Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Hydrology 279 (2003) 244–261

Journal of Hydrology

www.elsevier.com/locate/jhydrol

Hydrologic process simulation of a semiarid, endoreic catchment in Sahelian West Niger. 2. Model calibration and uncertainty characterization

Bernard Cappelaere[a,*], Baxter E. Vieux[b], Christophe Peugeot[a], Ana Maia[a], Luc Séguis[a]

[a]*UMR HydroSciences, IRD-BP 64501, 34394 Montpellier Cx 5, France*
[b]*University of Oklahoma, 202 West Boyd Street-CEC334, Norman, OK 73019, USA*

Received 17 September 2002; accepted 6 May 2003

Abstract

The Wankama endoreic system in the region of Niamey (Niger), monitored over the period 1992–2000, is studied with *r.water.fea*, a physically-based, spatially-distributed rainfall–runoff model. Catchment characteristics and data, together with model principles and construction, are described in Peugeot et al. [J. Hydrol., 2003], who used the uncalibrated model as one of several investigation tools for the screening of rainfall–runoff observations. This second paper focuses on model calibration and verification, namely the methods and criteria used to that end followed by the results thereby obtained. Based on a diagnostic function that combines errors in runoff volumes and in peak discharges, calibration is performed by exploring a 3D parameter space. A resampling-based cross-validation technique is used to investigate calibration stability with respect to data sample fluctuations, and to assess the predictive capabilities of the calibrated model. The issues of parameter uncertainty, sample representativeness, and presence of influential observations, are discussed. An empirical, non-parametric method is devised to characterize parameter uncertainty and to assign intervals to volume predictions. Model verification is performed against the data from the last two seasons. Internal catchment behavior, as produced by the model, appears qualitatively consistent with field information, including a weak upper-area contribution to catchment outflow due to large runoff abstraction by the conveying hydrographic network.

Keywords: Distributed hydrologic modeling; Calibration; Uncertainty; Channel infiltration; Hortonian runoff; Sahel; Africa

1. Introduction

The Sahel has been subject to sharp changes in climatic and environmental conditions since the late sixties, which strongly impact the water cycle. In this context, the goal of the study presented here and in the companion paper (Peugeot et al., 2003), hereafter referred to as '*C.Pap.*' (cf. Nomenclature), is to assess the impact of these changes on water resources in the Sahelian Niamey region, southwest Niger. Surface water consists of a large number of small pools, outlets of endoreic watersheds, which represent the main source of recharge of the underlying unconfined aquifer. Our studies focus on a 1500 km^2 target-area located 70 km east of Niamey, within the Hapex–Sahel Experiment area (Goutorbe et al., 1994). The Wankama endoreic system, composed of an

[*] Corresponding author. Fax: +33-467-144-774.
E-mail address: bernard.cappelaere@mpl.ird.fr (B. Cappelaere).

0022-1694/03/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0022-1694(03)00182-3

Nomenclature

C.Pap.	companion paper (Peugeot et al., 2003)
DEM	digital elevation model
EXO$^{(*)}$	event with proven exogenous inflow to Wankama pool (see *C.Pap.*)
GIN$^{(*)}$	event with proven general inundation of pool-kori system (*C.Pap.*)
GIS	geographical information system
IRD	Institut de Recherche pour le Développement (formerly ORSTOM), France
LOW$^{(*)}$	event with very low observed-to-simulated runoff ratio (*C.Pap.*)
pEXO$^{(*)}$	event with probable exogenous inflow to Wankama pool, rejected from reference data set after statistical and model-aided analyses (*C.Pap.*)
UMR	Unité mixte de recherche (CNRS-IRD-Universités de Montpellier 1&2, France)

(*) designates categories in the event classification of *C.Pap.*

elongated 1.9 km^2 catchment (roughly 2.7 km long by 0.7 km wide) that drains to a 5 ha pool and is considered representative of local catchments, is used as a reference for subsequent extension to the full study area. These two companion papers report on the modeling of surface runoff and pool recharge in this pilot watershed, using the physically-based, distributed model *r.water.fea* (Vieux and Gaur, 1994; Vieux, 2001). Physical characteristics of the watershed and analysis of rainfall–runoff data collected over the 1992-to-2000 rain seasons have been described in *C.Pap.*, together with model principles and construction. Using the uncalibrated model as an aid for event data analysis, a fraction of observed events in the 1992–1998 period was screened out from the reference set because of doubtful runoff volume values due to suspected exogenous inflow. The selection method was validated with the 1999–2000 data, for which no such uncertainty exists. A reference set of 97 rainfall–runoff events (73 and 24 for the two consecutive periods, respectively) was thus retained. The purpose of this paper is to present the tuning, verification and uncertainty analysis performed on the model based on the reference data set. Here again the 1999–2000 data is used a posteriori, only to test the model once calibrated with the 1992–1998 data.

Here is a brief summary of information about the catchment's hydrology, data, and model. Precipitation is limited to a short wet season (May–June to September–October) and for the most part is associated with west-bound mesoscale convective systems and squall-lines, causing short, often intense and very intermittent rainfall. Data are continuously available as 5 min intensities at the Mare raingauge near the pool. Hortonian overland runoff is largely linked with the presence/absence of surface crusts on essentially sandy soils, and therefore with the spatial distribution of cultivated millet fields and of natural savanna vegetation clearing. Average slope is around 2%. The hydrographic network is basically made up of a 2.5 km long, sandy and hence infiltrating main ravine, comprising two reaches that connect at mid-slope through a sand-clogged spreading zone 0.4 km long by 0.2 km wide. The lower reach feeds into the Wankama pool, which may occasionally interact with neighboring pools. Because of this and of catchment runoff flow values being derived from observed stage fluctuations of the Wankama pool, not all runoff events can be used for our catchment modeling study (*C.Pap.*). This results in an under-representation of large storms in the reference data sample. Large-event under-representation is a potential problem since this category may account for a substantial fraction of water resources in the area, but on the other hand Séguis et al. (2002) have shown that parameter calibration is more responsive to short and/or low intensity rainfall events. Event volumes (hereafter expressed as depths in mm after division by the catchment surface area) are known more precisely than actual discharge (m^3/s), and are also of much greater concern in the context of this resource-oriented study. The *r.water.fea* physically-based, spatially-distributed, Hortonian runoff event model couples Green-Ampt infiltration and kinematic wave routing equations to represent runoff and run-on distributions in time and space over complex terrain, land use, soil, and conveying-network conditions, within the *Grass* raster-GIS software (USACE, 1993). In conjunction with uniform rain intensity time-series, input for the Wankama catchment consists of 20 m resolution maps for DEM (from field survey) and hydraulic parameters (from SPOT imagery and field observations), including six channel reaches and a spreading zone. Spatially distributed hydraulic conductivity and Manning roughness are used as prior values before model calibration (see Section 2). For initial humidity at the beginning of a rain event, a uniform soil saturation ratio is computed, based on the API (Antecedent Precipitation Index) formulation of Kohler and Linsey (1951) and calibrated on soil moisture observations (*C.Pap.*).

The uncalibrated Wankama model was used in *C.Pap.*, together with more purely data-driven, statistical rainfall–runoff analyses, to perform an event classification from the recorded 1992–1998 raw sample, leading to selection of a subset referred to as the 'reference event sample'. As fortunate as it is fortuitous, it so happens that the uncalibrated model performs rather well, in terms of agreement between observed and simulated runoff volumes, against the reference event sample. This does not however obliterate the necessity for actually calibrating the model against sample data, in order to gain increased predictive accuracy but also to clarify parameter and prediction uncertainty. Model calibration includes tuning of the parameter set in order to obtain optimal model fit to observations, as well as characterization of parameter uncertainty and definition of an evaluation procedure for predictive uncertainty. Model verification is performed here mainly with the tuned ('optimal') parameter set, but the predictive-uncertainty estimation is also partially tested. Section 2 presents the methods used for model calibration and verification, while results are detailed in Section 3.

2. Calibration—verification method

2.1. Parameter space

Many physical characteristics included in the model (*C.Pap.*), particularly those related to geometry and topography, were defined as being invariant throughout the modeling procedure. Only the hydraulic conductivity and roughness parameters are subject to tuning, starting with prior reference values that were assigned on the basis of field and remote sensing information. The purpose of this subsection is to specify the parameterization framework used and the control-parameter space within which model calibration is performed. Some results from preliminary sensitivity analyses within this control-parameter space are also presented.

Given that for any physical parameter (e.g., hydraulic conductivity K_s) there are as many potential model parameters as there are grid cells in the discrete model, there is a need to reduce the parameter space dimension, which must be proportionate to the size and information content of the calibration data sample. This reduction can be done by using the distributed prior reference values as constraints on the relative spatial pattern for a given parameter physical type, leading to a single scaling factor per parameter type. Three such dimensionless scalar control parameters, K, C, and M, are actually defined: M is a uniform multiplier applied to the prior raster map of spatially distributed Manning's roughness values, while K and C are applied to hillslope cells and to channel cells of the prior K_s map, respectively. Hence, the model parameterization consisting of these three dimensionless scalar parameters K, C, and M, preserves the spatial structure for each of the associated infiltration or roughness properties.

