosion and soil productivity. Trans Am Soc Agric Eng 1984;27:129–44.