Parameter estimation in distributed hydrological modelling: comparison of global and local optimisation techniques

R.-S. Blasone[1]*, H. Madsen[2] and Dan Rosbjerg[1]

[1]Institute of Environment and Resources, Technical University of Denmark, Bygningstorvet 115, DK-2800 Kongens Lyngby, Denmark. *Corresponding author. E-mail: *rsb@er.dtu.dk*

[2]DHI Water - Environment - Health, DK-2970 Hørsholm, Denmark

Received 18 December 2006; accepted in revised form 5 September 2007

Abstract Much research has been spent in the last three decades in developing more effective and efficient automatic calibration procedures and in demonstrating their applicability to hydrological problems. Several problems have emerged when applying these procedures to calibration of conceptual rainfall–runoff and groundwater (GW) models, such as computational time, large number of calibration parameters, parameter identifiability, model response surface complexity, handling of multiple objectives and parameter equifinality. All these are expected to be much more severe for more complex models, for which comprehensive calibration studies have not so far been conducted. The scope of this paper is to investigate the performance of a global and a local optimisation technique, respectively, the Shuffled Complex Evolution algorithm and the gradient-based Gauss–Marquard–Levenberg algorithm, in calibration of physically based distributed models of different complexity. The models considered are a steady-state GW model, a transient GW model and a fully integrated model of the same catchment. The calibration is conducted in a multi-objective framework where two different aspects of the model response, the simulated runoff and the groundwater elevation are aggregated and simultaneously optimised. Different aggregated objective functions are used to give different weights to the calibration criteria. The results of the calibration procedures are compared in terms of effectiveness and efficiency and demonstrate the different performance of the methods. Moreover, a combination of the global and local techniques is investigated as an attempt to exploit the advantages of both procedures, while overcoming their drawbacks.

Keywords Distributed hydrological models; global and local optimisation methods; multi-objective calibration; parameter estimation

Introduction and scope

Traditionally, calibration of hydrological models has been performed manually by trial-and-error parameter adjustment. The process of manual calibration, however, requires a high degree of expert knowledge of the model and the system and is characterized by subjectivity in the strategy employed to adjust the parameter values, as well in the criteria (mainly visual) used to judge the goodness-of-fit of the model simulation. Moreover, manual calibration is a very tedious and time-consuming task (Boyle *et al.* 2000; Madsen *et al.* 2002). To overcome these problems much research has been spent in the last three decades in developing more effective and efficient automatic calibration procedures and in demonstrating their applicability to hydrological problems. Nowadays manual calibration is often substituted or supplemented by automatic procedures, which have found widespread use in hydrology. The main advantages of automatic techniques are the speeding up of the calibration process, the reduced subjectivity involved in the calibration procedure (for most of the techniques only

Nordic Hydrology **Vol 38 No 4–5 pp 451–476** © IWA Publishing 2007

doi: 10.2166/nh.2007.024

the parameter range needs to be specified by the modeller) and the availability of generic software that can be relatively easily linked to the model.

Automatic calibration procedures can be classified as either local or global search strategies (Sorooshian and Gupta 1995), depending on the evolving pattern of the solutions, i.e. the parameter sets. Local methods involve the estimate of a single parameter set following deterministic rules. These procedures direct the search in the parameter space by using only information on the objective function value (direct methods) or by including also derivative information (gradient-based methods). Global methods involve a population of solutions by using both deterministic and stochastic rules. Initially, the global procedures broadly search the parameter space and subsequently gradually converge into the subregion(s) containing the objective function's optima. The first applications of automatic calibration procedures to hydrological models were based on local, deterministic, gradient-based methodologies (Dawdy and O'Donnell 1965; Nash and Sutcliffe 1970; Johnston and Pilgrim 1976; Carrera and Neuman 1986). Since then, calibration procedures have evolved significantly, resulting in the development of robust and effective global search procedures. Among these, different population-evolution-based optimisation algorithms have been applied in hydrology, such as Genetic Algorithms (Wang 1991), the Shuffled Complex Evolution (SCE) algorithm (Duan *et al.* 1992), and the Simulated Annealing (Sumner *et al.* 1997).

Despite the various automatic calibration techniques available, their application is limited by computational constraints. In particular, the relationship between the computational time required to run the model and the efficiency of the search algorithm, i.e. the number of model runs required to calibrate the model, still remains one of the main constraints in the choice of the calibration methodology for the particular problem at hand. Because of their fast running time, lumped, conceptual rainfall–runoff (RR) models have been widely employed to test new calibration methods (Duan *et al.* 1992; Gupta *et al.* 1998; Vrugt *et al.* 2003*b*) as well as to investigate the application of existing methodologies to hydrological problems (Wang 1991; Franchini and Galeati 1997). They have also been used in several comparative studies to investigate the performance of different calibration techniques, mainly global methodologies (Gan and Biftu 1996; Cooper *et al.* 1997; Kuczera 1997; Franchini *et al.* 1998; Thyer *et al.* 1999; Madsen *et al.* 2002; Marshall *et al.* 2004; Skahill and Doherty 2006). Due to the higher computational cost of distributed groundwater (GW) models, faster calibration methods, mainly local procedures, based on nonlinear regression techniques, have been applied to these models (Carrera and Neumann 1986; Bentley 1993; Poeter and Hill 1997; Christensen and Cooley 1999; D'Agnese *et al.* 1999). Only a few studies have been conducted on GW models to compare the performance of different calibration techniques (Heidari and Manjithan 1998; Solomatine *et al.* 1999; Shigidi and Garcia 2003).