For the largest event in the reference sample (July 18, 1992, with rain depth 57 mm), Fig. 1 shows model sensitivity to the K, C, and M control parameters around the prior point $(1, 1, 1)$, with only one of the three being varied at a time. Fig. 1(a) shows the sensitivity of total volume for each parameter. While it might have been thought that runoff volumes are essentially controlled by the K and C infiltration parameters and that M would primarily control the timings of flow routing, Fig. 1(a) shows that M also has a strong impact on volumes. This is because, within the model as in reality, runoff production and routing interact: if surface flow is slowed down, due to hydraulically rougher surfaces, then there is more opportunity for infiltration. Sensitivity to C is comparatively lower: this probably results from the catchment channel only partially controlling

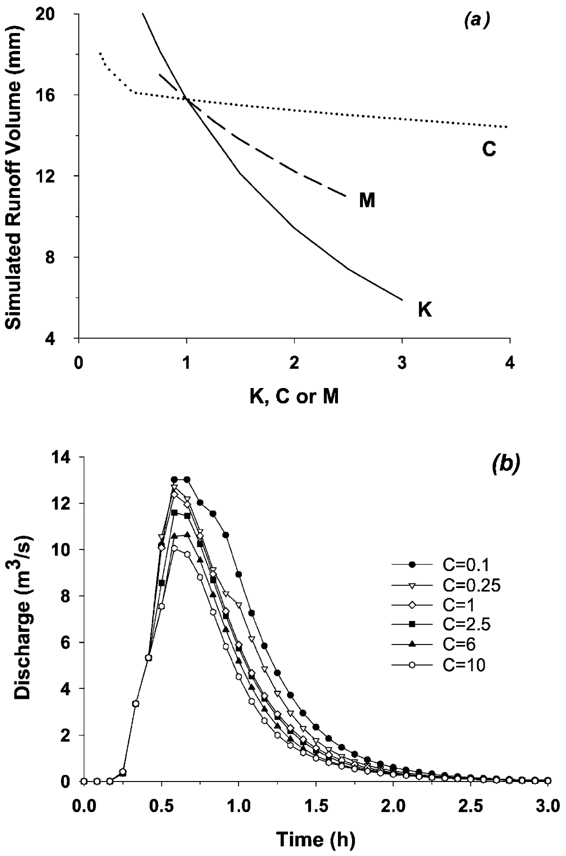

Fig. 1. Sensitivity to parameter values (July 18, 1992): (a) sensitivity of volume to *K* (solid line), *C* (dotted line) and *M* (dashed line); (b) hydrograph sensitivity to channel conductivity parameter *C*.

the catchment's surface-area (~55%), and from its rather high runoff abstraction efficiency, particularly in the mid-slope spreading zone, given its prior conductivity value (450 mm/h). The *C* impact on runoff hydrograph is shown in Fig. 1(b). The rising limb is much less affected than the peak and recession, a finding consistent with those reported by Woolhiser et al. (1996) for analytic kinematic-wave solutions with channel losses.

2.2. Data sample management

A satisfying calibrated model should show sufficient stability with respect to changes in the composition of the calibration data sample. When the available sample size is large enough, a common practice to check this stability is to divide the sample ('split-sampling') into two independent subsamples that both represent the overall population equally well, and to use one of them to perform calibration and the other one to test the calibrated model (see for instance: Kuczera and Parent, 1998; Feyen et al., 2000). Owing to an insufficient number of large storms, our 97-event reference data set does not allow two such equally representative subsamples to be obtained. The approach followed here is to use the entire 1992–1998 period of the reference set (73 events) to perform model calibration, and to test the calibration both by applying a resampling technique on the 1992–1998 subsample (hereafter referred to as cross-validation) and by verifying the calibrated model's performance against the 1999–2000 data (24 events). The latter is too short to fully represent the storm event population, but is quite informative in the sense that it did not have to go through the model-based event screening process devised for the 1992–1998 data (*C.Pap.*). In practice, these 'specialized' respective uses of the two successive data periods were performed independently at quite separate times, as each set became available.

2.3. Diagnostic function

Calibration is achieved through the minimization of a diagnostic function, hereafter denoted *F*, which quantifies model agreement with observations. This measure is also used for model verification, and serves as the basis for parameter uncertainty analysis. Various function formulations are possible (e.g. Sorooshian et al., 1983; Gan et al., 1997; Gupta et al., 1998; Feyen et al., 2000), depending on modeling purposes. For instance, if flood forecasting was the objective, then peak discharge and time-to-peak would be the most important criteria. Our perspective being water resources, runoff volume distribution over the storm event population is of primary concern. However, to maximize model reliability and robustness, it is desirable that the comparison between simulated and observed runoff also accounts for hydrograph shape. Due to the measurement method (*C.Pap.*), per-event runoff volumes are known with significantly better precision than instantaneous discharge rates. Hence, it is reasonable to base model performance assessment primarily on hydrograph volume and secondarily on its shape. In conjunction with the integrated volume, the latter is represented by the peak discharge (the choice of this variable as a shape descriptor is further discussed in Section 3.2) and is granted a lesser weight in the analysis. Because the distribution of rain depths in the reference data sample shows some distortion vis-à-vis the full set of observed rainstorms (*C.Pap.*), weighting coefficients are applied to each event in the diagnostic function expression, based on rain depth partitioning into 7 classes: for any event i in class j, the weight α_i is taken as the ratio of the number of events in class j for the two rain samples (full-sample over reference-sample) in order to reflect the full-sample class size through the reference-sample individuals. Resulting weight values are maximum for the upper storm-magnitude class. This event-number imbalance correction is felt all the more necessary as the contribution to seasonal pool recharge from every single large rain event is so much greater than from any small one.

For either one of the two variables (runoff volume V or peak discharge Q), the measure of model fit to a set of any n observations is built as the mean weighted sum of squared errors over these n events, normalized by the observed weighted variance over a fixed, reference sample of n_0 events, i.e.:

for runoff volume : F_V^2

$$= \frac{\sum_{i=1}^{n}\{\alpha_i[V_{\text{obs}}(i) - V_{\text{sim}}(i)]^2\}/\sum_{i=1}^{n}\alpha_i}{\sum_{i=1}^{n_0}\{\alpha_i[V_{\text{obs}}(i) - \overline{V_{\text{obs}}}]^2\}/\sum_{i=1}^{n_0}\alpha_i}, \tag{1}$$

for peak discharge : F_Q^2

$$= \frac{\sum_{i=1}^{n}\{\alpha_i[Q_{\text{obs}}(i) - Q_{\text{sim}}(i)]^2\}/\sum_{i=1}^{n}\alpha_i}{\sum_{i=1}^{n_0}\{\alpha_i[Q_{\text{obs}}(i) - \overline{Q_{\text{obs}}}]^2\}/\sum_{i=1}^{n_0}\alpha_i}. \tag{2}$$

Denominators are only used to make errors non-dimensional. It is the 1992–1998 reference sample of size $n_0 = 73$ which is used to compute these referential observed variances, irrespective of the n events on which model performance is measured. Normalization by constant variances eases inter-comparison of F_V- or F_Q-values, i.e. of mean quadratic errors, between dissimilar event samples, namely: 1992–1998 subsamples of unequal sizes (cross-validation in Sections 2.5 and 3.2) or distinct observation periods (Sections 2.7 and 3.5). Nil values of F_V and F_Q indicate perfect fit. Note that $1 - F_V^2$ (or $1 - F_Q^2$) would amount to the so-called efficiency criterion of Nash and Sutcliffe (1970) when the n and the n_0 events coincide. Combination of the two criteria into a global diagnostic function F is achieved by a weighted quadratic average of F_V and F_Q :

$$F = \sqrt{F_V^2 + (aF_Q)^2}, \qquad \text{with } 0 < a \le 1, \tag{3}$$

where the weighing factor a accounts for the lesser impact of peak discharge in the global quality assessment. A value of 0.5 is used for a. The curved lines in Fig. 2 represent the contours of the F surface as a function of F_V and F_Q; the straight line is the locus of equal sensitivity of F to F_V and F_Q. It can be seen that F, as a quadratic average, is much more sensitive to the more poorly satisfied of the two criteria when the difference between the two is large (i.e. near the axis corresponding to that poorly satisfied criterion). The priority given to the volume criterion appears clearly when one sees that F becomes really sensitive to F_Q only when F_V is low or when F_Q is very large. In other words, the peak discharge will come into significant play only when the fit on volumes is sufficient or when the discharge criterion is very bad. Hereafter F_0 designates the value of F for $n = n_0$, i.e. for the entire 1992–1998 reference sample, and is used as the tuning criterion (Section 2.4) and as a comparative measure for parameter acceptability (Section 2.6); F_0^2 is analogous to a weighted residual variance proportion. The addition of a penalty term to the diagnostic function was considered, to reflect a decreasing (K, C, M) likelihood with increasing distance from the prior $(1, 1, 1)$ value and thereby avoid an overly different calibrated parameter set, but tests showed that this is not necessary in the present case, partly due to the good performance ($F_0 = 0.196$, see Section 3.1) of this prior parameter set.

