Hydrol. Earth Syst. Sci., 18, 2065–2085, 2014
www.hydrol-earth-syst-sci.net/18/2065/2014/
doi:10.5194/hess-18-2065-2014

Hydrology and Earth System Sciences
Open Access

Impact of modellers' decisions on hydrological a priori predictions

H. M. Holländer[1,2], H. Bormann[3], T. Blume[4], W. Buytaert[5], G. B. Chirico[6], J.-F. Exbrayat[7,8,9], D. Gustafsson[10], H. Hölzel[1], T. Krauße[11], P. Kraft[6], S. Stoll[12], G. Blöschl[13], and H. Flühler[14]

[1]Chair of Hydrology and Water Resources Management, Brandenburg University of Technology, Cottbus, Germany
[2]Department of Civil Engineering, University of Manitoba, Winnipeg, Canada
[3]Department of Civil Engineering, University of Siegen, Siegen, Germany
[4]GFZ German Research Centre for Geosciences, Potsdam, Germany
[5]Department of Civil and Environmental Engineering and Grantham Institute for Climate Change, Imperial College London, London, UK
[6]Department of Agricultural Engineering, University di Naples Federico II, Naples, Italy
[7]Institute for Landscape Ecology and Resources Management, University of Giessen, Giessen, Germany
[8]Climate Change Research Centre and ARC Centre of Excellence for Climate System Science, University of New South Wales, Sydney, New South Wales, Australia
[9]School of GeoSciences and National Centre for Earth Observation, University of Edinburgh, Edinburgh, UK
[10]Department of Land and Water Resources Engineering, Royal Institute of Technology KTH, Stockholm, Sweden
[11]Institute of Hydrology and Meteorology, University of Technology Dresden, Dresden, Germany
[12]Institute of Environmental Engineering, ETH Zurich, Zurich, Switzerland
[13]Institute of Hydraulic Engineering and Water Resources Management, TU Vienna, Vienna, Austria
[14]Department of Environmental System Sciences, ETH Zurich, Zurich, Switzerland

Correspondence to: H. M. Holländer (hartmut.hollaender@umanitoba.ca)

Received: 12 June 2013 – Published in Hydrol. Earth Syst. Sci. Discuss.: 10 July 2013
Revised: 15 April 2014 – Accepted: 17 April 2014 – Published: 4 June 2014

Abstract. In practice, the catchment hydrologist is often confronted with the task of predicting discharge without having the needed records for calibration. Here, we report the discharge predictions of 10 modellers – using the model of their choice – for the man-made Chicken Creek catchment (6 ha, northeast Germany, Gerwin et al., 2009b) and we analyse how well they improved their prediction in three steps based on adding information prior to each following step. The modellers predicted the catchment's hydrological response in its initial phase without having access to the observed records. They used conceptually different physically based models and their modelling experience differed largely. Hence, they encountered two problems: (i) to simulate discharge for an ungauged catchment and (ii) using models that were developed for catchments, which are not in a state of landscape transformation. The prediction exercise was organized in three steps: (1) for the first prediction the modellers received a basic data set describing the catchment to a degree somewhat more complete than usually available for a priori predictions of ungauged catchments; they did not obtain information on stream flow, soil moisture, nor groundwater response and had therefore to guess the initial conditions; (2) before the second prediction they inspected the catchment on-site and discussed their first prediction attempt; (3) for their third prediction they were offered additional data by charging them pro forma with the costs for obtaining this additional information.

Holländer et al. (2009) discussed the range of predictions obtained in step (1). Here, we detail the modeller's assumptions and decisions in accounting for the various processes. We document the prediction progress as well as the learning process resulting from the availability of added information. For the second and third steps, the progress in prediction quality is evaluated in relation to individual modelling experience and costs of added information.

In this qualitative analysis of a statistically small number of predictions we learned (i) that soft information such as the modeller's system understanding is as important as

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

the model itself (hard information), (ii) that the sequence of modelling steps matters (field visit, interactions between differently experienced experts, choice of model, selection of available data, and methods for parameter guessing), and (iii) that added process understanding can be as efficient as adding data for improving parameters needed to satisfy model requirements.

1 Introduction

Predicting hydrological variables in ungauged catchments is one of the major challenges in hydrological sciences (Sivapalan et al., 2003). The success – or equivalently the uncertainty – of predicting the hydrological response of an ungauged catchment to external driving forces depends (i) on the quality and abundance of catchment data, (ii) on the availability of suitable models (Beven, 1999), (iii) on the existence of comparable catchments, and (iv) on the modeller her- or himself. In science one often tends to believe that prediction is an objective projection based on "hard" author-independent information. Seibert and McDonnell (2002) showed that improving the dialogue between modeller and experimentalist yields better predictions of the overall catchment behaviour because it brings in soft information such as a more realistic process understanding. Here, we look at how a group of modellers with similar system knowledge but with different modelling experience and philosophy address the problem of predicting a catchment response, while the observed response was made unavailable to them.

To estimate discharge from an ungauged catchment one has to address three major uncertainty sources: (i) the model and its structure, (ii) model parameters, and (iii) model inputs (initial and boundary conditions) (Blöschl, 2006). Most previous model comparisons focused on identifying the relative merits of alternative approaches to these three issues. Intercomparison studies in gauged catchments generally tested whether a particular model structure is superior relative to others (e.g. Naef, 1981; Goodrich, 1990; Reed et al., 2004; Breuer et al., 2009).

In the case of ungauged catchments the model comparisons often aimed at optimizing the methods of parameter estimation (Parajka et al., 2005; Oudin et al., 2008), as for instance using pedotransfer functions to guess the hydraulic soil parameters (Wösten et al., 2001). The role of the modeller was not systematically investigated in these studies or was even intentionally disregarded making the intercomparison as objective as possible assuming the modellers are interchangeable. The results by Holländer et al. (2009) and Bormann et al. (2011) indicate that the modeller per se is an intrinsic part of a modelling study and has a major bearing on the modelled results, even more so in ungauged catchments because of the more numerous degrees of freedom in making modelling decisions. Here we analyse the role of the modeller in repeatedly predicting the response of a particular catchment based on a stepwise improved database.

This is done by pretending that the artificial catchment "Chicken Creek" (Gerwin et al., 2009b) is ungauged. The 6 ha catchment was newly constructed and vegetation-free, and therefore it changed its structure in the course of the modelling period. The discharge record was only known to the organizers of this prediction exercise. Ten modellers were invited to predict the discharge from this catchment. In a first step, all of them received the same data set. They submitted their first-stage discharge prediction without having had the possibility to visit the catchment (Holländer et al., 2009). After presenting their first results to each other at a workshop and visiting the field site, they re-did their prediction using the same data set (second-stage prediction). For the third-stage prediction the modellers were offered additional data. They were then charged, in a virtual sense, with the costs of the parameters they actually selected. The expected improvement of the prediction quality could then be related to the additional investment required for a more detailed parameterization.

The objective of this paper is to investigate how modellers address the problem of a discharge prediction lacking calibration data in terms of (i) making assumptions about the dominant processes, (ii) choosing the model and its structure, (iii) identifying the model parameters, and (iv) defining the initial and boundary conditions. Furthermore, we discuss how the modeller's attitude changes with enhanced system understanding, and as a side benefit, we analyse the cost–benefit aspect of using models with increased parameter requirements.

The situation the modellers were confronted with was close to what happens in practice, having hardly ever sufficient information at the beginning of the prediction process. Often the model of their choice depends on its availability and flexibility to be adapted to the specifics of the considered case. Chicken Creek is a catchment in its initial phase of development and hence it is a system with a hydrologic behaviour particularly difficult to model. Many catchments are nowadays in a state transformation, but by far most models used in practice do not adequately account for changing conditions (Milly et al., 2008).

Due to the limited number of modellers and the difficulties just mentioned it is obvious that this study cannot be more than a qualitative test of the role of the modeller in the course of refining a first guess step by step by making use of additional information about the system.

2 Participants, their models, and the catchment

2.1 The participants

Ten modellers were invited to predict the discharge from the man-made catchment Chicken Creek (Table 1 and

Table 1. Prior modelling experience (SD = fully or partially) spatially distributed, L = lumped, PB = physically based, and C = conceptual).

Model name/modeller	Modelling experience	Other models	Regions, climates	SD	L	PB	C
Catflow (Maurer, 1997) T. Blume	PhD, postdoc	2	Alps, Andes	X		X	
CMF (Kraft et al., 2008) P. Kraft	teaching	2		X		X	
CoupModel (Jansson and Moon, 2001) D. Gustafsson	12 years	10	Scandinavia	X		X	
Hill-Vi/MIKE SHE (Weiler and McDonnell, 2004; DHI, 2007) S. Stoll	Diploma thesis	3	western Germany	X		X	
NetThales (Chirico et al., 2003) G. B. Chirico	temperate climates, no snow		temperate climates (Australia, southern Europe)		X	X	
SIMULAT (Bormann, 2008; Diekkrüger and Arning, 1995) H. Bormann	10 catchments 1-D to quasi-3-D	4	western Germany, western Africa	X	X	X	X
SWAT (Arnold et al., 1998) J.-F. Exbrayat	Master thesis	6	Scandinavia, Central Europe			X	X
Topmodel (Beven et al., 1995) W. Buytaert	PhD, postdoc	> 3	Ethiopia, Andes	X	X	X	X
WaSiM-ETH (Richards) (Schulla and Jasper, 2007) H. Hölzel	PhD thesis	3	western Germany, Cuba	X		X	
WaSiM-ETH (Topmodel) (Schulla and Jasper, 2007) T. Krauße	PhD thesis	2	Germany	X		X	

Supplement A). Neither of them had an extra budget nor time allocated for this modelling task. The time available for model selection and testing was therefore short. A detailed description of the models can be found in Holländer et al. (2009). In Supplement A, we document the main information about the models and the prediction process. Prior modelling experience varied among the participants. This was certainly relevant for the choice of the model, its implementation, and parameterization (Holländer et al., 2009). All modellers except the CMF user (**C**atchment **M**odelling **F**ramework) had experience with three to five different models (Table 1). CMF, which is a multi-model toolkit, was developed by the user himself, The Hill-Vi user was a member of the developer's group. None of the participants had experience either with artificial catchments or with applying hydrological models in the (semi-)continental climate of Lusatia (Table 1).

2.2 The catchment and monitoring devices

The 6 ha Chicken Creek catchment is located in the lignite mining area Wetzlow-Süd in northeastern Germany (for details see Gerwin et al., 2009b). It is 150 m wide and 400 m long and drains into a 3800 m^2 lake (Fig. 1). The maximum elevation difference is 15 m. The longitudinal slope varies between 1 and 5 % and between 0.5 and 2 % in the transverse direction. The subsoil is a compacted clay layer, which prevents water losses into the underlying geological formations. A V-shaped clay dam separated the catchment from the lake, built to funnel base flow through a narrow outlet into the lake. Coal-free tertiary sands from a nearby mining pit were deposited on top of the clay base and dam. The soil on the western slope had a slightly more sandy texture than the eastern slope. The soil depth varied from 4 m at the upper catchment area decreasing downslope to 2 m. Mean annual rainfall of the period 1961–1990 is 563 mm and mean annual temperature 8.9 °C (Gerwin et al., 2009b). The surface was initially bare (2005). The vegetation cover reached at the end of the observation period (2008) between 0.1 and 34.1 % but its composition underwent quite marked changes. Discharge from the lake was measured with a weir. Discharge from the catchment was calculated based on the lake level changes, and the lake's depth profiles, evaporative losses, and rain inputs. Weather data were recorded with standard equipment mounted on a weather station. Water storage was monitored with 17 groundwater wells (Fig. 1) and four soil pits equipped with TDR probes.