Compared to RR and GW models, fewer calibration attempts are reported for distributed and integrated models, which couple groundwater and surface water processes (Eckhardt and Arnold 2001; Madsen 2003; Mertens *et al.* 2004; Muleta and Nicklow 2005). There are several reasons why hydrologists rarely tackle the problem of calibration of these more complex models. One of these is the already mentioned concern about computational time, which grows substantially for these types of models. Another reason is the huge number of parameters to calibrate, due to the additional model complexity and the necessity to account for the spatial distribution of the hydrological properties within the catchment. Unless a proper procedure is applied to reduce the dimension of the problem, the higher dimensionality of the parameter space increases dramatically the number of model runs required by the calibration procedure. Moreover, the presence of parameter correlation and interaction, often encountered in hydrological models, can be even worse for integrated,

distributed models. This causes poor parameter identifiability, which might reduce the performance of search algorithms, or, in the worst cases, make the search unsuccessful (Duan *et al.* 1992; Poeter and Hill 1997; Kuczera 1997).

In addition, calibration of distributed, integrated models suffers from all the problems that have been encountered so far in parameter estimation applications in hydrology. Among these, the nature of the model response surface must be emphasised, which is characterized by roughness, non-smoothness, discontinuous derivatives, several regions of attraction with numerous local optima and a non-convex shape in the proximity of the optima. All these features, which have emerged from several studies conducted on lumped, conceptual RR models (Duan *et al.* 1992; Kuczera 1997; Skahill and Doherty 2006), are expected to be even more severe for complex, distributed and integrated models, which are characterized by a higher degree of non-linearity in the relationship between parameters and model responses.

Another feature of calibration of hydrological models is its intrinsic multi-objective nature (Gupta *et al.* 1998; Madsen 2000, 2003); in fact, in many applications, it is necessary to calibrate the model to more than one simulated response in order to obtain a hydrologically sound model. In general, multiple variables (e.g. groundwater elevation and river discharge), multiple response modes (e.g. in the case of river discharge, high flows, low flows and overall water balance) and multiple sites data (e.g. runoff at different river sections) may be included in the calibration process (Madsen 2003). This has the advantage of improving the identifiability of the estimated parameters, if the selected objectives contain new and independent information on the system (McLaughlin and Townley 1996; Gupta *et al.* 1998; Kuczera and Mroczkowski 1998). The multi-objective problem can be reduced to a single-objective problem by combining the targets to be optimised into an aggregated objective function. This can easily be done, for example, through a weighted sum of the different objective functions (Gupta *et al.* 1998), where the weights reflect the importance given by the modeller to the various aspects of the system behaviour. Other less subjective methods in aggregating different criteria have also been proposed, including the probability weighted method by van Griensven and Bauwens (2003) and the common distance scale method by Madsen (2003).

Even though the multi-objective optimisation can be reduced to a single-objective problem, the solution to the calibration problem is not expected to be unique. The inability to find a single solution and the existence of multiple solutions for a given model set-up are major issues in the calibration of hydrological models, which have been denoted "equifinality" by Beven and Binley (1992) and "multi-objective equivalence of parameter sets" by Gupta *et al.* (1998). The first definition is given in a statistical sense and it accounts for the probabilistic representation of parameter (model) uncertainty. The second one considers that, due to the different ways in which the best fit of a model to the data can be defined in a multi-objective optimisation context, the result of the calibration will be an ensemble of optimal solutions. Under the latter perspective, the performance of the solutions must be evaluated considering the trade-offs among the different calibration criteria. A simple and comprehensive way of doing so is by means of the Pareto criterion (Pareto 1906), which has been used in hydrology in several studies addressing multi-objective calibration (Gupta *et al.* 1998; Yapo *et al.* 1998; Madsen 2003; Vrugt *et al.* 2003*a*).

The scope of this paper is to investigate the performance of two different optimisation techniques, a global and a local methodology, in the calibration of distributed hydrological models. The global methodology is the Shuffled Complex Evolution (SCE) algorithm developed by Duan *et al.* (1992), while the local one is the gradient-based Gauss–Marquard–Levenberg (GML) algorithm, as implemented in the PEST software by Doherty (2005). The models considered are three physically based distributed models of different complexity: a steady-state GW model, a transient GW model and a fully integrated model of

the same catchment. The integrated MIKE SHE modelling system (Graham and Butts 2006) is used to model the catchment of the Danish river Karup, a system mainly dominated by infiltration and groundwater processes. The calibration is conducted in a multi-objective framework where two different aspects of the model response are aggregated and simultaneously optimised. These are the simulated runoff at the catchment outlet and the groundwater elevation at 17 different locations (monitoring wells). Six different aggregated objective functions are used to give different weights to the two calibration criteria. The results of applying the two optimisation methods to the various models are compared in terms of effectiveness and efficiency of the procedures. In this respect, the Pareto criterion is also used as a way of discriminating among dominant (optimal) and non-dominant (sub-optimal) solutions. We are aware that using different combinations of weights is not an efficient method to explore the Pareto front, for which other methods are better suited (Gupta *et al.* 2003), but the scope of this study is not to perform Pareto optimisation. The main focus of this paper is on comparing the calibrated results by use of SCE and PEST in a multi-objective framework, where the existence of multi-objective equivalence of parameter sets is dealt with by applying the Pareto criterion. The outcome of the research enlightens us to the different advantages and disadvantages of the two optimisation methods and the applicability of these techniques to the different models used in this study.