2.4. Function minimization, parameter sampling

Model calibration is performed by searching for minima of the F_0 function in the 3D K–C–M parameter space. Powerful automatic search algorithms may be required to find the global minimum when the problem dimension is large (e.g. Duan et al., 1992). When the dimension is small like in

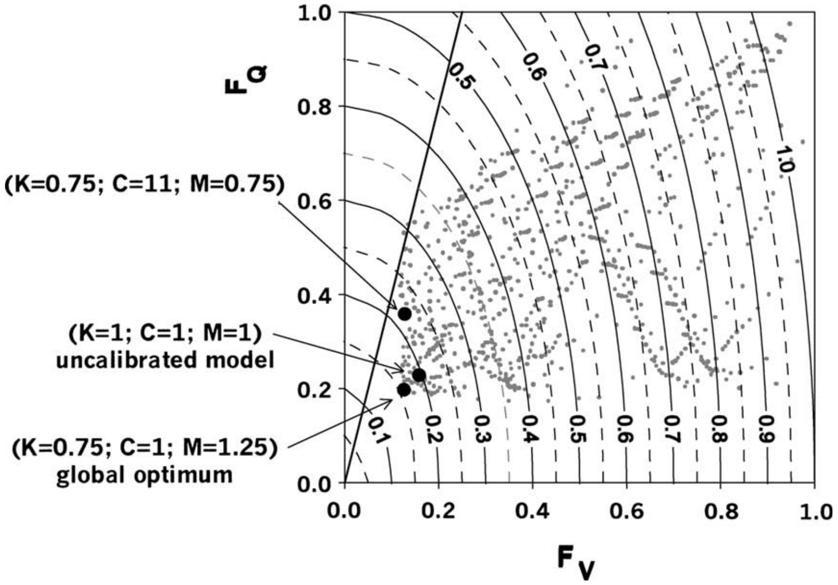

Fig. 2. Contours of diagnostic function F in (F_V, F_Q) space (straight line = equal sensitivity of F to F_V and F_Q), and F_0 locations (dots) of parameter sets explored in calibration and validation procedures.

the present case, direct analysis of the surface is possible, and is very informative as for possible multiple optima, parameter interactions and sensitivity to parameter values (e.g. Gaume et al., 1998), i.e. about uncertainty on calibrated parameters. Calibration is performed here by exploring the 3D parameter space at the nodes of an irregular, manually adapted grid with finer spacing around the function minimum and looser spacing away from the minimum, within what is considered to be a reasonable parameter subspace, centered around the $(1,1,1)$ prior point and spanning about one order of magnitude in K and M, and over two in C. To reflect the fact that the channel should not be less pervious than the hillslopes, a constraint is introduced that keeps the effective value of channel hydraulic conductivity (i.e. C times the prior value) larger than or equal to the lowest hillslope effective conductivity (i.e. K times the lowest K_s-map value). In order to obtain the solution to the minimization problem with adequate precision, the parameter grid is developed sufficiently in its vicinity to ensure an insignificant function gradient at the optimal node $(\hat{K}, \hat{C}, \hat{M})_0$. As put forward for instance by Beven and Binley (1992) under the concept of equifinality, the $(\hat{K}, \hat{C}, \hat{M})_0$ optimum is not the only parameter set worthy of interest, the neighboring sets whose F_0 performances are nearly as good should also be considered as potential candidates to represent the physical system. Therefore, the $F_0(K, C, M)$ surface around the $(\hat{K}, \hat{C}, \hat{M})_0$ optimum is used to represent this notion of multiple acceptable parameter sets, and serves as a basis for the parameter uncertainty characterization method of Section 2.6.

2.5. Resampling-based cross-validation

Because the available reference data set is not sufficient for fully-fledged split-sampling calibration/validation, the verification approach includes limited split-sample testing (data for the 1999 and 2000 seasons are used for verification only, see Section 2.7) as well as a resampling-based cross-validation scheme. This makes optimal use of the available data, since the size and representativeness of the calibration sample are maximized, while the quality of calibration output can still be checked. The cross-validation scheme used amounts to mimicking a validation sample in order to both test parameter stability with respect to the calibration sample, and assess the predictive capability of the calibrated model measured by the F criterion, much like the so-called *press* (short for 'predictive sum of squares') in regression methods (Draper and Smith, 1981). In this subsection, and in its follow-up Section 3.2, calibration refers only to the model's optimal parameter set, not to the parameter uncertainty aspect. Cross-validation with resampling is a commonly used technique to study the sensitivity of models to data sample fluctuations (see for instance Stone, 1974; Tukey, 1977). It basically consists in performing multiple calibrations using multiple data sets, each one being a subsample of the original data set obtained by leaving out a certain number of observations, one in our case ('leave out one observation' approach). Specifically, the model is re-calibrated for each subsample of size $n_0 - 1$ that can be obtained from the original calibration sample of size n_0, leading to n_0 new estimated parameter sets denoted $(\hat{K}, \hat{C}, \hat{M})_{\setminus i}$ where i designates the left-out event, between 1 and n_0. Comparison of the $(\hat{K}, \hat{C}, \hat{M})_{\setminus i}$ triplets with the globally-optimal, calibrated set $(\hat{K}, \hat{C}, \hat{M})_0$ informs about the stability of calibration-produced parameters. An F_0 like value is now computed by using for $V_{\text{sim}}(i)$ and $Q_{\text{sim}}(i)$ in Eqs. (1) and (2) the volume and peak discharge obtained by simulation of event i with parameter set $(\hat{K}, \hat{C}, \hat{M})_{\setminus i}$, which is independent from the $V_{\text{obs}}(i)$ and $Q_{obs}(i)$ observations for that event. This 'simulated' F_0 value (denoted F_0^{cv}) is a good indication of the predictive performance of the calibrated model, i.e. of its F score on a new, independent data sample representative of the storm event population. Some degradation relative to the calibration score $F_0^{\text{c}} = F_0((\hat{K}, \hat{C}, \hat{M})_o)$ is to be expected, but should remain moderate in order to accept the calibrated model. While cross-validation is commonly used for statistical or neural models (e.g. Prechelt, 1998; Coulibaly et al., 2000), very few applications can be found for conceptual or distributed hydrologic models (see for instance, Berri and Flamenco, 1999). The partial, recalibration steps are performed here using the same parameter-space sampling from Section 2.4, and may therefore be somewhat more approximate than the calibration step of Section 2.4 (no grid refining around new minimum), without weakening the conclusions from this cross-validation procedure.

2.6. Parameter uncertainty characterization

While the minimum of the F_0 function in the parameter space provides the single best parameter set, the neighboring region of parameter sets that produce close F_0 values can be viewed as nearly, albeit not quite, as acceptable as this optimal set. The farther a parameter set's F_0-value from the minimum $F_0^{\text{c}} = F_0((\hat{K}, \hat{C}, \hat{M})_0)$, the less consistent it is with the data, and vice-versa. Hence, the acceptability of a parameter set can be rated by its F_0-value, which means that a region of acceptable parameter sets may be defined in the parameter space as the compact subspace confined by some closed iso-F_0 surface. Parameter uncertainty can be characterized by analyzing how far from $(\hat{K}, \hat{C}, \hat{M})_0$ this region must extend in order that a prescribed fraction of the observed event volumes be properly simulated at least once when parameter sets are varied all over this region. Practically speaking, a parameter region is considered successful in simulating a given observed event when both underestimations and over-estimations of its runoff volume occur for the various parameter sets sampled over that region. This allows a graph to be built of F_0 against the fraction f_E of successfully simulated events within the parameter region defined by the F_0 value (Section 3.3). An experimental delineation (by an F_0 contour) is thus obtained of the parameter region to be considered for the inclusive reproduction of any given fraction f (%) of observed volumes. Conversely, the range of predicted volumes produced by this parameter region for any particular rain event (observed or not) can be considered as representing an empirical f-% confidence interval estimate for that event volume (see Section 3.4).

2.7. Test with the 1999–2000 event data

In order to check that performance of the calibrated model with the optimal parameter set $(\hat{K}, \hat{C}, \hat{M})_0$ or with its neighbors (acceptable parameter sets) is more or less preserved when applied to a different, fully independent sample of events, the 24 reference events of the 1999–2000 period are used to perform this test. Model performance is assessed with the same F function used for calibration and cross-validation, but with a smaller number n of events for the computation of quadratic errors in F_V and F_Q's numerators. As explained in Section 2.3, denominators in F_V and F_Q are left unchanged, i.e. taken as the volume and discharge variances of the 1992–1998 reference sample of $n_0 = 73$ events (see Eqs. (1) and (2)). Hence, comparing F-values (or F_V and F_Q separately) for the calibration sample, i.e. F_0, and for the 1999–2000 sample, denoted $F_{1999-2000}$, with the same model (same parameter set) amounts to comparing weighted mean quadratic error for the two samples. Results are shown in Section 3.5 for all parameter sets sampled in the $F_0 \leq 0.4$-region. The 1999–2000 data are also used to test the stability of the parameter uncertainty characterization scheme, by comparing, for the two distinct event samples (calibration and 1999–2000), the simulation 'success' rates of F_0-defined parameter regions, as defined in Section 2.6. These tests with the 1999–2000 sample are of course very partial, the sample being to short to fully represent the event population. In particular, it does not include any very large event. Therefore the test cannot be totally conclusive by itself, but complements the indications provided by the cross-validation test.