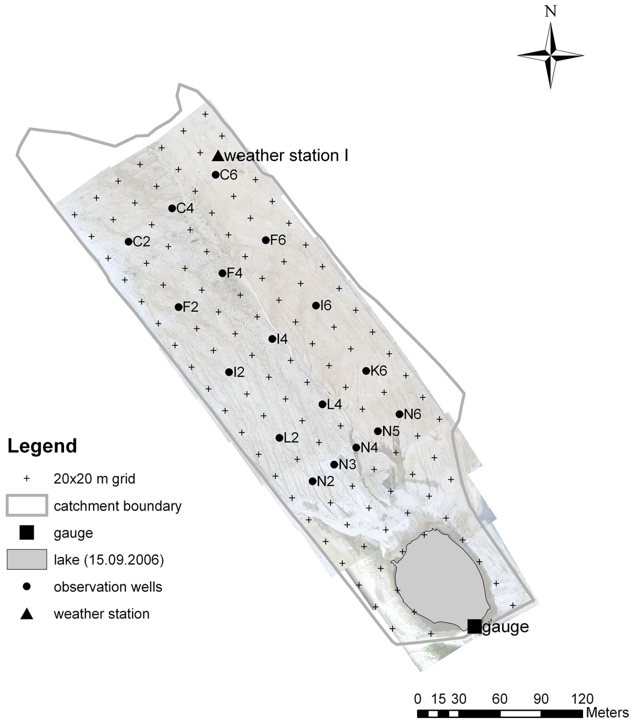

Figure 1. GIS framework of the Chicken Creek catchment.

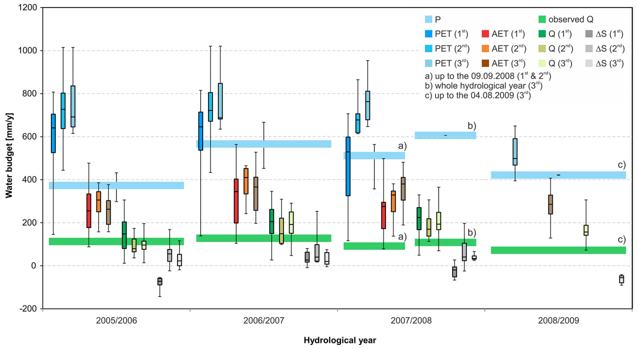

Figure 2. Mean and standard deviation of the water budget components simulated in the first, second and third prediction.

2.3 Provided initial data set

The participating modellers received the following data set: gridded information on elevation of the soil surface and also of the impermeable underlying clay layer (soil base), soil texture and soil depth at all 131 observation squares (Fig. 1), mean annual vegetation cover, hourly climate data (precipitation, temperature, relative humidity, global radiation, wind speed and direction), initial groundwater heads, an aerial photo taken in summer 2007, and a shape file of the clearly visible gully network. Hence, in many real-world applications the database would be even smaller than in the case of this experimental catchment.

2.4 Model selection

For all modellers it appeared to be a straightforward exercise, assuming that an artificial catchment dominated by sandy soils is a rather uniform and homogeneous system, although they had no analogue for expert similarity analyses. Therefore, all models except the Topmodel were parameter-rich physically based "bottom-up" approaches which assumed that the relevant processes can be realistically represented based on the provided data. This large data requirement can generally not be met when models are used for an a priori prediction.

Most of the modellers (CoupModel, CMF, Hill-Vi, SIMULAT, SWAT, Topmodel, WaSiM-ETH (Richards)) chose their model based on their earlier modelling experience whereas the Catflow, and NetThales user decided explicitly only after defining their modelling strategy to cope with the particularities of this application. The Hill-Vi modeller followed the suggestion of his group leader and used the group's model. Most modellers felt that this modelling exercise offered an opportunity to answer questions like (i) how does their previously used or developed model behave in a prediction context (CMF, SIMULAT, Topmodel, both WaSiM-ETH users) and (ii) how does their model perform in a region where it was not used before (CMF, Hill-Vi, and SWAT). The most common motives for selecting a particular model were immediate availability, familiarity with the code, lack of time, and unavailability of funding for extra work.

Not all model structures were perfectly suited to the Chicken Creek case. This applies especially to SWAT, a model developed to address hydrological modelling on a meso-scale or larger. The modeller wanted to test and understand SWAT's behaviour in small catchments: although SWAT is process based, parameters are more of a conceptual nature, e.g. the usage of the curve number approach to simulate infiltration and surface runoff. Therefore, the user minimized the influence of the modeller's decision and used default values with the objective to investigate how SWAT behaves in the case of small catchments.

The SIMULAT, Catflow, Topmodel, and WaSiM-ETH (Richards) user wanted to minimize the influence of the modeller's choice and did not decide a priori on the dominant process(es). This was a lesson the SIMULAT user had learned from previous studies where SIMULAT produced reliable results without prior calibration (Bormann et al., 1999). The Catflow, SWAT, Topmodel, and WaSiM-ETH (Richards) users had similar philosophies: the Topmodel user used the "standard" version of the model as a first approximation, knowing that some assumptions were invalid (e.g. exponential decrease of transmissivity with depth). The WaSiM-ETH (Richards) modeller also relied on the description of the physical processes in his model and thereby minimized the influence of his own decisions. For the Catflow user the model was primarily a platform for hypothesis testing. The underlying hypothesis was that the best possible representation of the catchment structure and physical properties would yield the best (possible) prediction. Although suspecting that soil surface crusts might be a dominant feature they were not included in the setup because the Catflow user understood that the modellers were supposed to use exclusively the provided data.

The majority of the models – except SWAT and Topmodel – have a spatially distributed structure suitable to make use of the accordingly structured catchment characteristics. All modellers who used spatially distributed models used either a rectangular or curvilinear grid with a 20 m or irregular spacing (CMF). All except the NetThales and SIMULAT user modelled at least a saturated and an unsaturated layer.

All modellers explicitly tried to identify the dominant controls of this catchment. However, they neglected infiltration excess knowing that the soils in this catchment are predominantly sandy. Some of the modellers included and others explicitly excluded certain processes, e.g. the existence of a snow pack or soil freezing when snow-free (Holländer et al., 2009), and none of them used the aerial picture which clearly showed the network of gullies.

3 The three prediction stages

3.1 First prediction

The importance of data availability and parameterization procedures for the users of distributed models – given the data as provided in this case – was ranked by the modellers as (1) terrain information such as lower and upper boundary elevation (soil depth), (2) soil texture, (3) vegetation coverage, and (4) for the initial phase soil water storage.

The main modelling challenges of this exercise as stated by the modellers were (a) soil parameterization, (b) evapotranspiration, and (c) the initial conditions.

(a) Soil parameterization: the assumption that an artificial catchment built with sandy soil material is homogeneous justified the use of average soil properties derived from texture data. All modeller's employed pedotransfer functions (PTFs) for assessing soil hydraulic properties: Catflow, Hill-Vi, NetThales, SIMULAT, SWAT, and Topmodel used published PTFs, but national soil databases were used for the CoupModel, CMF, SIMULAT (only bulk densities), and WaSiM-ETH (Richards). Most modellers did not consider the influence of soil freezing on the hydraulic conductivity (K_{sat}) because their model had no routine for snow and frost effects (Catflow, Hill-Vi, Topmodel) or the modeller did not have sufficient experience in this context (CMF, NetThales).

(b) Evapotranspiration: the evapotranspiration parameters were defined primarily based on prior experiences in areas with quite different climatic regimes. None of the modellers had worked with similar climatic conditions before. For instance, the NetThales user tried to match the annual water balance with that of catchments having a similar rainfall regime. The CoupModel user pointed out the importance of soil and snow evaporation based on a surface energy balance.

(c) Initial conditions: the water content of the deposited soil material was initially very low without any groundwater at the soil base. All modellers missed the fact that the catchment was initially far from steady state conditions. In order to determine the initial conditions, the Catflow, CMF, Hill-Vi, and the WaSiM-ETH (Richards) modeller used several warm-up runs to achieve steady state conditions. The other modellers assumed uniformly distributed soil moisture and groundwater levels. SIMULAT started with unsaturated conditions but allowed for partial saturation of the soil columns.

Only the SIMULAT and the SWAT user implemented the dynamics of the lake at the catchment outlet.

3.2 First workshop and field visit

The participants presented their first predictions at the first workshop[1] and visited the catchment (8 December 2008). The SIMULAT and the Topmodel modellers visited the catchment in spring 2009. The parameterization and results of the first-stage prediction are discussed in detail in Holländer et al. (2009).

The modellers realized that they had exploited the initial data set very differently. For instance, all of them neglected the initially low soil water content and the obvious traces of the eroded gullies.

The discussions about the major system controls during the workshop, field visit, and in the course of the manuscript preparation (Holländer et al., 2009) updated the system understanding of all project partners. The importance of the processes in the catchment and the modeller's decisions were ranked in the following order:

1. soil crusts enhancing surface runoff (workshop and field visit);
2. low initial soil water contents and transient conditions during the filling of the soil storage compartments (workshop and field visit);
3. influence of the V-shaped subsurface clay dam (workshop);
4. snowmelt and soil freezing (workshop and manuscript preparation);
5. vegetation development (field visit);
6. analysis of parameterization (manuscript preparation);
7. fast change of topography and soil surface structure (field visit).

[1] The WaSiM-ETH (Topmodel) user attended the workshop as an observer and subsequently joined the project.

3.3 Second prediction

For the second prediction the participants did not receive any additional data.

3.3.1 Redefinition of dominant controls, initial state, and parameters

Surface processes: most important for the second-stage prediction was the on-site inspection of the catchment when the modellers became aware of the deep gullies. The fact that the models did not predict infiltration excess suggested the necessity to implement soil crusts as observed by Fischer et al. (2010). The soil crusts had already developed within the first year after the construction of the catchment.

None of the modellers used the documented gully network. Possible causes of the observed erosion were discussed by the Catflow, SIMULAT, and the WaSiM-ETH (Richards) modellers. Based on the fast filling of the lake in early 2006 (Gerwin et al., 2009b) snowmelt and soil freezing were postulated as additional explanations for the observed erosive surface runoff. The WaSiM-ETH (Topmodel) modeller noticed that the rapid changes of the surface topography apparently depended on soil texture.

Subsurface structures: the V-shaped subterranean clay dam, which was meant to funnel the downslope subsurface flow at the catchment outlet into the lake was handled very differently (Table 2). The CoupModel, Topmodel and WaSiM-ETH (Richards) users did not consider the dam at all and the SIMULAT user assumed a shallow soil layer above the dam. Therefore, the modelled subsurface water storage and flow immediately uphill of the dam differed strongly. Since the predicted water budgets and runoff estimates varied greatly, the WaSiM-ETH (Topmodel) modeller concluded that in this case a physically based model for the unsaturated zone is not needed.

Vegetation: in December 2008 the CoupModel, the Hill-Vi, the NetThales, the Topmodel, and the two WaSiM-ETH users noticed the different vegetation development in the western and eastern half of the catchment. The SIMULAT and Topmodel users who visited the catchment only in spring 2009 noticed the fast development of the vegetation.

All modellers except the NetThales user integrated their findings reported in the joint publication (Holländer et al., 2009) for their second prediction. They analysed the scientific arguments and the parameterization of the other modellers to justify their process and parameter choices (Table 2). The Hill-Vi modeller made use of preliminary information about the catchment and the discharge magnitude as later published by Gerwin et al. (2009a).

3.3.2 Implementation of gained system understanding

The Hill-Vi user implemented most rigorously what he learned in the process. Recognizing that the original model could not adequately describe the actual evapotranspiration (AET), soil crusting, and the effect of the clay dam on the groundwater dynamics without major coding work, he switched to using MIKE SHE[2], redefined the initial conditions, and made small changes to reduce AET and potential evapotranspiration (PET) (Tables 3 and 4). The detailed description of all changes in the hydrological models with regard to model setup and parameterization can be found in Supplement A.