Moreover, a combination of the global and local techniques is investigated as an attempt to exploit the advantages of both methods, while overcoming their drawbacks. In this respect, the performance of two different ways of merging the global and local methodologies is analysed and compared. Previous trials to develop hybrid, two-stage optimisation procedures based on merging global and local methods have already been made with RR models (Franchini and Galeati 1997; Kuczera 1997) and with GW models (Heidari and Ranjithan 1998). Apart from these works, according to the knowledge of the authors, no attempts to couple global and local optimisation methodologies have so far been conducted with distributed, integrated models.

This paper is structured as follows. First, the case study, Karup catchment, is presented together with the three different hydrological models set up for this watershed and their respective calibration parameters. Then the two calibration techniques used in the study, SCE and PEST, are described together with the strategy applied for their implementation. Next follows the calibration results found by SCE and PEST for the three different models, when different objective functions are applied. The results are analysed according to parameter and objective function convergence. Subsequently, some preliminary results are presented showing how the global and local optimisation techniques can be merged to improve the efficiency and effectiveness of model calibration. Finally, a summary and an overall discussion of the findings from this work are provided.

Case study

The catchment

The catchment considered in this study is that of the Karup River, located in the western part of Denmark, which has been the object of previous studies by Refsgaard (1997) and Madsen (2003). The watershed has an area of about 440 km^2 and it is drained by the Karup River and about 20 tributaries. The catchment presents a quite homogeneous geology, characterised by sandy soils with high permeability. The aquifer is mainly unconfined and has a thickness varying between 90 m at the upstream part of the catchment and 10 m in the western and central areas. The geology of the catchment is defined by five main soil types, identified by the Danish National Water Resources model (Henriksen *et al.* 2003) and shown in Table 1. The data used in this study come from the comprehensive hydrological database available for the Karup catchment and previously employed by Refsgaard (1997). Rainfall from nine

Table 1 Soil types for the saturated zone

Soil code	Soil name	Description
1	Melt water sand	Quarternary and Post-Glacial sand and gravel
2	Clay	Glacial, Inter-Glacial and Post-Glacial clay and silt
3	Quartz sand	Miocene, medium to coarse grained sand and gravel
4	Mica sand	Miocene, fine to medium grained sand
5	Mica clay/silt	Pre-Quarternary clay and silt

stations (daily values), runoff at the river outlet (daily values), groundwater elevation data from 17 wells (recorded every 15 d) and temperature (daily values) are used in this work. Figure 1 shows the locations of the wells and the river outlet station in the Karup catchment. A 6-year period of data, from 1 January 1969 to 1 January 1975, is used to set up and calibrate the models.

Model conceptualizations and calibration parameters

The hydrological modelling software used in this study is the MIKE SHE modelling system (Graham and Butts 2005), a spatially distributed, physically based model, which has been used to set up a steady-state GW model, a transient GW model and a fully integrated model of the Karup catchment. The flexibility of the MIKE SHE modelling tool allows us to easily change the model formulation and to include in the description of the system the required degree of complexity. The number of parameters contained in the model and potentially subjected to calibration is huge and increases with model complexity. According to the "principle of parsimony" (Hill 1998), the calibration problem is better posed if its dimensionality is limited and, at the same time, the estimated parameters are sufficient to guarantee a satisfactory model fit. Thus, only a few parameters should be chosen for calibration. The selection of the parameters to estimate is made based on the results of a

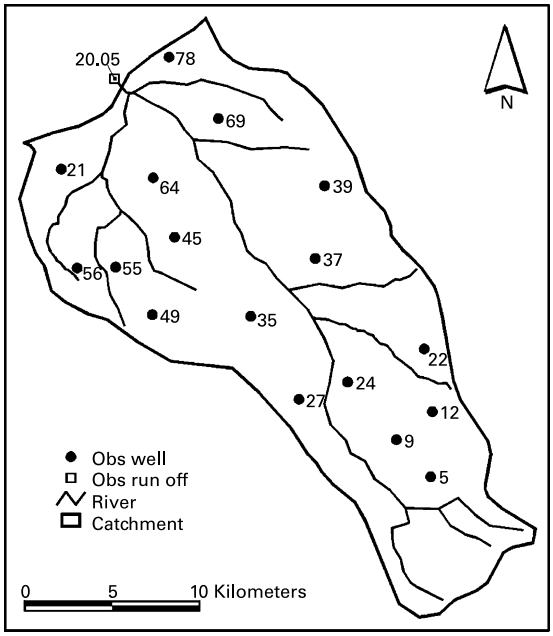

Figure 1 Karup catchment. Location of discharge gauging station and wells

previous extensive sensitivity analysis study (Dornes 2003), which was conducted on the three models of the Karup catchment used in this work. The parameters that were found to have the most pronounced effect on the model response are calibrated, while the others are fixed to their previously manually calibrated values. The conceptualisation of the steady-state GW model, the transient GW model and the integrated model is described in the following. Despite the conceptual differences among the models, some of the MIKE SHE modelling blocks are common to the set-up of the three different conceptualizations.

The horizontal computational grid of all the three models is defined with a spatial scale of 1 km × 1 km, and the geology of the saturated zone is represented with vertical and horizontal scales of 10 m and 1 km × 1 km, respectively. Each grid element is assigned one of the soil types specified in Table 1, identified by a given code and specific hydrogeological parameters. A two-dimensional GW model is applied and hence one computational layer is defined for the saturated zone by vertical weighting of the soil conductance in each horizontal grid element. The hydraulic conductivities of two soil types, melt water sand and quartz sand, are those with the largest impact on the model response and thus they are subjected to calibration. A constant anisotropy between horizontal and vertical hydraulic conductivity is assumed and only the horizontal conductivity is calibrated, *Kh*, while the vertical one, *Kv*, is set equal to one-tenth of the respective horizontal conductivity.