3. Results

3.1. Calibration

A total of 942 sets of three parameters (K, C, M) were processed, with steps in the ranges 0.05–0.5, 0.125–2.0, and 0.1–0.25, respectively, leading to a single function minimum and no alternative significant local minimum. The minimum F_0 value is $F_0^c = 0.160$, obtained for $(\hat{K}, \hat{C}, \hat{M})_0 = (0.75, 1.0, 1.25)$ considered as the globally optimal parameter set. It brings a noticeable improvement to the initial score of $F_0 = 0.196$ for the prior parameter set $(1, 1, 1)$. A second-order analysis performed around this point shows that any further refinement of the solution is not significant. The $F_0(K, C, M)$ diagnostic function surface appears to be well-conditioned (i.e. the model parameterization can be considered as yielding a well-posed calibration problem), with rapidly increasing values when crossing the boundaries of the explored parameter domain outwards. Only more model degradation is to be expected outside this domain, because of the monotonous marginal effect of any one of the three parameters on the runoff variables, from model construction. A partial graphical representation of this 4D hyper-surface is obtained by projections onto 2-parameter planes, in the neighborhood of the optimum: Fig. 3 shows the maps of F_0 contours, interpolated between computed values, for the 3D surfaces $F_0(K,C{=}1,M)$, $F_0(K,\log_{10}(C),M{=}1.25)$ and $F_0(K{=}0.75,\log_{10}(C),M)$. The C parameter is plotted with a logarithmic scale because of its noticeably lesser sensitivity, already pointed out in Section 2.1. The surfaces are rather smooth, with a single, well-confined region of low F values in the K–M subspace (Fig. 3(a)) and more elongated troughs in the K–C and M–C subspaces (Fig. 3(b) and (c)) extending over a large range of C values given the logarithmic axis, and nearly parallel to this axis especially in the K–C subspace. Hence, parameter uncertainty is the least for

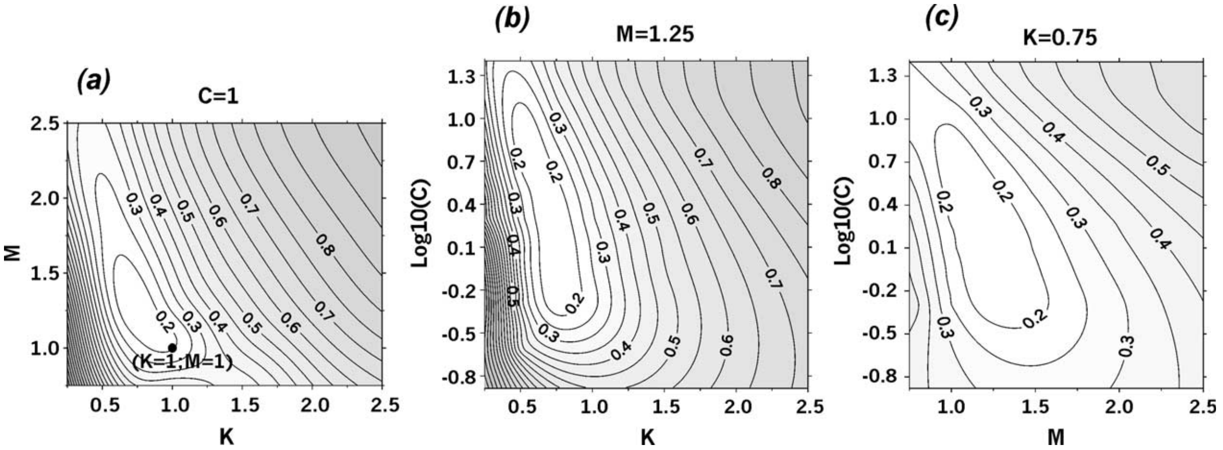

Fig. 3. 2D-projections of diagnostic function around global optimum: (a) $F_0(K, C = 1, M)$; (b) $F_0(K, C, M = 1.25)$; (c) $F_0(K = 0.75, C, M)$ (note that C is plotted with a logarithmic scale, in (b) and (c)).

K, then for M, and is much larger for C, as foreseen by the preliminary sensitivity analysis in Section 2.1. Fig. 3 shows that, although no alternate optima exist, multiple parameter sets may be nearly equally acceptable, especially for a broad range of C-values. Moreover, the tilt of the low F values especially in the (K,M) and (C,M) subspaces illustrates the interactions between soil conductivity and surface roughness parameters, embedded in the model's coupled formulation of the infiltration and transfer equations. Comparatively, K and C appear to be somewhat more independent from one another, in the tuning region. Although the optimal point $(\hat{K},\hat{C},\hat{M})_0$ found in the parameter space will be paid special attention, the existence of alternative parameter candidates will also be taken into account, through the procedure used for parameter uncertainty characterization (Section 3.3).

Model control by the parameterization scheme is further scrutinized in Fig. 2, where the explored parameter sets are plotted in the (F_V, F_Q) space (actually limited to values below 1, for legibility) with F_0 contours in the background. The cloud of points is bounded by non-zero F_V and F_Q minima (0.119 and 0.173), corresponding to the best possible simulations of either observed volumes or peak discharges, respectively. Excellent volume matching may be associated with bad peak discharge, and vice-versa. The optimal F_0 value combines only slightly suboptimal F_V and F_Q scores (0.127 and 0.197), and can therefore be considered to achieve an adequate compromise between the two criteria: volumes are not over-fitted at the expense of unrealistic hydrograph shape. Had the F_Q term not been included in the F_0 calibration criterion, then a much more fragile model would have been produced, for only a very slight gain in volume representation of the calibration sample. As a matter of fact, with the F_V criterion alone a much larger range of parameter values would have been acceptable candidates, all corresponding to quite different model behaviors as evidenced by the variations in F_Q scores. This emphasizes the importance of additional, behavioral variables in the diagnostic function besides the target variable itself, V in our case. The choice of the 'blending' coefficient (a in Eq. (3)) value is only of second-order importance, its role is merely to arbitrate between a few very close parameter sets (points clustered in the very lower-left corner of Fig. 2) that indeed appear as nearly equivalent when parameter uncertainty is considered as in Section 3.3.

Table 1
Compared global statistics (1992–1998 reference sample) for event runoff observation, calibration ($(\hat{K},\hat{C},\hat{M})_0$ optimal parameter set) and cross-validation; relative departure from observation in brackets

		Observed	Calibration	Cross-validation
Runoff volume (mm)	Average	1.3	1.2 (−9%)	1.1 (−10%)
	Std. deviation	2.4	2.4 (+1%)	2.3 (−3%)
	Largest	16.9	17.2 (+1%)	16.0 (−6%)
Peak discharge (m^3/s)	Average	0.9	0.7 (−22%)	0.8 (−17%)
	Std. deviation	1.6	1.5 (−5%)	1.9 (+18%)
	Largest	10.8	11.2 (+3%)	14.8 (+36%)
Diagnostic function	F_V		0.127	0.137
	F_Q		0.197	0.377
	F		0.160	0.233

In Table 1 (first two columns), some global statistics of simulated runoff volume and peak discharge produced by the $(\hat{K},\hat{C},\hat{M})_0$ optimal parameter set are compared with observations. While standard deviations and maxima (July 18, 1992) for both volume and peak discharge are quite close to observed values, mean peak discharge is underestimated by 22%. This reflects the lesser weight given to peak discharge in the objective function. Filled triangles in Fig. 4 compare observed and simulated event volumes (a) and peak discharges (b) for the $n_0 = 73$ events of the calibration sample. The latter indicates that the rather poor performance cited above for mean peak discharge mainly comes from a great number of small events. Displayed for illustrative purposes in Fig. 5 are the observed (bold line) and simulated (thin line) hydrographs for four selected events of varying, increasing magnitudes. Small-event peak-discharge underestimation is again evidenced (Fig. 5(a) and (b)), while bigger-event hydrographs appear very adequately simulated ((c) and (d)). Although to a lesser extent, this phenomenon also exists for volumes, leading to a 9.2% underestimation by the calibrated model of the volume mean over the sample (the underestimation is nearly

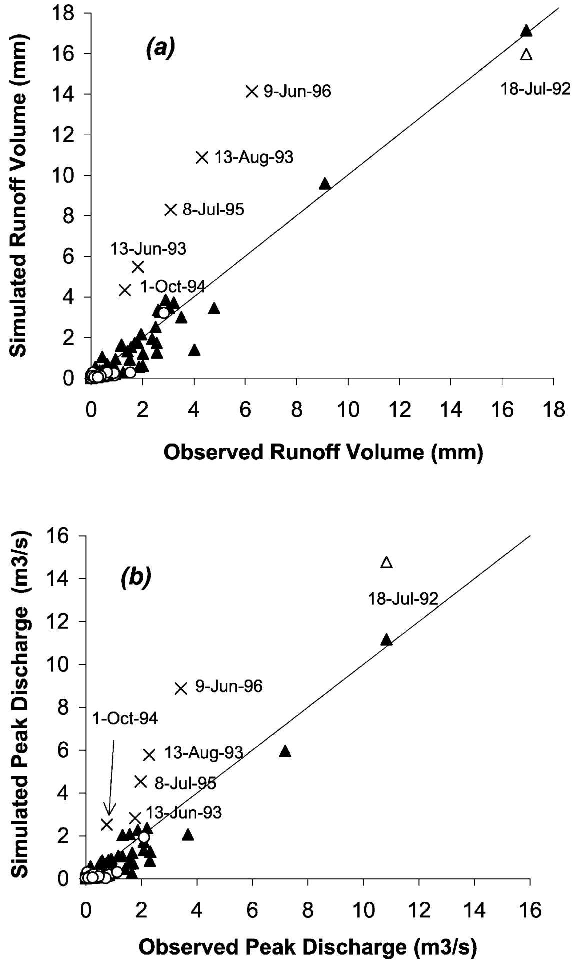

Fig. 4. Simulated vs observed volumes (a) and peak discharges (b) for: 73 reference events of 1992–1998 period with calibrated parameter set $(\hat{K},\hat{C},\hat{M})_0 = (0.75, 1.0, 1.25)$ (filled triangles) and with cross-validation scheme (empty triangles, coinciding with filled triangles for all but largest event of July 18, 1992); 24 events of 1999–2000, with $(\hat{K},\hat{C},\hat{M})_0 = (0.75, 1.0, 1.25)$ calibrated parameter set (circles); 5 'LOW' outlying events in classification of raw 1992–1998 sample (Peugeot et al., 2003), with $(\hat{K},\hat{C},\hat{M})_0$ calibrated parameter set (' × ' crosses).