3.4 Third-stage prediction: impact of an extended data set

A second modelling workshop was held in October 2009. Modellers had a better view of the dominant processes controlling the catchment response. Several modellers focused on the following key issues: (i) is it necessary to adapt the model structure to represent the dominant processes? (ii) What type of data is required for improving the model parameterization? (iii) How can the observed heterogeneity be accommodated?

3.4.1 The additional data set

The modelling period for the third-stage prediction was extended and lasted from 29 September 2005 to 4 August 2009. The database contained more and better data because most of the newly installed measuring devices had been installed in 2008. Detailed information on the field equipment and methods can be found in Gerwin et al. (2011) and Mazur et al. (2011). The extended record of weather station I included hourly data of precipitation, air temperature, wind speed and direction, humidity, global radiation, and the vegetation coverage of 2009. The following new data set – except discharge – was offered to the participants:

- K_{sat} measured by slug tests (at 15 grid points);
- K_{sat}, porosity, and bulk density measured in the laboratory on undisturbed samples taken at the four soil pits;
- soil water retention curves from two soil pits (four depths);
- carbon content of all observation points at several depths;
- infiltration rates measured in situ (19 measurements at 10 grid points);
- daily soil moisture measured in the four soil pits at 10, 30, 50, and 80 cm depth;

[2] The Hill-Vi user is referred to as MIKE SHE user hereafter.

Table 2. Process understanding gained and lessons learnt during the three steps of the project work.

Model	First workshop	Field trip	Manuscript preparation
Catflow	– extremely wide range of predictions – initial condition	– dominant processes (role of soil crusts and surface runoff) – uncertainties of field measured data due to monitoring set up	– large variations of soil hydraulic parameters obtained from various pedotransfer functions and their impact on the results
CMF	– initial condition	– gullies considered to be important	– reasons for different results with different models (assumptions and catchment perception)
CoupModel	– initial condition – construction of catchment – different assumptions and model structure	– vegetation characteristics – uncertainties with interpretation of field measurements	– reflections regarding own modelling approach – impact of soil hydraulic properties
Hill-Vi/MIKE SHE	– initial condition – important processes (soil freezing, clay wall)	– dominant process (surface runoff, gullies) – catchment size and characteristics – structures and spatial distribution of the vegetation	– development of the catchment – information about other modeller's decision and results – wrong assumption on PET
NetThales	– initial condition – dominant processes (soil freezing, clay wall)	– dominant processes (soil crusts and surface runoff dominated by infiltration excess) – spatial variability of soil ⇒ impact on soil moisture and vegetation	
SIMULAT	– impact of clay wall on groundwater – dominant process (surface runoff)	– soil crusts – impact of clay wall – vegetation coverage[3]	– justifications of other modellers for their decisions – importance of modeller's decisions during parameterization process
SWAT	– initial condition, lake volume – identical data are interpreted differently by modellers	– dominant processes (soil crusts and surface runoff)	– information about other models and their results
Topmodel	[2]	– clay wall – dominant process (surface runoff by infiltration excess) – vegetation coverage[3]	– information about other models and their results – good water balance ⇒ weak influence of model implementation
WaSiM-ETH (Richards)	– dominant processes (soil crusts and surface runoff) ⇒ K_{sat} most sensitive parameter – initial condition – model weakness (constant layer thickness, no clay wall, no lake)	– catchment size and characteristic and of structures (e.g. gullies) – spatial distribution of the vegetation, soil crust	– qualitative information (e.g. water budget)
WaSiM-ETH (Topmodel)	– physically based model for the unsaturated zone not needed[1]	– dominant process (surface runoff) – rapid catchment, land use and land form changes[1]	

[1] Participated only for the second prediction. [2] First workshop not attended. [3] Field visit in June 2009.

Table 3. Major modifications in the conceptualization of catchment processes and components between the first, second and third prediction.

Model	Initial condition		Soil crust		Clay wall		Other
Prediction stage	Second	Third	Second	Third	Second	Third	
Catflow	X		X		X[2]		
CMF	X	[3]		[3]		[3]	discharge into gullies at unsaturated conditions
CoupModel	X	[3]		[3]		[3]	
Hill-Vi/MIKE SHE[1]	X		X		X		Penman–Monteith, snowmelt
NetThales	X			X	X		soil freezing
SIMULAT		X	X		X	X	plant parameterization
SWAT	X						re-infiltration
Topmodel			X		[4]		
WaSiM-ETH (Richards)	X		X		X		soil thickness, soil cluster, lake
WaSiM-ETH (Topmodel)	[5]	[3]	[5]	[3]	[5]	[3]	

[1] Using another model; [2] Clay wall as no flow boundary instead of clay; [3] No prediction made in that prediction stage; [4] Use of lumped model did not allow implementation; [5] Only part of the second prediction.

Table 4. Major modifications in the parameterization between the first to second and second to third prediction.

Model	Initial condition		K_{sat} of soil crust		Other	
	Second	Third	Second	Third	Second	Third
Catflow			$0.06\,mm\,h^{-1}$	$0.06\,mm\,h^{-1}$	smaller K_{sat}	K_{sat} $110\,mm\,h^{-1}$
CMF	dry state (pF 2.5)	[2]		[2]	K_{sat} $60 \pm 30\,mm\,h^{-1}$, LAI = 1	[2]
CoupModel	dry state	[2]		[2]	smaller K_{sat}	[2]
Hill-Vi/MIKE SHE[3]	dry state		smaller	K_{sat}[1]		K_{sat}[1], larger vegetation
NetThales	dry state			$3\,mm\,h^{-1}$	smaller K_{sat}	K_{sat} $100\,mm\,h^{-1}$
SIMULAT		dry state	$2.1\,mm\,h^{-1}$	$11.6\,mm\,h^{-1}$	LAI > 1 (2008)	σ K_{sat}, min. water level for lower boundary condition
SWAT	dry state					vegetation[1], soil carbon content
Topmodel			$5\,mm\,h^{-1}$		smaller K_{sat}, larger vegetation, clay wall (Te $0.135\,m^2\,h^{-1}$)	K_{sat}
WaSiM-ETH (Richards)	quasi-dry state		$20\,mm\,h^{-1}$		smaller K_{sat}	
WaSiM-ETH (Topmodel)	wet conditions	[2]		[2]		[2]

[1] Properties derived directly from data set (Gerwin et al., 2011; Mazur et al., 2011); [2] No prediction made in that prediction stage.

- 10 min, hourly, daily, and monthly weather data monitored at weather station II: air temperature, wind speed and direction, humidity, net and global radiation, each measured at 2, 5, and 10 m elevation above surface, and precipitation, soil temperature and soil heat flux;
- detailed data about plant species and their distribution at all 131 observation squares;
- DEMs of soil surface elevation determined in November 2005, May 2006, November 2007, and August 2008.

The costs for instrument acquisition, installation, measurement campaigns, and maintenance were estimated according to LAWA (2005) starting at the beginning of the project in 2005. Costs for data inspection and storage are not included. The modellers were asked to select data based on their needs and their ad hoc cost–benefit analysis. The pro forma costs of data acquisition are documented in Table 5. The cost of the modeller's time was not accounted for.

3.4.2 Modellers' rationale of data selection

The CMF and the WaSiM-ETH (Topmodel) modeller left the modelling group after completion of their PhD programme. Although the WaSiM-ETH (Topmodel) user left the group due to time constraints, he requested additional data. Almost all remaining modellers selected K_{sat} derived from slug tests in the field and K_{sat} determined in the laboratory, porosity

Table 5. Virtual costs of data (in EUR) provided in the third prediction.

Measured property	Amount of observation locations	Available since	Costs ($\mathrm{EUR\,yr^{-1}}$)	Total costs (EUR)
Soil hydraulic conductivity (field K_{sat})	15	[4]		590
Soil hydraulic conductivity (laboratory K_{sat})	2	[4]		50
Porosity, bulk density	2	[4]		10
Water retention curves	2	[4]		510
Carbon content	129	[1]		660
Infiltration rates	10	[2,4]		410
Soil moisture (TDR)	4	2008	6200	9300
Weather station II	1	2008	4200	6300
Digital elevation model (DEM)	4			770
Vegetation	120	2006	5210[4]	15 630[4]

[1] Data taken in 2005. [2] Data taken in 2006. [3] Data taken in 2009. [4] Costs are average values.

(except WaSiM-ETH (Richards)), water retention (except SWAT and Topmodel), and soil moisture data (except Catflow, and MIKE SHE). The carbon content of the soil, infiltration rates, more detailed weather data, DEMs and vegetation data were considered of lesser importance and were selected only by few modellers (Table 6).

Generally, two strategies were followed to request additional data: (i) asking for all data which could be used in the specific model to aim for the best possible prediction and (ii) considering the pro forma costs to achieve a good benefit–cost ratio. Surprisingly, none of the modellers opted for the entire data set. Only the NetThales modeller tried to obtain the best fit with the best set of the available data. At this stage he was still aiming at improving his perception of the catchment dynamics and he was particularly interested in understanding the vertical dynamics (infiltration and soil water redistribution). He felt that he could infer a better parameterization of the model although he was already aware of the fact that the model could not mimic the actual catchment behaviour, in particular for what concerns the groundwater dynamics in the early stage of development. Similarly, the WaSiM-ETH (Richards) user still relied on the description of the physical processes, and therefore he requested data which made it possible to improve the model. Similarly, the SIMULAT modeller chose the data based on their usefulness for complementing or revising the model set-up. After a review of the data he used less data than he opted for.

All other modellers selected the data, which were a must or an optional input for their model. For instance, the Topmodel modeller was limited by the conceptual nature of the model, which made it difficult to integrate additional data. Therefore, both modellers chose the same soil and soil moisture data sets. Similarly, the SWAT user did not select the soil moisture data because the HRUs (Hydrologic Response Units) were separated from each other and the model lacked groundwater flow through the soil profile. The MIKE SHE, Catflow, and Topmodel users requested only the information that appeared to be most important for their model.

The soil moisture data could be used as input or as calibration data for defining soil parameters. Soil moisture data were not used by the Catflow and MIKE SHE modellers. Neither the NetThales nor SWAT modeller calibrated their models against soil moisture. The NetThales user argued that the limited point observations cannot be exploited for calibrating his distributed model at the element scale. In fact, there is a mismatch between the scale of computation (plot scale) and the scale of measurements (soil cores). The WaSiM-ETH (Richards) modeller used these data to have a control on the soil moisture dynamic, but not for calibration as done by SIMULAT and Topmodel modeller. The Topmodel user averaged all TDR values measured at 10 cm depth and the groundwater levels of observation well L4 (Holländer et al., 2009) as proxies for the storage deficit. He expected that the latter two observations are inversely related and performed a Monte Carlo sensitivity test based on the correlation coefficient as performance measure. He deduced from this that only the amount of water (expressed as a depth) which the soil can hold within the root zone ($\mathrm{Sr_{max}}$) is a sensitive parameter. Finally, he chose $\mathrm{Sr_{max}} = 0.02\,\mathrm{m}$. Subsequently, the initial root zone storage deficit ($\mathrm{Sr_0}$) and the initial subsurface flow per unit area (q_{s0}) were updated to be compatible with $\mathrm{Sr_{max}}$. The SIMULAT modeller used soil moisture in two steps. First, he evaluated his model with these data and, in a second step, he used them to calibrate the model by adjusting the lower boundary conditions of the soil columns (Bormann, 2011).

3.5 Implementation of data

All modeller used the additional data to revise K_{sat} of the soil and the soil crust. The SIMULAT, MIKE SHE, and the

Table 6. Data chosen by the modeller for the third prediction.