Surface runoff is routed down-gradient towards the river system using the diffusive wave approximation of the Saint Venant equations (two-dimensional Saint Venant equations). For the steady-state and transient GW models, overland flow is only generated when the groundwater level rises and reaches the surface level. For the integrated model, surface runoff is generated according to evaporation and infiltration processes along the flow path. Moreover, for the latter model, an additional contribution to overland flow is generated if there is saturation of the first centimetres of the soil. The drainage system defined within the catchment includes both natural and artificial drainage. The drainage flow is simulated using a linear reservoir description, which for each cell requires a drainage level, *DrnLev*, and a time constant (drainage coefficient), *DrnCoef*, that regulate how much and how fast water is drained. Both of these parameters are assumed uniformly distributed in the catchment and are subjected to calibration. The river system acts as a collector of both overland and saturated zone flow. The river flow is routed using a Muskingum routing scheme. The interaction between river and saturated zone is accounted for by a leakage coefficient, *LeakCoef*, which is assumed to be constant for all river branches and is subject to calibration.

For the steady-state and transient GW models an empirical root zone model (Henriksen *et al.* 2003) is applied to calculate the recharge to MIKE SHE. According to this method the recharge is calculated based on observed precipitation, potential evapotranspiration, land use information and estimates of the field capacity. The transient GW model has the same parametrisation as the steady-state model, but the processes are modelled in transient mode. A more detailed description of the MIKE SHE steady-state and transient GW model formulation can be found in Sonnenborg *et al.* (2003). The fully integrated model includes a conceptualization of the unsaturated zone by means of Richards' equation. The soil moisture retention curve (the relationship between the water content and the matric potential) and the hydraulic conductivity function are modelled using the van Genuchten equations. Four different soil types are identified in the unsaturated zone (loamy sand, fine sand, fine sand in heath areas and coarse sand), each characterised by its own specific van Genuchten parameters. The parameters of only two soil types, loamy sand and fine sand, are those that affect the model response the most and are therefore subjected to calibration. These are the saturated hydraulic conductivity, *Ks*, and the three empirical constants appearing in the van Genuchten formulas, α, *M* and *N*, describing the soil moisture retention curve and hydraulic conductivity function. The relationship linking two of these variables, *N* and *M*, allows

reducing the problem dimensionality by tying the value of M to N, thus only the last undergoes calibration. In the integrated model the actual evapotranspiration and the actual soil moisture status in the root zone are modelled according to the empirically derived equations developed by Kristensen and Jensen (1975), which account for the characteristics and the distribution of the different vegetation types present in the catchment (canopy, root distribution and vegetation density). The vegetation types present in the catchment are classified according to the following land use and vegetation typologies: agriculture (57%), forest (18%), heath (7%) and wetland (18%). The snow accumulation and snowmelt processes are also accounted for in the integrated model using a simple degree–day approach.

The parameters of each model included in the calibration are listed in Table 2. Those assumed to depend on the calibrated values of other parameters are indicated as "tied" parameters and the specified equations are also shown in Table 2.

Model calibration procedures

SCE and PEST optimisation algorithms

As previously mentioned, automatic calibration procedures can be grouped into local and global strategies. It has become common practice in hydrology to employ local methods for calibration of GW models, while global procedures are mainly used for RR models. As for distributed, integrated models, the increased complexity of the model response might suggest the failure of standard local calibration methods, mainly resulting in the inability to find the global optimum and the subsequent convergence to suboptimal regions of the parameter space. On the other side, local methods require a lower number of model simulations, which is a great advantage considering that the computational time of a model run of a distributed integrated model can be relatively high. The local and global techniques compared in this study are two procedures that are frequently applied in calibration of hydrological models. The global search methodology used is the Shuffled Complex Evolution (SCE) algorithm developed by Duan *et al.* (1992) and the local method is the Gauss–Marquardt–Levenberg (GML) non-linear regression method, as implemented in the PEST software by Doherty (2005).

The SCE algorithm is an evolutionary-based procedure that simultaneously evolves a population of solutions (parameter sets) towards better solutions in the search space, trying to converge to the global optimum of the objective function. The procedure starts with a random generation of an initial population of solutions in the feasible parameter space and evaluation of the objective function for each individual solution. The parameter sets are then divided into a certain number (defined by the modeller) of subgroups, called complexes, ensuring a similar quality of the solutions inside each subgroup. The solutions in each complex are evolved according to the simplex algorithm (Nelder and Mead 1965) in the attempt to substitute the worst solutions with better ones. After this phase, the solutions from the complexes are shuffled into a new population, from which new complexes are formed (still guaranteeing for each complex a similar quality of the parameter sets) and evolved as before. The "evolution of complexes" and "shuffling" steps are repeated until a convergence criterion is satisfied (maximum number of model runs or minimum relative change of parameter or objective function values between consecutives evolution steps). More details on the SCE algorithm can be found in Duan *et al.* (1992). The SCE algorithm has been extensively used in hydrology (e.g. Duan *et al.* 1992; Lee and Wang 1998; Hsieh and Wang 1999; Madsen 2003; Mertens *et al.* 2004) and, in the case of RR models, it has also been proved to be superior to other global search techniques, such as the Multiple Start Simplex, Genetic Algorithms and Simulated Annealing (Gan and Bitfu 1996; Cooper *et al.* 1997; Kuczera 1997; Franchini *et al.* 1998). Some extensions to the main SCE algorithm have been