25% for the uncalibrated model). This bias is a potential concern with respect to future model operation for predicting seasonal catchment yields. However, it is shown in Section 3.4 how the bias is essentially eliminated when the optimal parameter set is not the only one to be considered, i.e. when predictions are made as confidence intervals.

3.2. Cross-validation

The cross-validation scheme described in Section 2.5 was first used to investigate the stability of the calibration procedure, that is to say its sensitivity to the event sample. Successively for each of the $n_0 = 73$ subsamples of $n_0 - 1$ events from the 1992–1998 reference sample, 942 new F values (one for each of the previously explored parameter sets (K, C, M)) are computed (with $n = n_0 - 1$ in Eqs. (1) and (2)), the minimum of which yields the optimal parameter set for that subsample. Hence, these n_0 partial calibration steps lead to n_0 'partially'-optimal parameter sets $(\hat{K},\hat{C},\hat{M})_{\backslash i}$ ($i = 1$ to n_0); all but one turn out to be identical to the 'globally'-optimal set $(\hat{K},\hat{C},\hat{M})_0 = (0.75, 1.0, 1.25)$ produced by the full calibration step of Section 3.1 (strict identity results from the discrete parameter sampling scheme). The only exception occurs when withdrawing the largest event in the reference sample (July 18, 1992): in this case the partially calibrated parameter set is $(\hat{K},\hat{C},\hat{M})_{\backslash(\text{July 18, 1992})} = (0.75, 11.0, 0.75)$, i.e. C unrealistically departs from its prior and globally-optimal value ($\hat{C}_0 = 1.0$), unlike K and M. For this $(\hat{K},\hat{C},\hat{M})_{\backslash(\text{July 18, 1992})}$ parameter set, the F_0 value computed in Section 3.1 with the entire event sample is 0.220 ($F_V = 0.129$ and $F_Q = 0.357$, see Fig. 2). The model calibration may be called stable, being sensitive only to the absence of the largest event in the calibration sample, which does significantly impact the peak discharges produced by the resulting model but not the simulated runoff volumes, i.e. this study's prime concern.

This is confirmed when assessing the predictive capability of the calibrated model, as defined in Section 2.5: each event i is now simulated with the partial-calibration parameter set $(\hat{K},\hat{C},\hat{M})_{\backslash i}$ obtained without that event. The last column in Table 1 shows the statistics and F-value produced by this 'simulated' validation sample. For volumes the results are very similar to those obtained in the calibration step, whereas degradation in the F_Q component of the diagnostic function leads to an overall F_0^{cv} score of 0.233. In fact, since the largest event is the only one whose presence or absence in the calibration sample makes a difference in the estimated

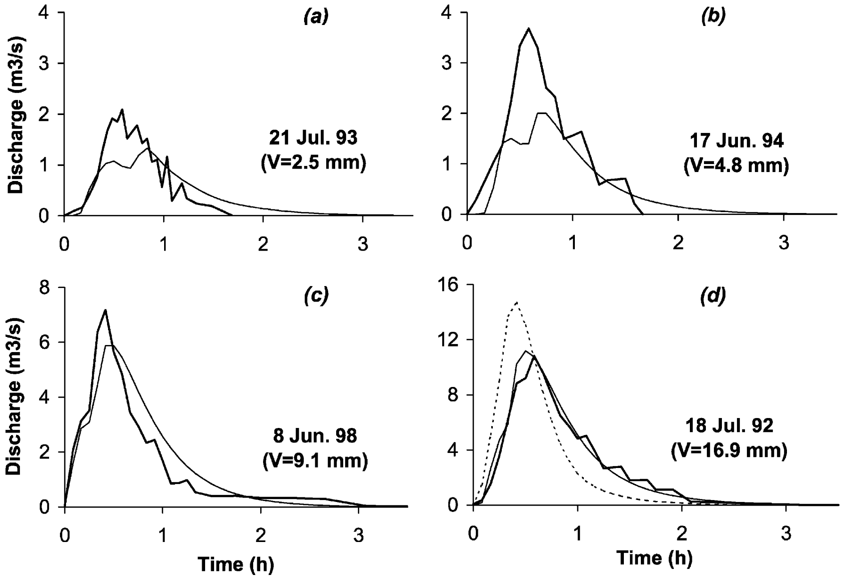

Fig. 5. Hydrographs for selected events: bold line = observation; thin line = simulation with calibrated parameter set $(\hat{K},\hat{C},\hat{M})_0 = (0.75, 1.0, 1.25)$; dotted line = cross-validation simulation with $(\hat{K},\hat{C},\hat{M})_{\backslash i}$ (coincides with thin line in (a) to (c)).

parameters, the F_0^{c}-to-F_0^{cv} deterioration comes entirely from the poorer $Q_{\max}$ plus the much more slightly degraded V for that remarkable event when it does not contribute to calibration (i.e. for $(\hat{K},\hat{C},\hat{M})_{\backslash(\text{July 18, 1992})}$, see ‘largest’ values in Table 1, empty triangles in the scatterplots of Fig. 4, and dotted-line hydrograph in Fig. 5(d)). This cross-validation exercise undoubtedly suffers from the scarcity of very large events in the reference data sample, only one belonging to that category. However again, our variable of main concern is the runoff volume which appears to be properly replicated by this validation step. Also, the relatively poor F_Q^{cv} score appears to be an overly pessimistic quantification of hydrograph shape deterioration for the validation sample: the only degraded hydrograph, i.e. July 18, 1992 (Fig. 5(d), dotted line) is not that different in shape from the observed curve, but it is the weight in F_Q of any error (even if limited in relative terms) on this event’s large peak discharge that tends to overemphasize the degradation. In fact F_Q suffers from not being a pure hydrograph shape-agreement indicator, but a variable that also carries information about event magnitude, as F_V already does. Furthermore, as already mentioned, precision is much less here for peak-discharge data than on volume or general hydrograph pattern. A better shape-criterion could probably have been devised, bearing for instance on relative peak-discharge errors, or on errors on some hydrograph-duration characteristic, e.g. the shortest duration for a given runoff volume fraction (the runoff time dimension is much less sensitive to event magnitude and to measurement error than discharge is), but this is beyond the scope of this paper.

From the above, it may be concluded that the simulation capability of the calibrated model is largely preserved through cross-validation resampling. Finally, it must be emphasized that, far from being an ‘outlier’, the large event of July 18, 1992, is an extremely informative one which should not be omitted from the event dataset as it carries a lot of information about the modeled system, especially in the prospect of seasonal runoff volume simulation. Being the only member of the under-represented class of large to very large events, its solitary presence in the reference sample penalizes cross-validation performance, but it is essential that model calibration fully accounts for this particular event, as is achieved through error weighting in the diagnostic function F_0 (α_i coefficients in Eqs. (1) and (2)).

3.3. Parameter uncertainty characterization

The F_0 measure of parameter acceptability was rated in terms of fraction of properly simulated volumes (f_E) as described in Section 2.6. Because of the heterogeneity between very small and very large events, the overall 1992–1998 reference sample of $n_0 = 73$ events was split into two halves, and the rating performed separately on each of these two subsamples. To make this uncertainty characterization readily usable for predictions, the splitting criterion used to differentiate event magnitude is the runoff volume $V_{\rm sim}^{\rm c}$ simulated with the optimal $(\hat{K}, \hat{C}, \hat{M})_0$ parameter set. Hence two separate F_0-vs-f_E graphs are constructed (Fig. 6), for the 37 and the 36 events with $V_{\rm sim}^{\rm c}$ above and below 500 m^3 (0.27 mm), respectively. In the determination of f_E, measurement uncertainty on pool recharge volumes associated with the 1 cm precision on stage recording is accounted for, as a function of pool stage. Each graph displays an expected overall shape with sharply contrasted, globally decreasing slopes (upward concavity), consistent with a roughly unimodal density distribution around the $(\hat{K}, \hat{C}, \hat{M})_0$ optimal parameter set. Improved parameter space sampling would logically result in smoother curves. The heterogeneity between small and big events is clearly highlighted, with a larger parameter uncertainty for the former. Using the one-to-one relationships defined by Fig. 6, the hypersurface displayed as projections in Fig. 3 could be redrawn with parameter sets now rated in terms of fraction f_E of correctly simulated events, for each of the two subsamples. The use of this method of parameter uncertainty characterization in prediction is presented below (Section 3.4).

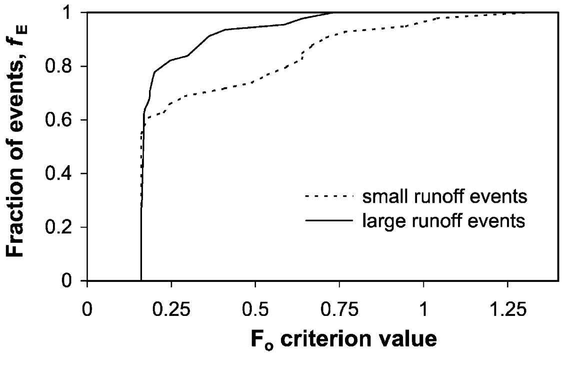

Fig. 6. Parameter uncertainty characterization, as $f_E(F_0)$ (fraction of successfully simulated events f_E within the parameter subregion encompassed by an F_0 contour), for calibration events with $V_{\rm sim}^{\rm c}$ (simulated runoff volume using optimal parameter set $(\hat{K}, \hat{C}, \hat{M})_0$), above (solid line) or below (dotted line) threshold of 500 m^3 (0.27 mm).