Model	K_{sat} (field)	K_{sat} (lab.)	Porosity	Water retention	Carbon content	Infiltration rates	Soil moisture	Weather station II	DEM	Total vegetation	Costs [EUR]
Catflow	X	X	X	X		X					1570
Hill-Vi/MIKE SHE	X	X	X	X							1160
NetThales	X	X	X	X	X	X	X				11 530
SIMULAT	X	X	X	X[1]	X[1]	X	X[2]			X	27 930
SWAT	X	X	X		X		X	X		X	32 540
Topmodel	X	X	X				X[2]				9950
WaSiM-ETH (Richards)	X	X		X			X	X			16 700

[1] Review of data without using them later in the modelling process. [2] Soil moisture data were used for model evaluation and model calibration.

SWAT increased the impact of the vegetation. The details of data implementation are given in Supplement A.

3.6 Metrics of prediction success

We used the root mean square error (RMSE) and the Nash–Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970) to compare the quality of the discharge predictions. RMSE quantifies the standard deviation of residuals, the differences of predicted from observed discharge, $Q_{\mathrm{p}}\,(t)$ [m^3 d^{-1}] and Q_{o} (t) [m^3 d^{-1}] respectively, with t being the time of prediction and observation [d]:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \times \sum_{i=1}^{n} \left(Q_{\mathrm{o}}(t) - Q_{\mathrm{p}}(t)\right)^2}, \tag{1}$$

where n is the amount of data points [-]. Large discharge values dominate this prediction quality index:

$$\mathrm{NSE} = 1 - \frac{\sum\limits_{i=1}^{n} \left(Q_{\mathrm{o}}(t) - Q_{\mathrm{p}}(t)\right)^2}{\sum\limits_{i=1}^{n} \left(Q_{\mathrm{o}}(t) - \overline{Q_{\mathrm{o}}}\right)^2}, \tag{2}$$

where $\overline{Q_{\mathrm{o}}}$ is the average observed discharge [m^3 d^{-1}]. This index equals 1.0 if the predictions perfectly fit the observations. NSE ≤ 0.0 means that using the average of the observed data is as good or a better predictor as the predicted values. Hence, this index weights the deviation of predicted from observed value relative to the difference of observations from the observed mean.

Looking for a measure which enables us to value the improvement or diminishment of prediction quality, we calculated the relative change of the RMSE Δ_{rel} RMSE [-] for each hydrological year i between the two predictions $j-1$ and j as defined by

$$\Delta_{\mathrm{rel}}\mathrm{RMSE} = \frac{\mathrm{RMSE}_{i,j-1} - \mathrm{RMSE}_{i,j}}{\mathrm{RMSE}_{i,j-1}}. \tag{3}$$

Therefore, the larger Δ_{rel}RMSE the larger is the relative prediction improvement. Similarly, the relative change of the NSE Δ_{rel}NSE [-] is defined by

$$\Delta_{\mathrm{rel}}\mathrm{NSE} = \frac{-\left(\mathrm{NSE}_{i,j-1} - \mathrm{NSE}_{i,j}\right)}{1 - \mathrm{NSE}_{i,j-1}}. \tag{4}$$

In order to look at the potential role of the modellers' experience we use an index for rating their experience taking into account five attributes (Table 1):

1. number of different models they used before,
2. amount of modelling years of each modeller,
3. amount of different regions where the modellers were active,
4. number of years they worked with the model used in this comparison, and finally,
5. closeness of contact of the modeller to the developer team of their model.

All attributes were rated on a scale from 1 to 3 where 1 is little and 3 is top experience. Only the last attribute could be rated with a zero in the case of no or a very minor connection to the model developers. The overall experience is the sum of the five ratings (Table 7).

A final judgement of the consecutive model improvements of the ensemble of all prediction incorporates two criteria:

1. Are the dominant physical processes addressed by the numerical modellers?
2. How does the ensemble of all prediction represent the observed discharge?

Both criteria are ranked from 0 to 4 where 0 represents a poor, 1 represents minor, 2 represents major, and 3 represents a good agreement. The results of all criteria are averaged to the final judgement.

Table 7. Indexed experience of the modeller[1].

Model	Amount of models[2]	Different regions[3]	Modelling years[4]	Years with the used model[5]	Model development[6]	Total (max. 15)
Catflow	1	2	2	3	1	9
CMF	1	1	2	2	3	9
CoupModel	3	1	3	3	2	12
MIKE SHE	2	1	1	1	2	7
NetThales	1	2	3	3	2	11
SIMULAT	2	2	3	3	1	11
SWAT	3	2	2	2	0	9
Topmodel	3	2	2	2	2	11
WaSiM-ETH (Richards)	2	2	1	2	0	7
WaSiM-ETH (Topmodel)	1	1	1	1	0	4

[1] The rating of the prediction "success" reported in this paper does neither qualify the model nor the professionalism of the modeller. This was the unanimous agreement among all participants at the very start of the project. [2] Rating of amount of models: 1: < 3 models, 2: 3–5 models, 3: > 5 models. [3] Rating of different regions: 1: 1 region, 2: 2–5 regions, 3: > 5 regions. [4] Rating of modelling years: 1: during PhD, 2: < 5 years, 3: > 5 years. [5] Rating of years with the used model: 1: < 2 years, 2: 2–3 years, > 3 years. [6] Rating of model development: 0: no or email contact, 1: developers within the range of the modeller, 2: being a developer, 3: being the main developer.

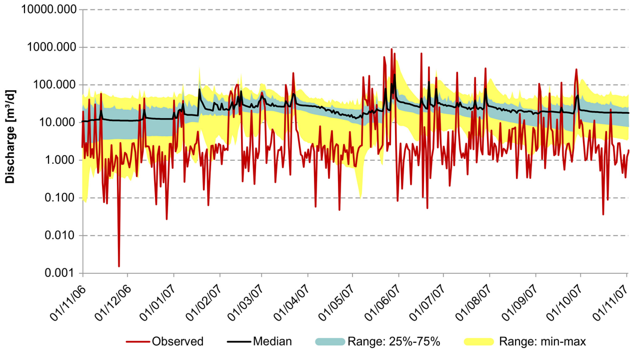

Figure 3. Discharge predicted for the hydrological year 2006/2007 (first prediction).

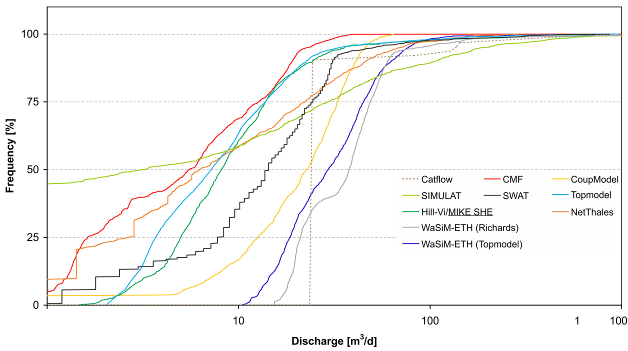

Figure 4. Discharge–frequency relationship of the second prediction.

4 Results

4.1 Main results from the first prediction (minimal data set)

The main result of the first prediction was the huge variability of predicted variables among the models, notably regarding the simulated water budgets (Fig. 2) and daily discharges (Fig. 3). This has been discussed in detail by Holländer et al. (2009). The ratio of predicted to observed annual discharge ranged from 20 to 300 %. This means that the frequency distribution of simulated daily discharge varied drastically as well. It is remarkable that the variability in simulated actual evapotranspiration (AET) was large, ranging from 88 to 579 mm yr^{-1} although most models used the Penman–Monteith approach to predict potential evapotranspiration (PET).

Most models overestimated discharge caused by continuous subsurface flow while the gully network suggested massive surface runoff (Fig. 3). The lateral subsurface flow into the gullies the subsurface slowed the storage change. The change in catchment storage was not well predicted by any of the models. Note that the initial soil water content was not defined by the provided data.

The differences among the model predictions could be mainly attributed to different modeller decisions regarding conceptualization and parameterization of their models. The modellers' decisions during the phase of model implementation and parameterization are strongly influenced by their experience from previous modelling studies. The detailed description of the modelling results and its discussion of this phase can be found in Holländer et al. (2009).

4.2 Main results from the second prediction (after field visit)

4.2.1 Water budget

Relative to the first predictions, the second predictions resulted in the average in larger PET ($+103\,\mathrm{mm\,yr^{-1}}$), AET ($+56\,\mathrm{mm\,yr^{-1}}$), catchment water storage (ΔS; $+49\,\mathrm{mm\,yr^{-1}}$), and lower discharge ($Q$; $-60\,\mathrm{mm\,yr^{-1}}$) (Fig. 2). Note that the errors in mass balance decrease considerably between the two predictions. The variability among the predictions of PET, AET, and Q decreased considerably but the variation of ΔS increased strongly. This can also be seen in the frequency–discharge relationship (Fig. 4). The discharges calculated in the second prediction by all modellers except SIMULAT and WaSiM-ETH (Richards) varied much less than in the first prediction. However, maximum discharge $Q_{\max}$ and the shape of the frequency–discharge relationship still differed widely (e.g. $Q_{\max}$: CoupModel $65\,\mathrm{m^3\,d^{-1}}$ and Topmodel $949\,\mathrm{m^3\,d^{-1}}$).

4.2.2 Evapotranspiration

Most predicted PET ranged from 600 to $800\,\mathrm{mm\,yr^{-1}}$, just CMF and Topmodel predicted more extreme values (e.g. values for the second year) (Table 8). Most models predicted AET in the order of 300, 410 and $330\,\mathrm{mm\,yr^{-1}}$ for the first, second, and third year, respectively. Only SIMULAT (157, 266 and $260\,\mathrm{mm\,yr^{-1}}$) and WaSiM-ETH (Richards) (234, 273 and $285\,\mathrm{mm\,yr^{-1}}$) predicted significantly lower AET.

4.2.3 Discharge

The range of discharge during the three years was 32 to 154 %, 94 to 311 %, and 111 to 271 % of the measured discharge. Most modellers except the MIKE SHE ($\Delta S = -13\,\mathrm{mm\,yr^{-1}}$) and SWAT ($\Delta S - 24\,\mathrm{mm\,yr^{-1}}$) user started their model runs with the initially dry state of the catchment and predicted positive storage changes ΔS throughout the entire simulation period. All other models predicted storage changes in the first year of about 50 to $85\,\mathrm{mm\,yr^{-1}}$. In the second year CoupModel, MIKE SHE, SIMULAT, and SWAT predicted about 20 to $40\,\mathrm{mm\,yr^{-1}}$ of storage change, CMF ($108\,\mathrm{mm\,yr^{-1}}$), NetThales ($88\,\mathrm{mm\,yr^{-1}}$), and WaSiM-ETH (Richards) ($53\,\mathrm{mm\,yr^{-1}}$) predicted larger values . Similar values were obtained for the third year.

The discharge of the second prediction is illustrated in Fig. 5 for the second year. SIMULAT was the only model, which had a zero flow period during 40 % of the simulation period (Fig. 4, Supplement B). CoupModel (2 %) and NetThales (8 %) predicted also periods with zero discharge. Times with base flow below $1\,\mathrm{m^3\,d^{-1}}$ were also found by CMF. All other models consistently predicted base flow larger than $2\,\mathrm{m^3\,d^{-1}}$ (Fig. 4). The largest base flow was predicted by WaSiM-ETH (Richards) and WaSiM-ETH (Topmodel) always at least 15 and $10\,\mathrm{m^3\,d^{-1}}$, respectively. Although SIMULAT predicted zero discharge for many days, it rose rapidly and produced the largest peak discharge of all models ($1433\,\mathrm{m^3\,d^{-1}}$ in the second year). Also CoupModel and SWAT predicted large discharges with Q_{25} equal to 11 and $9\,\mathrm{m^3\,d^{-1}}$, respectively. All other models predicted lower total discharges and hence a low base flow. The CMF, SWAT, and CoupModel predicted peak discharges of less than $100\,\mathrm{m^3\,d^{-1}}$, followed by NetThales ($287\,\mathrm{m^3\,d^{-1}}$), WaSiM-ETH (Topmodel) ($701\,\mathrm{m^3\,d^{-1}}$), MIKE SHE ($742\,\mathrm{m^3\,d^{-1}}$), Topmodel ($958\,\mathrm{m^3\,d^{-1}}$), WaSiM-ETH (Richards) ($1028\,\mathrm{m^3\,d^{-1}}$), and finally SIMULAT (Fig. 4). All models with a peak discharge larger than $100\,\mathrm{m^3\,d^{-1}}$ have in common that there is a strong discharge reduction between Q_{100} and Q_{95} of nearly two orders of magnitude.