Table 2 Parameters subject to calibration and tied parameters. Relationships between tied and calibrated parameters: (*) $Kv = 0.1Kh$; (**) $M = 1 - 1/N$

				Models		
Model blocks	**Parameter**	**Range**	**Units**	**Steady-state**	**Transient**	**Integrated**
Saturated zone (hydraulic conductivity)	Horizontal, soil 1, *Kh1*	$5 \times 10^{-5} - 5 \times 10^{-3}$	[m/s]	Calibrated	Calibrated	Calibrated
	Vertical, soil 1, *Kv1*	$5 \times 10^{-6} - 5 \times 10^{-4}$	[m/s]	tied*	tied*	tied*
	Horizontal, soil 3, *Kh3*	$1 \times 10^{-4} - 1 \times 10^{-2}$	[m/s]	Calibrated	Calibrated	Calibrated
	Vertical soil 3, *Kv3*	$1 \times 10^{-5} - 1 \times 10^{-3}$	[m/s]	tied*	tied*	tied*
Drainage	Drainage level, *DrnLev*	$-1.3 - -8$	[m]	Calibrated	Calibrated	Calibrated
	Drainage constant, *DrnCoef*	$1 \times 10^{-8} - 1 \times 10^{-6}$	[s^{-1}]	Calibrated	Calibrated	Calibrated
River-aquifer interaction	Leakage coefficient, *LeakCoef*	$1 \times 10^{-8} - 1 \times 10^{-6}$	[s^{-1}]	Calibrated	Calibrated	Calibrated
Unsaturated zone (van Genuchten parameters)	α, soil 1, $\alpha 1$	0.05–0.5	[cm^{-1}]			Calibrated
	M, soil 1, *M1*	0.6–1.67	[–]			tied**
	N, soil 1, *N1*	1.2–2.5	[–]			Calibrated
	Ks, soil 1, *Ks1*	$1 \times 10^{-6} - 1 \times 10^{-4}$	[m/s]			Calibrated
	α, soil 2, $\alpha 2$	0.01–0.1	[cm^{-1}]			Calibrated
	M, soil 2, *M2*	0.6–1.67	[–]			tied**
	N, soil 2, *N2*	1.2–2.5	[–]			Calibrated
	Ks, soil 2, *Ks2*	$5 \times 10^{-5} - 5 \times 10^{-3}$	[m/s]			Calibrated

proposed. Yapo *et al.* (1998) expanded the procedure to multi-objective optimisation and Vrugt *et al.* (2003*b*) modified the algorithm, incorporating an adaptive Markov chain Monte Carlo procedure to provide estimates of parameter and model response uncertainty as part of the optimisation. The SCE version used in this study is the one implemented in the AUTOCAL software (DHI 2004).

The GLM method is a recursive gradient-based optimisation strategy that combines the Gauss–Newton algorithm and the method of gradient descent to provide a solution to the mathematical problem of minimising a sum of squared deviations between model outputs and corresponding observations. In the basic gradient method a new parameter estimate is found from the current set by shifting the actual parameters of a given amount along the direction of the maximum improvement of the objective function. The Levenberg–Marquardt variation of the method introduces a correction in the direction and length of the parameter upgrade vector in order to avoid the search being trapped in the proximity of the optimum without reaching it, as typically occurs when the optimum is located in elongated valleys of the objective function (hemstitching phenomenon). The software implementing the GML method used in this study is PEST, the parameter estimation package developed by Doherty (2005). Due to the availability of free software packages like PEST and UCODE (Poeter and Hill 1998), local procedures for nonlinear parameter estimation have been widely applied in the past few years in groundwater hydrology. PEST has been applied to several hydrological studies, mainly in calibration of GW models (see, among others, Zyvoloski *et al.* (2003) and Moore and Doherty (2006)), to a MIKE SHE distributed model of the saturated zone (Islam *et al.* 2006), as well as to model sorption and degradation processes of pesticides (Dubus *et al.* 2004). The main advantages and drawbacks of the PEST procedure, compared to a global estimation method like SCE, are enlightened by Skahill and Doherty (2006). PEST is particularly efficient in terms of the number of model runs; moreover, there is the possibility to modify the procedure in cases where potential numerical problems emerge due to parameter insensitivity as well as correlation. The main problems of PEST are those common to most of the gradient-based methods: the possibility of the procedure being trapped in local objective function minima as well as the dependence of the optimisation result on the point from which the procedure is started. To overcome the latter, Poeter and Hill (1997) recommended restarting the local search procedure many times from different sets of starting values, in order to check whether the calibrated result is the global solution. Based on the same idea, Skahill and Doherty (2006) have further developed the PEST algorithm by coupling the calibration routine with a multi-start method to conduct multiple consecutive calibration runs. This scheme is implemented in a way which ensures the coverage of the parts of the parameter space that have not been visited by the previous model runs. The advantage of this feature is that of preventing PEST to converge to local objective function optima.

In this work, a multi-starting PEST calibration is conducted, with the purposes of overcoming the problem of convergence to local optima, getting more insight into how the starting point affects the result of the search and obtaining more than one solution for comparison with the SCE results.