3.4. Prediction intervals

Based on parameter uncertainty as expressed by Fig. 6, it is possible to derive empirical confidence limits on volume prediction for any simulated event. This obviously is of great value for operational model use with unmonitored events, but in the current model development phase it is also interesting to explore these generated confidence intervals for the reference sample events. The idea is that, as a result of the F_0-vs-f_E calibration rule, a volume prediction interval at a given f_E confidence level can be obtained by simulating with all parameters sets within the $F_0(f_E)$ region. Practically, for any given rainfall event, the model is first run with the optimal $(\hat{K}, \hat{C}, \hat{M})_0$ parameter set to allow identification of the runoff event type, either small or big (Section 3.3), and selection of the associated parameter uncertainty curve from Fig. 6. Then, out of the 942 parameter sets, only those with F_0 values below the upper limit ($f_E < 1$) of the appropriate uncertainty curve need to be simulated, and computed $\{F_0$, Volume$\}$ couples are sorted in increasing-F_0 order. The gradual growth of the volume interval with increasing F_0 is represented through the evolution of its lower and upper bounds, obtained as the running extremes along the sorted $\{F_0$, Volume$\}$ series. These two curves of F_0 vs volume-bound can then be translated directly into graphs of volume confidence limits by replacing the F_0 axis with the event fraction f_E axis, using the proper parameter uncertainty curve from Fig. 6. Hence confidence intervals on volume prediction for an event are obtained not only for a particular confidence level, but across the whole range of f_E confidence values, as a function of the latter. Fig. 7 shows two contrasting examples of such confidence graphs, for the predicted runoff volumes of July 18, 1992, and June 27, 1998, respectively. The latter displays strong asymmetry typical of smaller events.

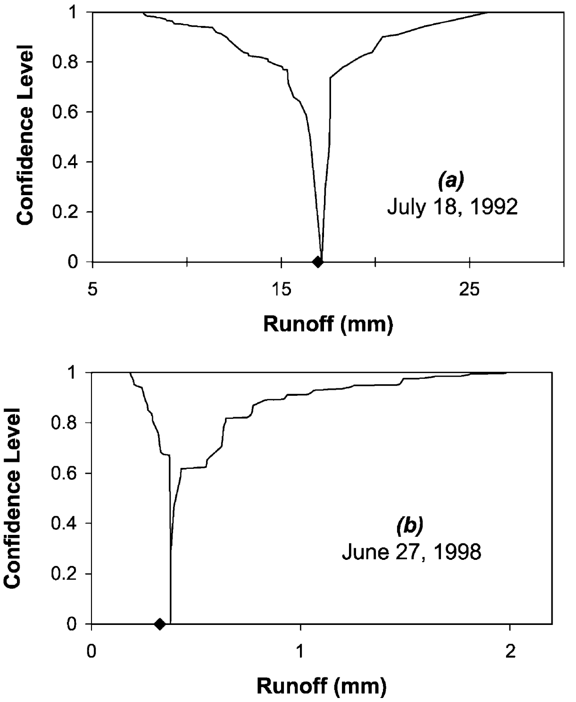

Fig. 7. Examples of predicted runoff confidence graphs (limits of predicted volume interval, on abscissa, for a given confidence level, on ordinate) for two individual events, big and small: (a) July 18, 1992; (b) June 27, 1998; diamonds are observed volumes.

The above analysis quantifies the effect of parameter calibration uncertainty on runoff volume predictions for individual events, but the model's main practical use will eventually be to estimate seasonal volumes. It is therefore highly desirable to be able to produce confidence intervals for any linear combination, like the sum or the mean, of predicted runoff volumes from a collection of events. It will be seen hereafter how this approach further highlights and helps to get round major shortcomings of the use of the optimal parameter set alone, namely: bias on cumulative volume prediction due to non-nil mean event-wise error (already mentioned in Section 3.1), as well as unlikeliness of this aggregated prediction in the overall range of possible values (see below). Prediction errors on individual events may reasonably be assumed to be independent, thus confidence intervals on combined volumes can be produced through Monte Carlo simulations using independent probability distributions for individual event volumes, which can be derived from the confidence graphs exemplified in Fig. 7. This was performed for the mean volume of the 73 events from the 1992–1998 reference sample, hereafter called 'combined volume'. A cumulative distribution function was generated for each event from its own confidence graph, and 10,000 samples of 73 event volumes were simulated by randomly drawing from each of these distributions. The resulting histogram for the combined volume is shown in Fig. 8. As expected the distribution is quasi-gaussian, with a mean value of 2465 $\mathrm{m}^3 = 1.31$ mm (roughly equal to the mode), and a 6.4% variation coefficient. This combined volume expectation over the 1992–1998 reference sample is only 2.4% off the observed value, significantly less than the above variation coefficient, meaning that when confidence intervals on event volumes are considered instead of only their most probable values (those produced by the optimal parameter set) then the bias on volume estimations vanishes. In fact, the combined volume produced over the reference sample by the optimally tuned model appears to be a very unlikely value when prediction intervals are considered, since it is located 1.8 standard-deviation away from the expected value (Fig. 8); probability of a lower combined volume is only 2.6%. This is due to the largely asymmetrical distributions for most individual events. Illustrated here is the danger of considering only a single, most 'reasonable' set of model parameters, not only because it disregards the full range of possible output values, but also because those predictions may, as in our case, take relatively unlikely values and lie relatively far from the most

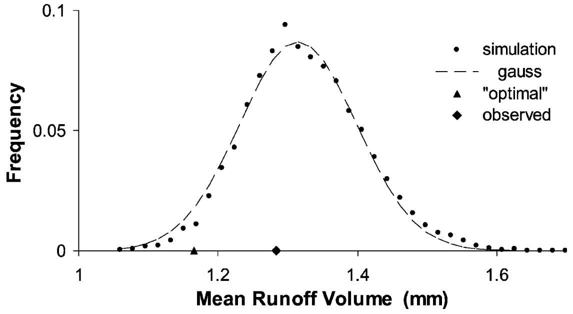

Fig. 8. Distribution of Monte Carlo-simulated event volume mean over 1992–1998 reference sample (diamond: observed mean; triangle: computed with optimal parameter set $(\hat{K}, \hat{C}, \hat{M})_0$).

probable ones. Making predictions with confidence intervals here allows us to obtain de-biased estimations for expected volumes, and consequently to more safely extend those predictions to seasonal event sequences for water resources investigation.

3.5. Test with 1999–2000 data

The simulated volumes and peak discharges obtained with the calibrated parameter set $(\hat{K}, \hat{C}, \hat{M})_0 = (0.75, 1.0, 1.25)$ for the $n = 24$ reference events of the 1999 and 2000 rainy seasons are compared with observations in Fig. 4 (circles). It can be seen that the agreement is generally fine, with values of $F_V = 0.096$ and $F_Q = 0.121$, yielding $F = 0.114$ for the 1999–2000 storm subsample. The response hyper-surfaces $F_0(K, C, M)$ and $F_{1999-2000}(K, C, M)$ are compared via the scatterplot of Fig. 9 which crosses the F scores obtained with the two data samples (calibration and 1999–2000 reference samples), for all parameter sets sampled in the $F_0 \leq 0.4$-region. The two variables are bound in a triangle that has the following properties: its vertex corresponds to the above optimal parameter set $(\hat{K}, \hat{C}, \hat{M})_0$, its lower edge is horizontal and is associated with the minimum $F_{1999-2000}$ value (~ 0.1), and its upper edge rises with a roughly 1:1 slope. Hence, when F_0 increases, the mean and standard deviation of the $F_{1999-2000}(F_0)$ conditional distribution both grow steadily, while this distribution remains entirely below F_0 : the performance tested with this independent sample is always at least as good as that obtained with the calibration sample, for any F_0 region taken around the calibrated parameter set. For instance, the range of $F_{1999-2000}$ is approximately 0.1–0.3 for F_0 up to 0.4 (please note that the $F_0 \leq 0.4$-region is associated with a volume-confidence level slightly above 80% as rated by the overall fraction of successfully simulated events from the calibration sample, see Section 3.3).

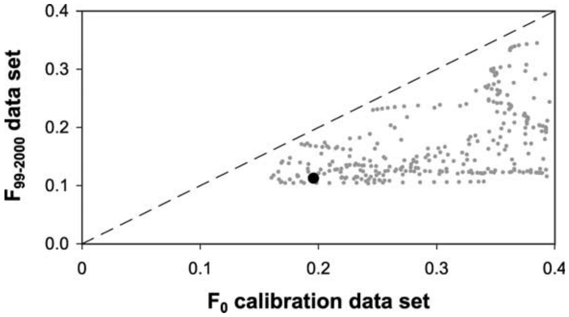

Fig. 9. Comparing parameter performance on calibration (1992–1998) and verification (1999–2000) reference data sets; filled circle is prior parameter set (1, 1, 1).