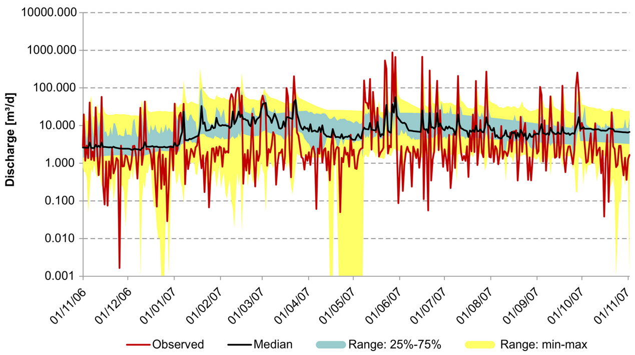

Figure 5. Discharge predicted for the hydrological year 2006/2007 (second prediction).

4.3 Main results from the third prediction (extended data set)

4.3.1 Water budget

Due to a longer simulation period (until 3 August 2009), the predictions for hydrological year 2007/2008 are not directly comparable with those of the earlier predictions. The changes in the water budget until 3 August 2009 from the second predictions to third prediction are directly opposite to the changes from the first predictions to the second predictions: the discharge increased (Q; $+39\,\mathrm{mm\,yr^{-1}}$) but PET ($-15\,\mathrm{mm\,yr^{-1}}$), AET ($-21\,\mathrm{mm\,yr^{-1}}$), and catchment water storage (ΔS; $-18\,\mathrm{mm\,yr^{-1}}$) decreased (Fig. 2). The variability among the predictions of PET and ΔS decreased considerably but the variation of AET and Q increased.

4.3.2 Evapotranspiration

Figure 2 shows that for the first two hydrological years the average PET (third prediction) was slightly reduced when the modellers had access to the larger data sets. Due to a longer simulation period (until 3 August 2009), the predictions for hydrological year 2007/2008 are not directly comparable

Table 8. Water budget components of the second year (second prediction) (NA: not available).

Model	P (mm yr^{-1})	PET (mm yr^{-1})	AET (mm yr^{-1})	Discharge (mm yr^{-1})	Storage (mm yr^{-1}	Balance (mm yr^{-1})
Catflow	[1]	[1]	[1]	[1]	[1]	[1]
CMF	565	433	334	109	108	14
CoupModel	635	NA	417	179	39	−0
MIKE SHE	565	621	452	103	20	−10
NetThales	566	NA	374	104	88	0
SIMULAT	565	688	266	269	17	13
SWAT	565	815	410	148	23	−16
Topmodel	565	1021	465	99	NA	1
WaSiM-ETH (Richards)	565	705	280	327	42	−84
WaSiM-ETH (Topmodel)	716	801	459	221	NA	36

[1] Catflow did not predict a complete water budget due to numerical problems (see Sect. 4.4).

with those of the earlier predictions. Despite the longer simulation period, PET was smaller than in the second prediction. The reduction in PET resulted in a lowered AET. The reduction of PET was nearly equal to that of AET. Since the storage changes were in average also smaller than in the second prediction, the resulting discharge was larger.

The results for the additional simulation period (2008/2009) produced similar results as obtained for the preceding years. AET was on average of the order of 300 mm yr^{-1} and discharge 90 mm yr^{-1}. Only PET and the storage changes showed considerably deviations from previous estimates (PET $\sim$500 mm yr^{-1} and a negative $\Delta S \sim -50$ mm yr^{-1}).

PET simulated for 2006/2007 did not change from the second to the third prediction in the case of NetThales, SIMULAT, and Topmodel because they did not make use of the additional weather data. MIKE SHE reduced PET slightly (-13 mm yr^{-1}) although a denser vegetation was assumed (Table 4). Only the SWAT and WaSiM-ETH (Richards) user requested the data of weather station II. The results of the PET changes were opposite: SWAT predicted about 30 mm yr^{-1} more PET (847 mm yr^{-1}) whereas WaSiM-ETH (Richards) reduced PET by about 30 to 682 mm yr^{-1}.

The changes in AET corresponded to the changes in PET. The models, which did not use new weather data, calculated AET values, which differed by <10 mm yr^{-1} from AET in the second prediction. Only the Topmodel calculated considerably smaller AET (420 mm yr^{-1}, previously 465 mm yr^{-1}). SWAT predicted the largest changes in AET. Although PET only increased by 33 mm yr^{-1}, AET increased by 118 mm yr^{-1} during the second hydrological year. The large AET changes predicted by SWAT are probably due to the parameter changes of the vegetation. AET calculated by MIKE SHE decreased by about 50 mm yr^{-1} despite a larger PET.

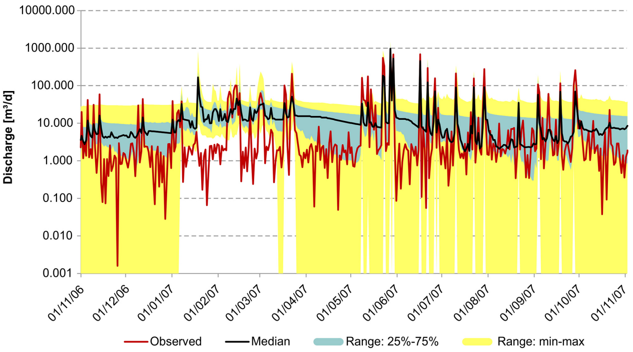

Figure 6. Discharge predicted for the hydrological year 2006/2007 (third prediction).

4.3.3 Discharge

The largest changes in the water budget are those of discharge and storage since all modellers made use of additional soil property data: MIKE SHE (+53 mm yr^{-1}), NetThales (+88 mm yr^{-1}), SIMULAT (+21 mm yr^{-1}), and Topmodel (+48 mm yr^{-1}) predicted larger discharge, whereas SWAT (-101 mm yr^{-1}) and WaSiM-ETH (Richards) (-65 mm yr^{-1}) calculated less discharge , the latter showing an increase in AET. Catflow simulated 255 mm yr^{-1} discharge and therefore the second largest discharge in 2006/2007 (SIMULAT 291 mm yr^{-1}). The changes in discharge were due to changes in K_{sat}, e.g. Catflow used a larger K_{sat} of 110 mm h^{-1} and NetThales increased K_{sat} to 100 mm h^{-1} (Table 4). But K_{sat} of the soil crust was a new input: the NetThales modeller implemented soil freezing and the soil crust into his model using a K_{sat} of 3 mm h^{-1}, thereby reducing the infiltration. Similarly, the MIKE SHE user reduced infiltration by changing K_{sat} of the soil crust.

Figure 6 shows the discharge of the third prediction for the hydrological year 2006/2007. The measured peak discharge on 27 May 2007 was 897 m^3 d^{-1}. The range of predictions

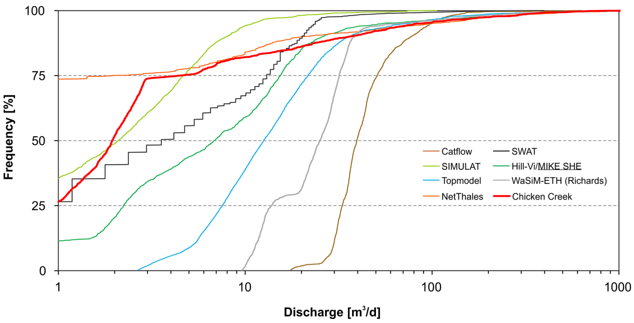

Figure 7. Discharge–frequency relationship of observed and simulated daily discharge (third prediction 2005–2008).

was large, from 106 $m^3 d^{-1}$ (SIMULAT) to 1481 $m^3 d^{-1}$ (NetThales) (Supplement C), but somewhat smaller than during the second prediction ((Supplement B; 24 $m^3 d^{-1}$ (CMF) to 1433 $m^3 d^{-1}$ (SIMULAT)). A similar behaviour is observed for other events during 2006/2007.

The discharge–frequency relationships of MIKE SHE, NetThales, Topmodel, and WaSiM-ETH (Richards) shown in Fig. 7 are quite similar for high discharge (> 40 $m^3 d^{-1}$ at 90 % of all events). Only SIMULAT and SWAT predicted considerably smaller discharges and Catflow larger ones. Low flow was more frequent among models compared to those of the second prediction: NetThales predicted for about 75 % of all events a discharge of less than 1 $m^3 d^{-1}$, whereas Catflow predicts for 5 % of all events a discharge of more than 20 $m^3 d^{-1}$. The discharge characteristic of the Topmodel predictions showed the slightest change.

The detailed results including the measures for RMSE and NSE of all models are given in Supplement A.

4.3.4 Subsurface water storage

The storage changes from the second to third prediction were smaller than from the first to second prediction in case of WaSiM-ETH (Richards) (-171 mm yr^{-1}), NetThales (-49 mm yr^{-1}), and MIKE SHE (-6 mm yr^{-1}) while they were larger in the case of SWAT (+21 mm yr^{-1}) and SIMULAT (+20 mm yr^{-1}). Topmodel and WaSiM-ETH (Topmodel) did not account for storage changes. The large storage changes of WaSiM-ETH (Richards) were mainly caused by correcting the mass balance error as described in the preceding section.

4.4 Prediction success

We used the RMSE and the Nash–Sutcliffe index (NSE) (Nash and Sutcliffe, 1970) to compare the discharge predictions (Figs. 8 and 9). The prediction improvements for the first year were relatively poor throughout all three prediction stages as shown by the RMSE (Fig. 8) and Nash–Sutcliffe index (Fig. 9) mainly because of the large error related to the intensive snowmelt event on 20 and 21 January 2006 (Holländer et al., 2009; Gerwin et al., 2009b). Only NetThales predicted a significant amount of snowmelt in the third prediction, and the predicted discharge was consistently less than measured.

Excluding the snowmelt event results in smaller RMSE (Fig. 8). However, the NSE shows a different picture. Only a few discharge predictions were rated better. Therefore, starting with quasi-dry or dry conditions – as most modellers did in the second prediction – had no positive impact when rated with NSE. RMSE shows similar results. However, the impact on the water budget predictions was positive. Only SIMULAT, SWAT, and Topmodel reduced the RMSE from the first to the second prediction.

For five models RMSE shows an improvement in the second year but a negative impact for four models (Fig. 8). The WaSiM-ETH (Richards) predictions improved strongly from the first to the third prediction but the second prediction of the second year was the worst. The RMSE shows the best results in the third year when the catchment gradually became stabilized. This agrees with the common understanding that catchment models are mainly designed for catchments in steady-state conditions. For the third year, the first predictions were best. However, the average RMSE is still very large (52 $m^3 d^{-1}$). The NSE shows the best results for the first prediction in the third year while the third prediction is the worst. From this we take that the hydrological regime of a particular year may be best predicted by a certain model, which might rank totally differently when predicting the regime of another year. If this is a permissible conclusion – based on the small data set – it would be advantageous to rely on predictions of an ensemble of different models. This agrees with findings of Viney et al. (2009).

5 Discussion

The successive predictions changed mainly due to modified process descriptions (first to second prediction) and due to the availability of additional data (second to third prediction), which affected the parameterization.