Initialisation of the procedures and termination criteria

The initialisation of the SCE algorithm requires the generation of an initial population of parameter sets, which are then evolved according to the SCE algorithm. These points are here used as initial points for N_0 independent PEST runs. The size of the population, N_0, is determined by the algorithmic parameters of SCE. In this study three complexes are used, with the number of points in each complex depending on the number of calibrated parameters as recommended by Duan *et al.* (1994), resulting in $N_0 = 33$ for the GW models

and $N_0 = 69$ for the integrated one. The initial population of parameter sets is generated using the Latin Hypercube Sampling (LHS) approach (McKay *et al.* 1979) in order to guarantee a good coverage of the parameter space. The number of complexes is the key parameter of the SCE algorithm which trades-off efficiency and effectiveness. That is, more complexes reduce the chance of converging into a local optimum but at the expense of a larger number of model evaluations and vice versa. Preferably, sensitivity tests with different number of complexes should have been performed before running the calibration experiments. In particular, it is expected that more complexes are needed for the distributed integrated model, since more parameters are included in the calibration. However, due to computational time requirements, this investigation was not conducted and the number of complexes was kept the same for the three models.

The result of the local search is one single solution for each independent optimisation, thus generating a total of N_0 solutions for every combination of weights (as explained below). SCE provides a unique best solution for each specified weight combination, plus a number of near-optimal solutions with objective function values only slightly worse than the optimum. To compare the results with those of PEST, the best N_0 SCE results are considered.

The termination criteria for the local method are convergence conditions in both objective function and parameter space. The PEST stopping criteria are met if it is not possible to decrease the objective function by a relative amount of 0.005 over four successive iterations, or if the value of the parameter vector changes less than a relative amount of 0.005 over four successive iterations. The PEST routine for temporary immobilisation of insensitive parameters at their current value during a few iterations, while calibrating the others, is used in the execution of the procedure. This feature does not only allow dealing with parameter non-uniqueness, but it also conveys the advantages of preventing the search being trapped in local optima (Skahill and Doherty 2006) and of speeding up the convergence of the procedure. The SCE calibration process will end when a maximum number of model runs is reached which, based on previous calibration experiments conducted on the same case studies, is set to 700 for the GW models and 1400 for the integrated model.

Multi-objective optimisation

As previously mentioned, the calibration of the three models is conducted in a multi-objective context, i.e. different model responses, which may be distributed in space, are simultaneously optimised. In this work the objectives to optimise are defined in terms of groundwater elevation heads from 17 wells and discharge data at the catchment outlet. The fitness of the simulated model outputs versus the recorded observations is evaluated by two objective functions. These are the sum of the mean squared errors related to groundwater levels (m) at the 17 well locations, MSE^w, and the streamflow ($m^3\ s^{-1}$) at the catchment outlet, MSE^r. The two error functions are aggregated into a single objective function, F_{aggr}, by weighting them as shown in Equation (1):

$$F_{aggr}(\theta) = w^r MSE^r(\theta) + w^w \sum_{i=1}^{17} MSE_i^w(\theta) \tag{1}$$

where θ is the parameter vector, and w^r and w^w are the weights assigned to the objective functions of the river runoff and the groundwater levels, respectively. The weighting method is the simplest way to direct the search of the calibration procedure towards regions of the solution space with different performance of the two optimisation criteria. Moreover, the use of the weights allows accounting for the different scales of magnitude of the hydrological variables. Six different combinations of weights, chosen by a trial-and-error process, are used for the steady-state and the transient model, while, due to the high computational time

required, only one combination of weights is applied for the integrated model. This is chosen as the one providing the more balanced solutions according to the MSE^w and MSE^r criteria. The combinations of weights associated with the various objective functions are shown in Table 4.

Comparison of calibration results

Parameter convergence

As described above, the number of calibration parameters varies with the model complexity. As shown in Table 2, for the GW models 5 parameters are calibrated, and for the integrated model 11 parameters are calibrated. In the following the relationship between the values of a particular parameter and the objective function at the convergence stage of the search is analysed for each model formulation and each combination of weights. Only the N_0 final points of the search are considered for PEST, while the best N_0 solutions from all the parameter sets generated by SCE are used.

The locations of the calibrated parameters versus the respective aggregated objective function value are considered and compared. These regions are referred to as "convergence regions" of the parameters, despite the fact that they are not strictly such for the local procedure, for which each calibrated point is the result of an independent optimisation run. The convergence regions are analysed because the dispersion of a calibrated parameter gives an indication of the sensitivity of the model to that particular parameter. Tight convergence regions (small parameter ranges for the lowest values of the objective function) are associated with parameters that can be well determined through the calibration process, while broad convergence regions (larger parameter ranges) are typical for parameters that have only a minor effect on the model response (if varied alone). The latter can be the effect of either parameter insensitivity, high parameter correlation or both. This interpretation always holds for the results obtained from the global procedure while, in our opinion, this is not always true in the case of PEST, as shown in the following. In fact, when the local method is used, the presence of multiple convergence regions may indicate parameter insensitivity, but it can also simply be an effect of the tendency of the procedure of being trapped in local minima during the calibration iterations. This is clearly illustrated in Figure 2, which shows the convergence regions of the parameter *Kh3* of the steady-state GW model obtained by SCE and PEST for the six objective functions used. A major difference emerges from the plots when comparing the results obtained from PEST and SCE. SCE normally converges to a narrow region, while the convergence points by PEST are generally either spread all over the parameter space or localised around a few regions of attraction (some of these being suboptimal according to the chosen calibration criterion). These different convergence patterns of SCE and PEST in the parameter space are common for the three models and they hold for most of the objective functions and parameters.

In the cases where a parameter calibrated by PEST converges to a unique, narrow region, a similar behaviour for the same parameter is also obtained by SCE. Despite this, the opposite is not true; in fact, some parameters show a narrow convergence according to SCE, but not to PEST. Based on these results, the global method demonstrates the ability to identify more parameters that can be well determined than the local method.