The volume-confidence rating of parameter regions produced by calibration (Sections 2.6 and 3.3, Fig. 6) can be tested against success frequencies obtained with the 1999–2000 event sample, when using the same, F_0-controlled regions. Only the smaller-event curve can be tested since all but one of the 24 events in the 1999–2000 sample fall in that category. While the curve produced with this latter sample (not shown) does generally associate to any given F_0-region a somewhat lower level of confidence than the calibrated curve, this difference remains reasonable, always below 10% (mean difference is 5.6%). The bias resorption effect of using prediction intervals is also obtained with the 1999–2000 reference sample, even more so than with the 1992–1998 calibration data: while the ratio of simulated-over-observed combined volume is only 0.56 when the $(\hat{K}, \hat{C}, \hat{M})_0$ optimal parameter set is used alone, its expected value is 0.83 when parameter uncertainty is accounted for, with a 27% variation coefficient. Albeit the limitation stemming from sample representativeness, these results contribute to corroborating the model calibration obtained in the explored parameter space, both in terms of the optimal parameter set $(\hat{K}, \hat{C}, \hat{M})_0$ and of quantification of parameter uncertainty around this optimum.

3.6. Discussion: internal catchment behavior, and posterior analysis of screened-out events

Except for catchment outflows, no quantitative data is available for confrontation with the various state variables of the model. Nevertheless, it is of interest to examine a few of these distributed variables to get some insight into the model's internal behavior, and check it against qualitative information. For lack of spatially distributed storm-period observations, only field-estimated conveying capacities can be compared with simulated water depths and discharges, for the calibrated parameter set $(\hat{K}, \hat{C}, \hat{M})_0$. Values produced by the largest reference event (July 18, 1992) are most pertinent for this purpose.

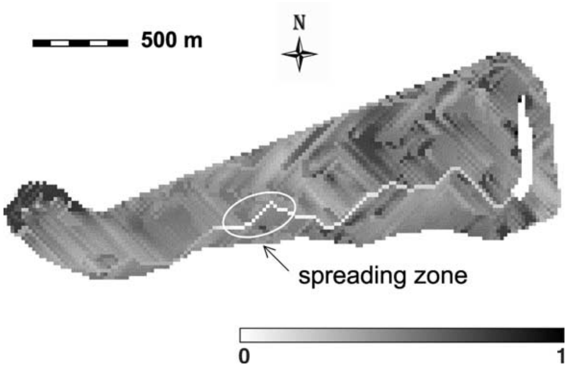

Fig. 10. Map of simulated runoff coefficient for July 18, 1992, with calibrated parameter set $(\hat{K},\hat{C},\hat{M})_0 = (0.75, 1.0, 1.25)$; pond is white.

The largest simulated overland discharge is 2.3 m^3/s, for a water depth of 27 cm, in eroding car-tracks (just slightly visible along the very northern basin boundary in Fig. 10) where such values are to be expected. Everywhere else on hillslopes, water depths are at most 2 cm. In the main channel, the largest discharge is 4.4 m^3/s while maximum water depth is 1.9 m, in line with the bank-full depth of the ravine and with field-observed flood marks. It is interesting to note that the ravine, which represents 55% of the catchment area, contributes only 31% of the total pool recharge for this event, and even less for most other simulated events. This is yet more pronounced for the runoff contribution from the upper 28% of the catchment, through the mid-slope spreading zone. As already seen with the uncalibrated model (*C.Pap.*), only insignificant volumes run off from the spreading zone even for the largest storms of the reference sample with the optimal parameter set. This result remains true for a wide parameter region around the optimal set, including the cross-validation $(\hat{K},\hat{C},\hat{M})_{\backslash(\text{July 18, 1992})}$ parameter set, and is consistent with field observations. Fig. 10 shows the spatial distribution of runoff coefficient (ratio of runoff volume to upstream-precipitated rainfall) for July 18, 1992, as simulated with the $(\hat{K},\hat{C},\hat{M})_0$ optimal parameter set (whole catchment value is 0.30).

While the runoff data screening performed in *C.Pap.* made use of the uncalibrated model, it is not uninteresting to investigate a posteriori how model parameter values affect the event discrimination achieved. First, it is found that none of the 942 parameter sets explored for model calibration is able to produce acceptable simulated outputs both for the reference 1992–1998 data sample actually used for calibration and for the events that were rejected from that sample: the pattern of relative locations of those two event groups in a simulated-vs-observed volume scatterplot is always very similar to that presented in *C.Pap.* for the prior parameter set $(1, 1, 1)$. Second, when volume prediction intervals are calculated with the calibrated model for all 57 rejected events of the 1992–2000 period (all rejection categories) and compared to observations, only 11% of them fall in the 80%-confidence intervals, 16% in the 90% intervals. Among these, none of the events that were ultimately screened-out based on statistical and model analyses (‘pEXO’ and ‘LOW’ categories, see *C.Pap.*) fall in the 90% intervals. All this strengthens the conclusion that it is very unlikely that the entire original set of observations belongs to a single, homogeneous population, thereby making selection of reference events necessary. Conversely, five large events with proven pond overflow (from GIN and EXO categories in *C.Pap.*) do appear well simulated against ‘apparent’ observed runoff, within the present calibration framework; among these, the three GINs produce more predicted runoff than any other of the 154 events simulated in this study (i.e. all categories). It may be pointed out that, had the evidence of pond overflow not been available for these five events, they would have been accepted as valid by the classification method of *C.Pap.*. This supports the calibrated model in the range of very large events, where the reference data set is rather weak.

In *C.Pap.*, the hypothesis is put forward that the five LOW events occurred under specific, distinct conditions which justify that they be excluded from the reference data set used for model calibration/validation. Indeed, simulation of these events with the calibrated parameters $(\hat{K},\hat{C},\hat{M})_0$ (see five ‘×’—crosses in Fig. 4) largely overestimates volumes compared to data, by a factor of 2.3–3.4, thereby corroborating the assumption that some unaccounted-for feature(s) comparatively dampen(s) runoff for these particular events. It is suggested in *C.Pap.* that for most of these events this may largely be due to temporarily rougher and more pervious farmed areas, from hoe-weeding. This interpretation is tested here by applying, within the cultivated parts of the catchment exclusively, two further, greater-than-one multiplicative scalar factors denoted K_{fields} and M_{fields} to the calibrated hydraulic conductivity and roughness maps, respectively, i.e. 'on top of' the optimal $K-C-M$ parameter set $(\hat{K}, \hat{C}, \hat{M})_0$. All channel reaches and unfarmed overland-flow areas are left unchanged for this test. Farming operations being tightly linked with the stage in the rain season and in millet-plant development, it is reasonable to assume total correlation between the states of all fields. Using an exploration step of 0.5 for these two extra, partial-space parameters, it is found that all five outlying events are now best, very closely and simultaneously reproduced by a set of K_{fields} and M_{fields} values nearing 2.5 and 3 respectively. These figures are consistent with reported infiltration rate increases under tillage, observed at the plot scale in the Sahel (respectively 6-fold and 1.5- to 7-fold by Lamachère, 1991, and Casenave and Valentin, 1992, under simulated rain; 7.2-fold by Stroosnijder and Hoogmoed, 1984). A large increase in hydraulic roughness is to be expected when one considers that in addition to the effect on the soil surface per say, tillage and weeding also result in above-ground accumulation of vegetation debris that hinder overland flow. Altogether, these figures seem quite realistic, and confirm that runoff sensitivity to enhanced field-area infiltration suffices to explain the occasionally-observed marked fluctuations in relative catchment yield as resulting from cultivation. It will be recalled that in this area the effects on soil structure disappear soon after each field operation, due to very fast crust restoration under a few centimeters of subsequent rainfall (Stroosnijder and Hoogmoed, 1984; Lamachère, 1991; Peugeot et al., 1997).

Spatial rainfall heterogeneity is also mentioned in *C.Pap.* as a possible source of occasional departure from the main-trend catchment behavior represented by the calibrated, uniform-rainfall model with the Mare hyetograph as sole input. Hyetograph comparison of the Mare and Ouest raingauges, which are located well apart in the catchment, over the 1993–1994 period of simultaneous recording, showed that differences can only rarely be large (*C.Pap.*). The strongly infiltrating channel and spreading zone, as evidenced by field observations and model calibration, result in a strong dampening effect on any rainfall heterogeneity that may occur over the basin. Tests with those three events, among the five 'outliers' (LOW), that belong to the 1993–1994 period for which Ouest data is available, show that for two of them (June 13 and August 13, 1993) hyetograph overestimation in the simulation with the calibrated model cannot explain the observed runoff volume overshoot. Only for the smallest of all five storms (October 1, 1994) is the discrepancy between the two hyetographs large enough to provide a partial explanation for the model error; note that large-scale cultivation is unlikely at that point in the growing season. Hence it is believed that rainfall non-uniformity is only rarely a significant source of runoff mis-modeling by the calibrated model, and that it should not represent a real problem in the context of seasonal runoff yield prediction for a catchment of that size.