5.1 Impact of changing process assumptions and descriptions

Half of the modellers tried to identify dominant processes before their first prediction. In this phase the modeller's experience was crucial (Holländer et al., 2009). Defining the major controls was the major issue during the first workshop. The discussions during the workshop and field visit about the role of soil crusts, initial soil water content, and the role of the V-shaped subsurface dam resulted in more consistent predictions of the water budgets (Fig. 2). For instance, the standard deviation of the simulated AET decreased 33 to 52 %. The annual mean of PET, AET systematically increased, but Q decreased. The exception was water storage.

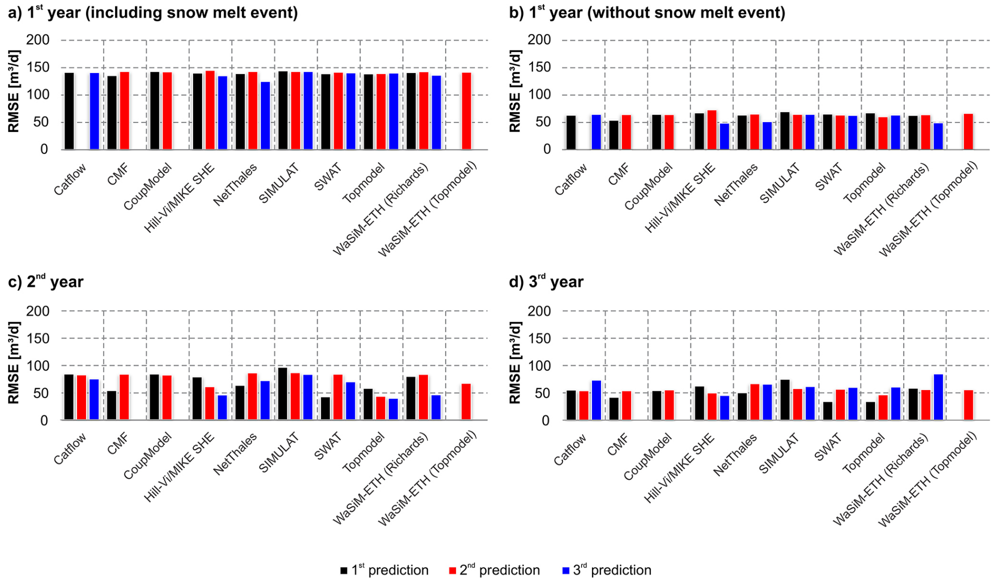

Figure 8. RMSE of each simulated discharge prediction against the observed discharge for **(a)** the first year including the snowmelt event, **(b)** the first year without the snowmelt event, **(c)** the second year, and **(d)** the third year.

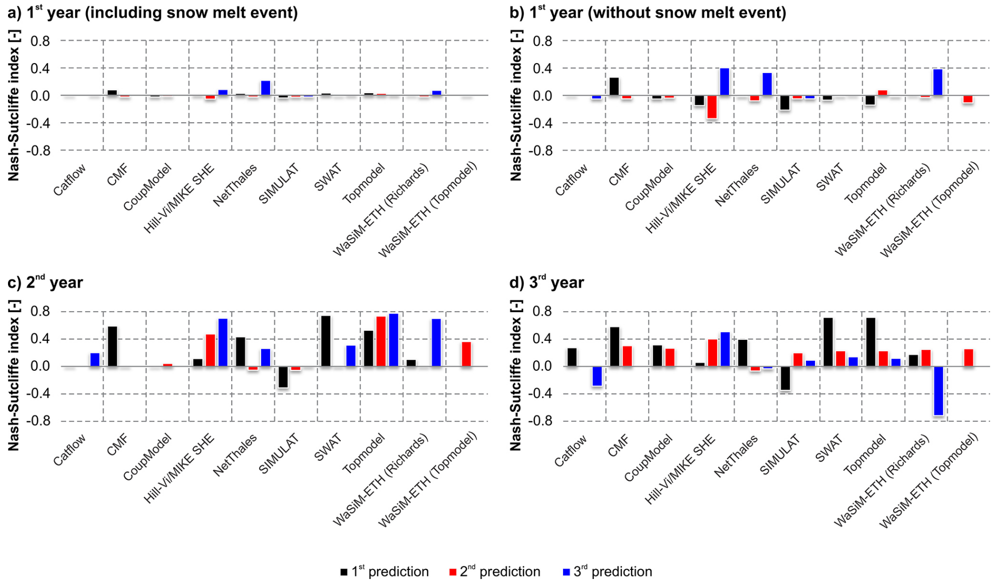

Figure 9. Nash–Sutcliffe index of each simulated discharge prediction against the observed discharge for **(a)** the first year including the snowmelt event, **(b)** the first year without the snowmelt event, **(c)** the second year, and **(d)** the third year.

Several modellers did not account for the initially dry soil conditions and the rising water table because all modellers had previously dealt only with natural "mature" catchments (Table 1).

In the first prediction base flow was overestimated by nearly all modellers although all of them used PTFs based on the same soil texture data to estimate K_{sat}. These estimates varied from 50 (NetThales) to 420 mm h^{-1} (CMF). Most modellers adapted the parameterization of their model for the second prediction. Reducing K_{sat} obviously reduced the base flow and raised the groundwater table (Supplement D). For example, CMF reduced K_{sat} from 420 to 60 mm h^{-1}. It decreased discharge by about one-third. Base flow was also addressed by accounting for the dam, which increased subsurface storage because groundwater could not seep away fast enough.

During the first workshop the soil crust was recognized as the most crucial property in the early phase of the developing catchment. This concept superseded the views prevailing for the first prediction when most modellers tried to reduce

Table 9. Water budget components of the second year (third prediction) (NA: not available).

Model	P (mm yr^{-1})	PET (mm yr^{-1})	AET (mm yr^{-1})	Discharge (mm yr^{-1})	Storage (mm yr^{-1}	Balance (mm yr^{-1})
Catflow	565	NA	197	255	75	38
CMF	[1]	[1]	[1]	[1]	[1]	[1]
CoupModel	[1]	[1]	[1]	[1]	[1]	[1]
MIKE SHE	565	635	403	155	9	−2
NetThales	565	NA	262	192	14	−1
SIMULAT	565	688	268	291	−5	11
SWAT	565	847	528	47	7	−17
Topmodel	565	1021	420	146	NA	−1
WaSiM-ETH (Richards)	565	682	248	244	72	1
WaSiM-ETH (Topmodel)	[1]	[1]	[1]	[1]	[1]	[1]

[1] CMF, CoupModel, and WaSiM-ETH (Topmodel) did not take part of the third prediction.

Table 10. Model-specific runoff components of the second year (third prediction) (all numbers in % of total runoff).

<table>
<tr><th></th><th colspan="3">Second prediction</th><th colspan="3">Third prediction</th></tr>
<tr><th></th><th>Surface runoff</th><th>Interflow</th><th>Base flow</th><th>Surface runoff</th><th>Interflow</th><th>Base flow</th></tr>
<tr><td>Catflow</td><td></td><td></td><td></td><td>11</td><td colspan="2">89</td></tr>
<tr><td>CMF</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>CoupModel</td><td>54</td><td></td><td>46</td><td></td><td></td><td></td></tr>
<tr><td>MIKE SHE</td><td>55</td><td></td><td>45</td><td>87</td><td colspan="2">13</td></tr>
<tr><td>NetThales</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>SIMULAT</td><td>79</td><td>~ 0</td><td>21</td><td>22</td><td>4</td><td>74</td></tr>
<tr><td>SWAT</td><td>39</td><td>61</td><td></td><td>40</td><td>60</td><td></td></tr>
<tr><td>Topmodel</td><td>39</td><td colspan="2">61</td><td>40</td><td colspan="2">60</td></tr>
<tr><td>WaSiM-ETH (Richards)</td><td>5</td><td>56</td><td>39</td><td>36</td><td>38</td><td>26</td></tr>
<tr><td>WaSiM-ETH (Topmodel)</td><td>21</td><td>19</td><td>60</td><td></td><td></td><td></td></tr>
</table>

infiltration by modifying the van Genuchten parameters. Catflow implemented a soil crust with K_{sat} of only 0.06 mm h^{-1} whereas WaSiM-ETH (Richards) used a value of 20 mm h^{-1} (Table 4). The latter value made infiltration and discharge the highest of all models (Table 8) causing a base flow of 51 to 67 % of the discharge. Implementing a soil crust drastically reduced base flow, but the maximum discharges still reached a similar magnitude. Note that precipitation P varied between the models because the modellers interpreted the P record differently so that some modellers used data correction functions (Holländer et al., 2009).

Soil freezing and snowmelt was discussed during the workshop as an important process. CoupModel already contained a snow-and-frost routine for the first prediction and NetThales and MIKE SHE users added it to their model. In the case of NetThales the snowmelt-induced discharge event (January 2006) was quite effectively predicted, but CoupModel predicted very little and MIKE SHE no discharge for this event.

Several modellers (Table 5) accounted for the larger vegetation cover by using a larger leaf area index. This had only a minor impact on PET. Only the MIKE SHE user calculated a significantly different PET, since he switched from using the Turc to the Penman–Monteith model. Similarly, modifications related to the gullies (CMF and SWAT) had a minor influence on the results.

Generally, the impact of the modeller's experience was much less pronounced in the second prediction, very likely because the discussions during the workshop and field visit concerning the dominant system controls harmonized the modellers' views. As a consequence, the water budget components (Fig. 2) showed a smaller variation and a smaller spread in the discharge–frequency relationships (Fig. 4). However, the differences among the model simulations remained substantial. The process assumptions for the third prediction remained largely the same as for the preceding prediction. Only the NetThales user added soil crusting to his model. All parameter changes were based on the additional data.

5.2 Impact of additional data (third prediction only)

The range of AET predictions did not change significantly from the second to the third prediction, but that of PET

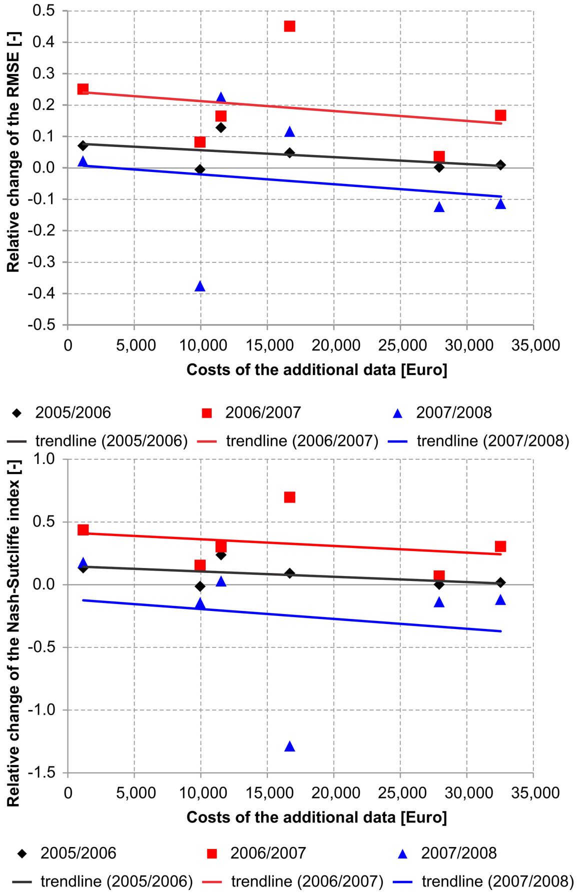

Figure 10. (a) Relationship between the costs of the additional data and the relative change of the RMSE (comparison of second and third prediction). **(b)** Relationship between the costs of the additional data and the relative change of the Nash–Sutcliffe index (comparison of second and third prediction).