It is difficult to judge which procedure gives true information on the sensitivity of the model to the parameters, as we are working with real case studies, affected by several error sources. In fact, it is possible that the global method over-fits the best solution and disregards parameter sets, which are closer to the unknown optimal solution, but suboptimal due to the presence of errors. On the other hand, the additional solutions by PEST, which are more spread in the parameter space, can either indicate the presence of "equifinal" solutions or

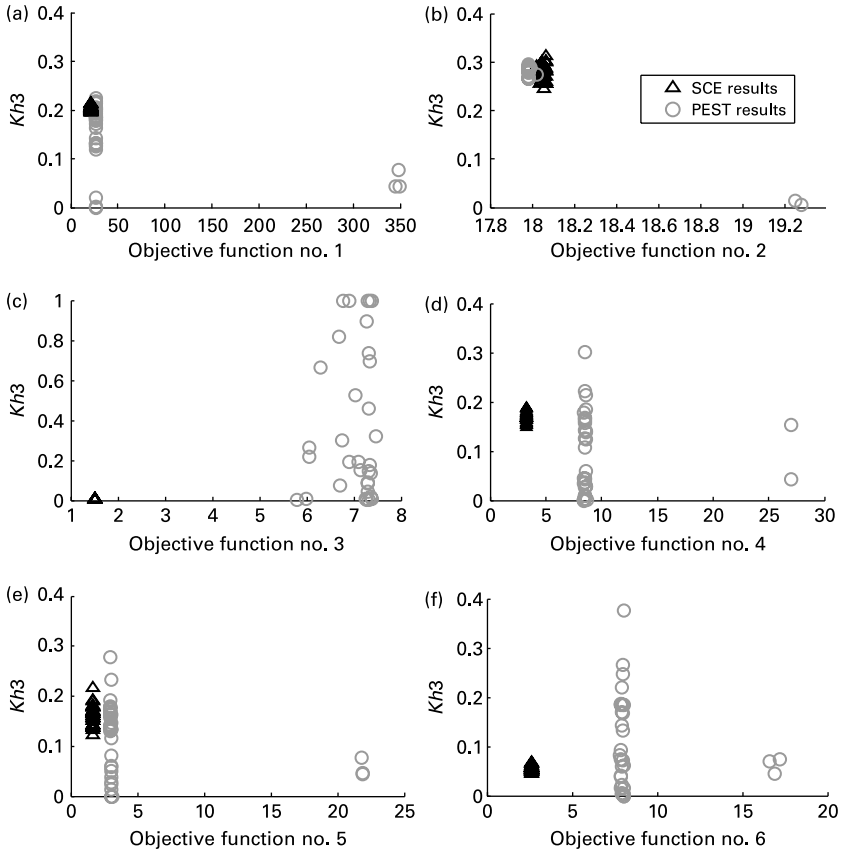

Figure 2 Convergence points of PEST and SCE executions of the normalized value of parameter *Kh3*; steady-state GW model and various objective functions used

they can be caused by the problem of the procedure to be trapped in local regions of attraction.

A factor, which deeply affects the regions of convergence in parameter space, is the choice of the objective function determined, in this case, by the particular combination of weights assigned to the river discharge and to the groundwater MSE. The impact of changing the weights appears evident in Figure 2 for the steady-state GW model (similar results are obtained for the transient GW model). The estimate of a parameter may vary when different criteria are chosen to calibrate the model. Moreover, the N_0 final results can be more or less spread in the parameter space, depending on the objective function, indicating that a parameter can be well determined according to a particular combination of the calibration criteria, while it can be non-identifiable for another combination. These variations in the parameter convergence pattern should be expected when using various objective functions. In fact, due to the different dependences of the runoff and the groundwater elevation heads on the modelled processes and the associated parameters, the objective function also has a different shape in the parameter space depending on the applied weights. No tendencies of particular objective function combinations in generating narrower or more spread parameter convergence are found, the results being mainly dependent on the model and the particular parameter considered.

The model conceptualisation is also one of the factors affecting the convergence patterns of the parameters. This is evident when comparing, for the same parameter and the same

objective function, the convergence points of the calibration obtained for the various models. Figure 3 shows the convergence regions of two parameters common to the three models, *Kh3* and *LeakCoef*, when calibrated using the same objective function (objective function no. 5), as their behaviour is representative of most of the parameters/cases. The plots in Figure 3 demonstrate that a particular parameter can converge to a different region when a different model is used, even if the same objective function is optimised. The general tendency of the parameter convergence intervals is that of getting narrower when the model complexity increases from the steady-state to the transient GW model. This indicates that the parameters are more identifiable when the dynamic behaviour of the GW model is calibrated, rather than just the steady-state behaviour. The convergence trends obtained for the integrated model can be compared to the results of the other models only for the parameters common to the three model conceptualisations. The convergence intervals of some of these parameters are also quite large (Figure 3(c)). Such poor parameter identifiability can be caused by the fact that additional complexity and more calibration parameters are included in the integrated model without adding new data into the estimation process.

These convergence trends are common to both optimisation methods. The SCE results are more sensitive to the model variation, while, as mentioned earlier, the convergence points from PEST are generally quite spread in the parameter space for all the models and objective functions used.