4. Conclusion and prospects

Based on a reference rainfall–runoff data sample built from a 7-year long (1992–1998) record of rain intensities and pool level fluctuations in the 1.9 km^2 Wankama catchment (*C.Pap.*), the *r.water.fea* distributed, physically-based hydrological model set up for this catchment (*C.Pap.*) was calibrated through model control by three non-dimensional scaling parameters applied to prior maps of hydraulic conductivity and roughness in overland flow areas and channel reaches. A special diagnostic function was built to meet the specific needs and conditions of the problem in hand, in particular the emphasis on runoff volumes and the question of representativeness of the full event population by the reference data sample. A single calibration optimum was identified, making the equifinality dilemma inherent to model calibration less acute than in multiple-optima cases. This optimum is rather close to the prior parameter set. Parameter uncertainty around the optimum was characterized by rating the diagnostic function value, which defines a region in the parameter space, against successes in replicating individual observed event volumes, separately for small and large events. This empirical approach to calibration uncertainty enables the production of prediction intervals, a highly desirable feature for reasoned model operation. Event volume bias associated with the 'optimal' model is virtually eliminated when prediction intervals are considered, allowing safe aggregative upscaling over time, from event-wise to seasonal pool recharge. This method of uncertainty characterization is believed to have potential for use in various modeling applications.

Model verification was performed successively through a resampling-based cross-validation technique applied to the reference data sample, through fully calibration-independent runs on a separate event subsample obtained for the years 1999 and 2000 (split-sample approach), and through the analysis of internal simulated variables (peak flow discharges and depths over hillslopes and in channel reaches) and their confrontation to known qualitative field information. The calibrated model highlights some key processes controlling runoff in the catchment. Runoff production is high due to widespread soil crusts, but is dampened by intense channel and spreading-zone infiltration, leaving only small fractions of upper-catchment runoff actually contributed to pool recharge. Significant though short-lived pool recharge reduction may result from occasional tillage operations over the cultivated areas at the catchment scale.

Event-wise runoff simulation and model calibration, as performed in this study, appears to be a necessary step for seasonal-scale analyses of water resource renewal in the area. Model improvement is currently sought through dynamic interfacing with a vegetation growth model and an energy balance (SVAT) model. The uncertainty analysis component needs to be refined with better parameter space sampling, using for instance Kuczera and Parent's (1998) sampling scheme. Next, the modeling tool developed based on the data for the Wankama pool and catchment is intended to be used for extending recharge estimates to a much longer period of time and to a regional area containing several such runoff–collecting systems. The specific objective is to investigate, over significant time and space scales, the sensitivity to rainfall variability and to environment alteration. Subject to further testing, the physically based hydrological model has the potential to be transposed to other endoreic catchments in the same landscape, to changing land surface conditions associated with land use modifications, and to non-stationary climatic situations. Applying the model to the past few decades, simulated runoff evolution will be compared with the observed increase in groundwater recharge (Favreau et al., 2002), which is attributed to the extension of cropped surfaces over this rain-deficient period through enhanced runoff production and concentration. A preliminary investigation of possible climatic impacts on water resources in the area was made by Vieux et al. (1998), based on the uncalibrated Wankama model and on hypothesized rainfall reduction scenarios through possible cuts in event numbers, shown by Le Barbé and Lebel (1997) to be the prime mode of seasonal rain deficit in the region. In current studies with the calibrated model (Séguis et al., submitted), more elaborate models of rainfall chronicles over the past decades are considered, together with reconstitution of environmental evolution.

Acknowledgements

This study was funded by IRD (formerly ORSTOM) and by the Programme National de Recherche en Hydrologie, France. G. Favreau, H. Niel, and the referees are thanked for their reviews.

References

Berri, G.J., Flamenco, E.A., 1999. Seasonal volume forecast of the Diamante River, Argentina, based on El Nino observations and predictions. Water Resour. Res. 35 (12), 3803–3810.

Beven, K., Binley, A., 1992. The future of distributed models: model calibration and uncertainty prediction. Hydrol. Process. 6, 279–298.

Casenave, A., Valentin, C., 1992. A runoff capability classification system based on surface features criteria in semiarid areas of West Africa. J. Hydrol. 130, 231–249.

Coulibaly, P., Anctil, F., Bobée, B., 2000. Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. J. Hydrol. 230, 244–257.

Draper, N., Smith, H., 1981. Applied regression analysis, Wiley-Interscience, New York, 709 pp.

Duan, Q., Sorooshian, S., Gupta, V., 1992. Effective and efficient global optimization for conceptual rainfall–runoff models. Water Resour. Res. 28 (4), 1015–1031.

Favreau, G., Leduc, C., Marlin, C., Dray, M., Taupin, J.D., Massault, M., Le Gal La Salle, C., Babic, M., 2002. Estimate of recharge of a rising water table in semiarid Niger from ^{3}H and ^{14}C modeling. Ground Water 40 (2), 144–151.

Feyen, L., Vazquez, R., Christiaens, K., Sels, O., Feyen, J., 2000. Application of a distributed physically-based hydrological model to a medium size catchment. Hydrol. Earth Syst. Sci. 4 (1), 47–63.

Gan, T.Y., Dlamini, E.M., Biftu, G.F., 1997. Effects of model complexity and structure, data quality, and objective functions on hydrologic modeling. J. Hydrol. 192 (1–4), 81–103.

Gaume, E., Villeneuve, J.P., Desbordes, M., 1998. Uncertainty assessment and analysis of the calibrated parameter values of an urban storm water quality model. J. Hydrol. 210, 38–50.

Goutorbe, J.P., Lebel, T., Tinga, A., Bessemoulin, P., Brouwer, J., Dolman, A.J., Engman, E.T., Gash, J.H.C., Hoepffner, M., Kabat, P., Kerr, Y.H., Monteny, B., Prince, S., Said, F., Sellers, P., Wallace, J.S., 1994. Hapex–Sahel: a large scale study of land–atmosphere interactions in the semiarid tropics. Ann. Geophys. 12, 5364.

Gupta, H.V., Sorooshian, S., Yapo, P.O., 1998. Towards improved calibration of hydrologic models: Multiple and noncommensurable measures of information. Water Resour. Res. 34 (4), 751–763.

Kohler, M.A., Linsey, R.K., 1951. Predicting the runoff from storm rainfall, Weather Bureau, US Dept. of Commerce. Res. Paper 34, Washington, USA,.

Kuczera, G., Parent, E., 1998. Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm. J. Hydrol. 211, 69–85.

Lamachère, J.M., 1991. Aptitude au ruissellement et à l'infiltration d'un sol sableux fin après sarclage. In: Sivakumar, M.V.K., Wallace, J.S., Renard, C., Giroux, C. (Eds.), Soil Water Balance in the Sudano–Sahelian Zone, Proc. Int. Workshop, February 1991, Niamey, Niger. IAHS Publ. 199, pp. 109–119.

Le Barbé, L., Lebel, T., 1997. Rainfall climatology of the Hapex–Sahel region during the years 1950–1990. J. Hydrol. 188-189, 43–73.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models; Part I—a discussion of principles. J. Hydrol. 10, 282–290.

Peugeot, C., Esteves, M., Galle, S., Rajot, J.L., Vandervaere, J.P., 1997. Runoff generation processes: results and analysis of field data collected at the East Central SuperSite of the Hapex–Sahel experiment. J. Hydrol. 188–189, 179–202.

Peugeot, C., Cappelaere, B., Vieux, B.E., Séguis, L., Maia, A. 2003 Hydrologic process simulation of a semiarid, endoreic catchment in Sahelian West Niger, Africa: 1. Model-aided data analysis and screening. J. Hydrol.

Prechelt, L., 1998. Automatic early stopping using cross validation: quantifying the criteria. Neural Networks 11 (4), 761–767.

Séguis, L., Cappelaere, B., Peugeot, C., Vieux, B., 2002. Impact on Sahelian runoff of stochastic and elevation-induced spatial distributions of soil parameters. Hydrol. Process. 16, 313–332.

Séguis, L., Milési, G., Cappelaere, B., Peugeot, C., Massuel, S., Favreau, G. Assessing the impacts of climate and land-clearing on runoff in a small Sahelian catchment (Southwest Niger). Hydrol. Process., submitted for publication

Sorooshian, S., Gupta, H.V., Fulton, J.L., 1983. Evaluation of maximum likelihood parameter estimation techniques for conceptual rainfall–runoff models: influence of calibration data variability and length on model credibility. Water Resour. Res. 19 (1), 251–259.

Stone, M., 1974. Cross-validation choice and assessment of statistical predictions. J. Roy. Statist. Soc. B-36, 111–147.

Stroosnijder, L., Hoogmoed, W.B., 1984. Crust formation on sandy soils in the Sahel. II. Tillage and its effect on the water balance. Soil Tillage Res. 4, 321–331.

Tukey, J.W., 1977. Exploratory data analysis, Addison-Wesley, Reading, MA, USA.

USACE, 1993. *Grass 4.1* user's reference manual, U.S. Army Corps of Engineers Construction Engineering Research Laboratories, Champaign, IL, USA, 556 pp.

Vieux, B.E., 2001. Distributed hydrologic modeling using GIS, Water Science and Technology Library, vol. 38, Kluwer Academic, 293 pp.

Vieux, B.E., Gaur, N., 1994. Finite-element modeling of storm water runoff using *Grass* GIS. Microcomput. Civil Engng 9 (4), 263–270.

Vieux, B.E., Looper, J.P., Cappelaere, B., Peugeot, C., Maia, A., 1998. Climatic impacts on water resources in West Niger, Proceedings of the International Conference Water resources variability in Africa during the XXth Century, Abidjan, Nov. 1998. AISH Publ. 252, pp. 347–354.

Woolhiser, D.A., Smith, R.E., Giraldez, J.V., 1996. Effects of spatial variability of saturated hydraulic conductivity on Hortonian overland flow. Water Resour. Res. 32 (3), 671–678.