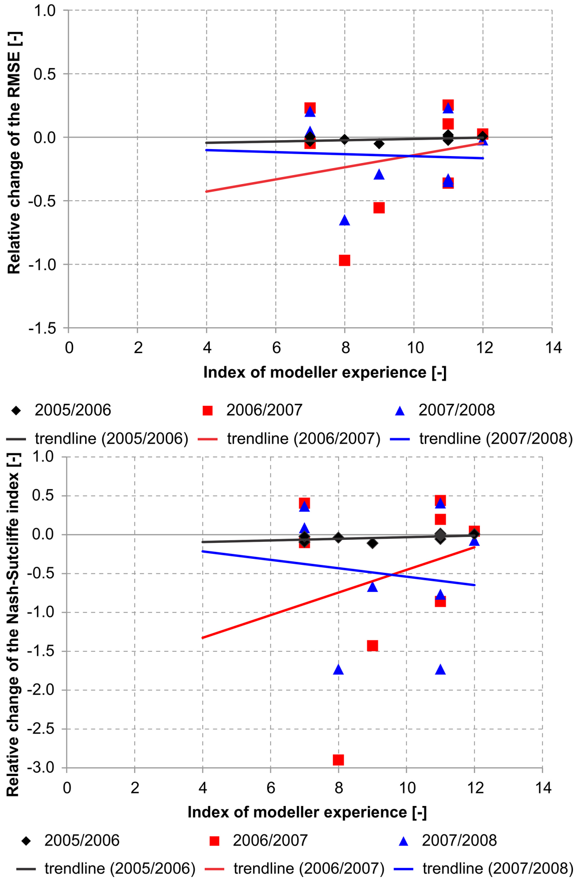

Figure 11. (a) Relationship between the indexed modeller experience and the relative change of the RMSE (comparison of first and second prediction). **(b)** Relationship between the indexed modeller experience and the relative change of the Nash–Sutcliffe index (comparison of first and second prediction).

increased. The reduced variation of discharge and storage (Fig. 2) was clearly a consequence of the updated soil parameterization.

Only the SWAT modeller used additional vegetation data (Table 6). The SWAT user raised AET significantly from 410 to 528 mm yr^{-1}. Since PET only slightly changed, the AET increase was probably not related to the updated vegetation data but rather to the implementation of re-infiltration, which increased soil water storage. WaSiM-ETH (Richards) also used additional weather data, which caused an opposite trend in PET compared to that of SWAT. All other models except MIKE SHE used the same plant parameterization resulting in smaller AET, primarily due to the soil parameterization. MIKE SHE increased the vegetation density, which resulted in a minor AET increase of only 2 % in the second year.

Changes in simulated discharge into the lake are mostly opposite to changes in AET. Most models used the additional data to adjust the soil parameters and the changes in subsurface storage were small (Table 9). Hence, runoff generation was very likely the main driver for changes in discharge.

Depending on model philosophy and parameterization strategy, the various models respond differently in terms of changes in runoff generation mechanisms (surface runoff, interflow, base flow (Table 10 and Supplement A). The runoff generation of Topmodel did not change from second to third prediction, but SIMULAT and SWAT show a remarkable decrease in surface runoff. In case of SIMULAT this is due to the change of soil crust K_{sat} which were justified based on the infiltration tests (Table 4). In case of WaSiM-ETH (Richards) the contribution of base flow increased at the cost of interflow, which may be caused by the choice of soil K_{sat}.

The parameterization, which changed the runoff components, also affected peak flow and the duration of zero-flow in some models. For example, the peak flows (Supplements B and C) simulated by SIMULAT decreased significantly due to changes of the surface crust K_{sat}, which diminished the contribution of surface runoff. In contrast, SWAT generated higher peak flows due to a significant increase in interflow mainly at the cost of base flow. Similar changes due to parameterization of the surface crust are seen for NetThales predictions where the various discharge components cannot be extracted from the hydrographs. CMF, MIKE SHE, NetThales, SIMULAT and SWAT which predict increased fast discharge show zero- or low flow periods while Catflow, Topmodel, and WaSiM-ETH (Richards) still simulate continuous flow. Due to the increased soil crust K_{sat} the zero-flow periods predicted by SIMULAT were shortened.

5.3 Relation between additional data chosen (cost) and the improvement of the model performance (benefit)

Examining this process of iteratively "improving" predictions raises the question of the significance of the modeller's experience versus the modelling strategy per se. None of the modellers had any experience with artificial catchments and the particular climatic conditions of this area. Their experience differed mainly regarding the number of models they were acquainted with and the number of modelled catchments. Having used conceptually different models did not matter in this case because all modellers chose process- and physically based models. An interesting detail is the fact that the most experienced modellers chose the simplest models, either in terms of dimensionality (CoupModel and SIMULAT), or in terms of physical process representation (Topmodel) in order to represent the hydrological behaviour of the Chicken Creek catchment.

Figure 10a and b show the relative improvement from the second to the third prediction related to the costs of additional data (Table 6) using the relative RMSE and NSE, respectively. For the first and second year the prediction quality increased in the average (Δ_{rel}RMSE > 0, Fig. 10a), but making use of more costly additional data did not improve the prediction quality. The results for the third year show a slightly negative trend except in case of MIKE SHE and NetThales which used less costly data. Changes in NSE (Fig. 10b) show for all three years the same trend. Obviously, the trend lines are statistically a weak statement, but on average over all Δ_{rel} values increased data costs did not pay off.

Most of the differences can be attributed to the additional soil data, some of them less costly in total such as K_{sat} (EUR 640), bulk densities (EUR 10), and infiltration rates (EUR 410) and some very expensive (soil moisture: EUR 9300) (Table 6). Both types of data seem to be equally valuable for improving the model parameterization and for an adequate description of the initial conditions.

We compare the indexed modeller's experience against the relative change of the RMSE and NSE between the first and second prediction (Fig. 11a and b). Improvements from the first to the second prediction (without additional data) were larger than those of the following step. The costs for the second prediction – field visit and exchanging ideas during the workshop – were definitely lower than the data costs for the third prediction, but the latter are too arbitrary to be quantified for a similar comparison because they depend on the modeller's travel and work time costs. These results suggest that the sequence of modelling steps could or should follow cost efficiency criteria. The improvement of the predictions from first to the second stage can be explained by a more detailed view on the particular features of the site, by collective learning about the dominant controls as discussed by Holländer et al. (2009), and the modeller's experience in grasping the important features and assimilating convincing arguments brought up by colleagues.

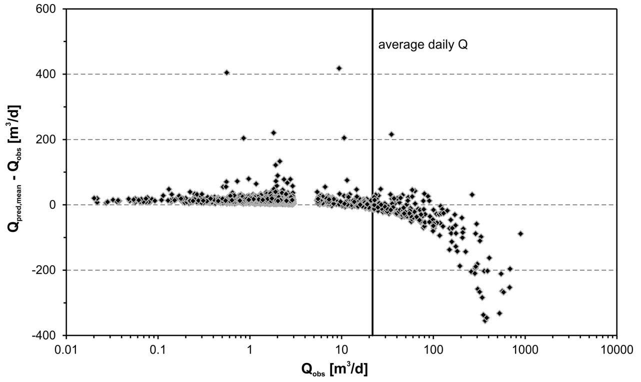

Figure 12. Deviation of the predicted ensemble means $Q_{\text{pred,mean}}$ of daily discharge from the observed discharge Q_{obs}. To plot the log Q_{obs} we added $5 \times 10^{-5}\,\text{m}^3\,\text{d}^{-1}$ to the average daily discharge.

The relative change of the RMSE (Fig. 11a) yields a partly expected picture: almost no impact in the first year since significance of the snowmelt event overwhelms this statistics. Most Δ_{rel}RMSE for the second year were close to zero or slightly positive and those for the third year close to zero or negative mainly due to the predictions by two modellers with medium experience.

It is obvious that single negative or positive Δ_{rel} largely define the slope of trend lines for such a small database. These data intuitively support the hypotheses that the pay-off of more and more costly data was small in this prediction exercise but more experience was essential for the first and second guess. We take this as a justification for proposing future and better funded studies, which do not only compare the suitability of the models, but also the role of the modeller.

In Fig. 12 we compare the deviation of the ensemble mean of predicted daily discharge $Q_{\text{pred,mean}}$ from the actually observed discharge Q_{obs}. In this context we use the predicted mean of the daily means $Q_{\text{pred,mean}}$ ($24.2\,\text{m}^3\,\text{d}^{-1}$) instead of the mean of the daily medians $Q_{\text{pred,median}}$ ($14.3\,\text{m}^3\,\text{d}^{-1}$) because the former corresponds well with the observed mean of daily discharge Q_{obs} ($21.8\,\text{m}^3\,\text{d}^{-1}$). Hence the mean of the ensemble daily means was the better predictor than the mean of the daily medians. This is consistent with the conclusions about the mean being the best predictor drawn by Surowiecki (2004) based on many statistical exercises.

Table 11. Rating of prediction progress compared to measurements in the course of consecutive model improvements (0 = poor to 3 = good).

	Development stage of the catchment		
Prediction stage	first year highly dynamic surface processes, gully formation and snowmelt event	second year increasing role of feedbacks, crust formation, emerging plant cover	third year partly stabilized surface, approaching quasi-steady state
first prediction	0	1	1
second prediction	0	2	3
third prediction	0	2	2

Discharge is systematically over-predicted for discharge rates below the average daily Q_{obs} (Fig. 12). The larger the discharge above this rate, the more pronounced is the under-prediction by the ensemble mean $Q_{\mathrm{pred,mean}}$ down to about two-fifths of the actual Q_{obs}. In the specific case of Chicken Creek – in its initial development phase – the models therefore tended, on average, to over-estimate the non-event discharge and to massively under-predict the event discharge. In Fig. 12 we excluded the extreme influence of the singular snowmelt event (January 2006) and the discharge in the first months after leaving the completed catchment surface to be shaped by nature.

6 Conclusions

Anticipating the hydrological response of a catchment to external forces is the ultimate goal of catchment hydrology. Here we show how expertise and added information affects the quality of such predictions. It is obvious that we back up the hypothesis that ensemble modelling might be a better investment than satisfying the demand for additional, possibly useless, data for making reliable forecasts rather than presenting hard facts because the small number of models and modellers is a statistically weak basis. But based on the results, we advocate a continuation of such prediction exercises. The accuracy of scenarios modelled under given or assumed initial and boundary conditions depends (i) on the availability of pertinent data, (ii) on the suitability of available models – which should include the major system controls – and (iii) on the modeller's expertise to choose and adapt a suitable model. Sufficient modelling experience and a profound system understanding are indispensable. All of the above requirements consume resources. Therefore, it is essential to know more about the gain of prediction quality relative to the needed investments of time and funding.

Each catchment is unique. A predictive model must be tailored to the case-specific features. The man-made Chicken Creek catchment challenged the modellers because of the unusual initial conditions (dry soil material) and the dynamic transition from the state of its construction to that of converging toward a quasi-equilibrium (third year).

Table 11 summarizes the discharge prediction "success" for the first three years. The first prediction was a difficult task because the modellers were confronted with three special features of the newly constructed catchment: (i) the initially dry soil, and (ii) the impact of an unusually intensive snowmelt event, (iii) which enhanced the gully formation on the not yet stabilized bare surface. The gully network was, however, known to the modellers beforehand (aerial picture in the initial data set). The first predictions for the second year were somewhat better. The real progress was made with the second prediction after the field visit and the discussions among the modellers during the first workshop. Adding additional data for the third prediction improved those made for the second but not those for the third year and even decreased the predictive accuracy of several models in the last step. The modellers chose additional data based on two philosophies: using low-investment data to optimize cost efficiency or to perfect parameter guessing by maximizing the database. The former had a better effect for the model performance than using expensive data such as detailed information on vegetation, weather data and newer digital elevation models.

The differences between the first and second predictions were definitely larger compared to those between the second and third predictions. This underpins the value of soft information (field visit, workshop discussions, and experience). However, local measurements such as infiltration tests – provided for the third prediction – certainly contributed to the improved predictions as well. They were apparently better suited to define the soil parameters than those estimated based on PTFs or also laboratory data (Holländer et al., 2009; Bormann et al., 2011).

Most modellers struggled with estimating the initial soil moisture conditions since it is not something that needs to be done in natural "mature" catchments. This corrupted the catchment's storage behaviour during the first year of development. Hard information on soil moisture was therefore essential to define the initial conditions (see also Bormann, 2011; Bormann et al., 2011).

From this modelling exercise we conclude (i) that soft information such as the modeller's system understanding is as important as the model itself, (ii) that the sequence of different modelling steps impacts the relative improvement attributed to the different steps (e.g. field visit, expert discussion, choice of model, selection of available data, parameter estimation protocols), and (iii) that additional process understanding gained during the modelling process can be as efficient as improving data availability for optimizing parameters needed to satisfy model requirements.

7 Implications

Being faced with the request for a real-world prediction we have different options to start with: on-site inspection which is important as we showed in this study, getting a better handle on available data, using local knowledge of residents, or asking differently experienced colleagues to join the team for a first guessing phase, some of them being the real expert and others being recently educated, scientifically up to date and not blocked or blinded by their previous experiences. Such a team might be a better investment (and likely to come at smaller financial cost) than to demand additional, possibly useless data to satisfy parametric needs of the chosen model.

Hence the sequence of modelling steps for making a forecast – a real prediction and not a re-prediction – has to be carefully planned. There is no universal recipe for the "right" strategy. It depends on the case, which might be similar or differ from already encountered cases.

Another important lesson is what became routine in recent years in climate modelling. It is the ensemble of reasonable and well founded predictions, which yields the envelope of the possible outcomes. Such an ensemble is not solely a matter of how good the model is, but how well the steps of making a prediction are being sequenced and being based on solid knowledge.

The Supplement related to this article is available online at doi:10.5194/hess-18-2065-2014-supplement.

Acknowledgements. These investigations were supported by the German Research Foundation (DFG) and the Brandenburg Ministry of Science, Research and Culture (MWFK, Potsdam) in the framework of the Transregional Collaborative Research Centre 38 (SFB/TRR 38). Werner Gerwin gave tremendous help on supporting this work. This study would not have been possible without the careful, persevering, and methodologically creative work of all collaborators at this centre. We thank Vattenfall Europe Mining AG for providing and constructing the artificial catchment. We acknowledge the valuable input of the two anonymous reviewers.

Edited by: R. Merz

References

Arnold, J. G., Srinivasan, R., Muttiah, R. S., and Williams, J. R.: Large Area Hydrologic Modeling and Assessment Part I: Model Development, J. Am. Water Resour. Assoc., 34, 73–89, doi:10.1111/j.1752-1688.1998.tb05961.x, 1998.

Beven, K. J.: Uniqueness of place and process representations in hydrological modelling, Hydrol. Earth Syst. Sci., 4, 203–213, doi:10.5194/hess-4-203-2000, 1999.

Beven, K. J., Lamb, R., Quinn, P., Romanowicz, R., and Freer, J. E.: Topmodel, in: Computer Models of Watershed Hydrology, Colorado, 627–668, 1995.

Blöschl, G.: Rainfall-Runoff Modeling of Ungauged Catchments, in: Encyclopedia of Hydrological Sciences, edited by: Anderson, M. G., John Wiley & Sons, Ltd, Chichester, 2061–2080, 2006.

Bormann, H.: Sensitivity of a soil-vegetation-atmosphere-transfer scheme to input data resolution and data classification, J. Hydrol., 351, 154–169, doi:10.1016/j.jhydrol.2007.12.011, 2008.

Bormann, H.: Treating an artificial catchment as ungauged: Increasing the plausibility of an uncalibrated, process-based SVAT scheme by using additional soft and hard data, Phys. Chem. Earth, Parts A/B/C, 36, 615–629, doi:10.1016/j.pce.2011.04.006, 2011.

Bormann, H., Diekkrüger, B., and Renschler, C.: Regionalisation concept for hydrological modelling on different scales using a physically based model: Results and evaluation, Phys. Chem. Earth, Part B, 24, 799–804, doi:10.1016/s1464-1909(99)00083-0, 1999.

Bormann, H., Holländer, H. M., Blume, T., Buytaert, W., Chirico, G. B., Exbrayat, J.-F., Gustafsson, D., Hölzel, H., Kraft, P., Krauße, T., Nazemi, A., Stamm, C., Stoll, S., Blöschl, G., and Flühler, H.: Comparative discharge prediction from a small artificial catchment without model calibration: Representation of initial hydrological catchment development, Die Bodenkultur – J. Land Manage., Food and Environment, 62, 23–29, 2011.

Breuer, L., Huisman, J. A., Willems, P., Bormann, H., Bronstert, A., Croke, B. F. W., Frede, H.-G., Gräff, T., Hubrechts, L., Jakeman, A. J., Kite, G., Lanini, J., Leavesley, G., Lettenmaier, D. P., Lindström, G., Seibert, J., Sivapalan, M., and Viney, N. R.: Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM). I: Model intercomparison with current land use, Adv. Water Resour., 32, 129–146, doi:10.1016/j.advwatres.2008.10.003, 2009.

Chirico, G. B., Grayson, R. B., and Western, A. W.: On the computation of the quasi-dynamic wetness index with multiple-flow-direction algorithms, Water Resour. Res., 39, 1115, doi:10.1029/2002wr001754, 2003.

DHI: MIKE SHE user manual, 386 pp., Vol. 2, Reference Guide, 2007.

Diekkrüger, B. and Arning, M.: Simulation of water fluxes using different methods for estimating soil parameters, Ecol. Modell., 81, 83–95, doi:10.1016/0304-3800(94)00162-B, 1995.

Fischer, T., Veste, M., Schaaf, W., Dümig, A., Kögel-Knabner, I., Wiehe, W., Bens, O., and Hüttl, R. F.: Initial pedogenesis in a topsoil crust 3 years after construction of an artificial catchment in Brandenburg, NE Germany, Biogeochemistry, 101, 165–176, doi:10.1007/s10533-010-9464-z, 2010.

Gerwin, W., Raab, T., Biemelt, D., Bens, O., and Hüttl, R. F.: The artificial water catchment "Chicken Creek" as an observatory for critical zone processes and structures, Hydrol. Earth Syst. Sci. Discuss., 6, 1769–1795, doi:10.5194/hessd-6-1769-2009, 2009a.

Gerwin, W., Schaaf, W., Biemelt, D., Fischer, A., Winter, S., and Hüttl, R. F.: The artificial catchment "Chicken Creek" (Lusatia, Germany)–A landscape laboratory for interdisciplinary studies of initial ecosystem development, Ecol. Eng., 35, 1786–1796, doi:10.1016/j.ecoleng.2009.09.003, 2009b.

Gerwin, W., Schaaf, W., Biemelt, D., Winter, S., Fischer, A., Veste, M., and Hüttl, R. F.: Overview and first results of ecological monitoring at the artificial watershed Chicken Creek (Germany), Phys. Chem. Earth, Parts A/B/C, 36, 61–73, doi:10.1016/j.pce.2010.11.003, 2011.

Goodrich, D. C.: Geometric simplification of a distributed rainfall-runoff model over a range of basin scales, Technical Reports NO. HWR 91-010, Hydrology Department, University of Arizona, 361 pp., 1990.

Holländer, H. M., Blume, T., Bormann, H., Buytaert, W., Chirico, G. B., Exbrayat, J.-F., Gustafsson, D., Hölzel, H., Kraft, P., Stamm, C., Stoll, S., Blöschl, G., and Flühler, H.: Comparative predictions of discharge from an artificial catchment (Chicken Creek) using sparse data, Hydrol. Earth Syst. Sci., 13, 2069–2094, doi:10.5194/hess-13-2069-2009, 2009.

Jansson, P.-E. and Moon, D. S.: A coupled model of water, heat and mass transfer using object orientation to improve flexibility and functionality, Environ. Modell. Softw., 16, 37–46, doi:10.1016/s1364-8152(00)00062-1, 2001.

Kraft, P., Vaché, K. B., Breuer, L., and Frede, H.-G.: A solute and water flux library for catchment models, Proceedings of the iEMSs Fourth Biennial Meeting: International Congress on Environmental Modelling and Software Barcelona, 2008.

LAWA: Leitlinien zur Durchführung dynamischer Kostenvergleichsrechnungen 7th Edn., edited by: Deutsche Vereinigung für Wasserwirtschaft, A. u. A. e. V., Berlin, 186 pp., 2005.

Maurer, T.: Physikalisch begründetete, zeitkontinuierliche Modellierung des Wassertransports in kleinen ländlichenen Einzugsgebieten, Inst. für Hydrologie und Wasserwirtschaft (IHW), Universität Karlsruhe, Karlsruhe, 1997.

Mazur, K., Schoenheinz, D., Biemelt, D., Schaaf, W., and Grünewald, U.: Observation of hydrological processes and structures in the artificial Chicken Creek catchment, Phys. Chem. Earth, Parts A/B/C, 36, 74–86, doi:10.1016/j.pce.2010.10.001, 2011.

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., and Stouffer, R. J.: Stationarity Is Dead: Whither Water Management?, Science, 319, 573–574, doi:10.1126/science.1151915, 2008.

Naef, F.: Can we model the rainfall-runoff process today? / Peut-on actuellement mettre en modèle le processus pluie-écoulement?, Hydrol. Sci. J., 26, 281–289, doi:10.1080/02626668109490887, 1981.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I – A discussion of principles, J. Hydrol., 10, 282–290, doi:10.1016/0022-1694(70)90255-6, 1970.

Oudin, L., Andréassian, V., Perrin, C., Michel, C., and Le Moine, N.: Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments, Water Resour. Res., 44, W03413, doi:10.1029/2007wr006240, 2008.

Parajka, J., Merz, R., and Blöschl, G.: A comparison of regionalisation methods for catchment model parameters, Hydrol. Earth Syst. Sci., 9, 157–171, doi:10.5194/hess-9-157-2005, 2005.

Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., Seo, D.-J., and DMIP Participants: Overall distributed model intercomparison project results, J. Hydrol., 298, 27–60, doi:10.1016/j.jhydrol.2004.03.031, 2004.

Schulla, J. and Jasper, K.: Model Description WaSiM, ETH Zurich, Zurich, 181 pp., 2007.

Seibert, J. and McDonnell, J. J.: On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration, Water Resour. Res., 38, 1241, doi:10.1029/2001wr000978, 2002.

Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J. J., Mendiondo, E. M., O'Connell, P. E., Oki, T., Pomeroy, J. W., Schertzer, D., Uhlenbrook, S., and Zehe, E.: IAHS decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences, Hydrol. Sci. J., 48, 857–880, doi:10.1623/hysj.48.6.857.51421, 2003.

Surowiecki, J.: The wisdom of crowds, Anchor Books, New York, 2004.

Viney, N. R., Bormann, H., Breuer, L., Bronstert, A., Croke, B. F. W., Frede, H., Gräff, T., Hubrechts, L., Huisman, J. A., Jakeman, A. J., Kite, G. W., Lanini, J., Leavesley, G., Lettenmaier, D. P., Lindström, G., Seibert, J., Sivapalan, M., and Willems, P.: Assessing the impact of land use change on hydrology by ensemble modelling (LUCHEM) II: Ensemble combinations and predictions, Adv. Water Resour., 32, 147–158, doi:10.1016/j.advwatres.2008.05.006, 2009.

Weiler, M. and McDonnell, J. J.: Virtual experiments: a new approach for improving process conceptualization in hillslope hydrology, J. Hydrol., 285, 3–18, doi:10.1016/s0022-1694(03)00271-3, 2004.

Wösten, J. H. M., Pachepsky, Y. A., and Rawls, W. J.: Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics, J. Hydrol., 251, 123–150, doi:10.1016/s0022-1694(01)00464-4, 2001.