Parameter correlations

The correlation among the calibrated parameters of each optimisation experiment is also investigated to check the differences between SCE and PEST results and to check whether there are trends when varying model complexity or objective function. The correlation coefficients $\rho(i,j)$ are calculated for each couple of parameters (θ_i, θ_j) as

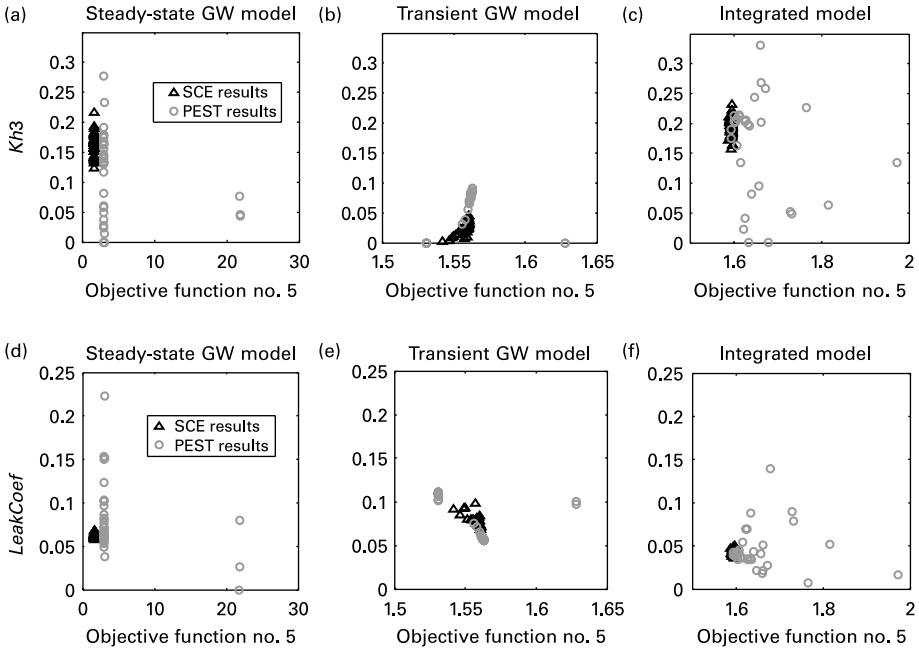

Figure 3 Convergence points of PEST and SCE executions of the normalized parameters *Kh3*: (a), (b) and (c); and *LeakCoef*: (d), (e) and (f); objective function no. 5; steady-state GW model, transient GW model and integrated model

$$\rho(i,j) = \frac{\mathrm{cov}\,(\theta_i, \theta_j)}{\sigma(\theta_i)\sigma(\theta_j)} \tag{2}$$

where $\mathrm{cov}(\theta_i, \theta_j)$ is the covariance between the parameters θ_i and θ_j and $\sigma(\theta_i)$ and $\sigma(\theta_j)$ are the standard deviations of θ_i and θ_j, respectively. All these quantities are calculated based on the ensemble of N_0 solutions found by each method. The calculated $\rho(i,j)$ are global correlation values that should not be confused with the parameter correlation at the convergence point (as calculated as part of the GML optimisation).

Very different correlation patterns among the parameters are found for the SCE and PEST solutions for all the models and the various objective functions used. Large differences are seen with respect to the pairs of parameters with high correlation as well as the absolute value and the sign of the correlation coefficients. For both GW models and the various objective functions considered, the correlation in the parameter estimates is larger for the set of PEST solutions than for those from SCE. Table 3 shows the correlation matrices obtained for SCE and PEST for the transient GW model and the objective function including only the *MSE* of the groundwater levels as calibration criterion (objective function no. 2). Similar results are obtained in the other cases. In general, for the different models and objective functions considered, few pairs of parameters have an absolute value of the correlation coefficient above 0.75 and only very few have absolute correlation values above 0.90. For both calibration procedures, the model with the largest correlation of estimated parameters is the transient GW model. In the case of the integrated model, lower correlation coefficients are found and this is true for both SCE and PEST results. Only *DrnLev* and *DrnCoef* estimates from the SCE results have a large correlation, around 0.86, while the correlation of the same parameters from the PEST results is 0.55. Moreover, based on our results, the parameter correlation is found to be affected not only by the calibration procedure and the model conceptualisation employed, but also by the objective function. This is because giving more weight to the groundwater or to the surface water processes affects the performance of different parts of the model and therefore it also affects the estimated values of the parameters related to each modelling component. For example, for the PEST results, higher correlations among the parameters is found when only the groundwater output is optimised (objective function no. 2) and this is more evident for the steady-state GW model. Including different kinds of data in the calibration process (recommended, among others, by Hill (1998)) is, in this case, achieved by adding also the observed discharge in the optimisation

Table 3 Parameter correlation matrices; transient GW model; objective function no. 2

		Kh1	***Kh3***	***DrnLev***	***DrnCoef***	***LeakCoef***
PEST						
	Kh1	1	0.65	0.71	−0.39	−0.72
	Kh3	0.65	1	0.96	0.13	−0.95
	DrnLev	0.71	0.96	1	−0.13	−0.85
	DrnCoef	−0.39	0.13	−0.13	1	−0.24
	LeakCoef	−0.72	−0.95	−0.85	−0.24	1
SCE						
	Kh1	1	−0.06	0.27	−0.05	−0.13
	Kh3	−0.06	1	0.71	0.04	−0.42
	DrnLev	0.27	0.71	1	0.05	−0.03
	DrnCoef	−0.05	0.04	0.05	1	0.16
	LeakCoef	−0.13	−0.42	−0.03	0.16	1

objectives and it is found to be an effective way of reducing parameter correlation of the PEST solutions.

The parameters of GW and distributed models have often been found to be highly correlated, a feature which represents an obstacle for effective and efficient parameter estimation (Poeter and Hill 1997; Kuczera 1997). High correlation coefficients (i.e. above 0.75) are rarely found in our case studies. Moreover, it seems that the degree of correlation among the final estimates of the parameters is more the product of the estimation procedure employed, rather than being an intrinsic feature of the particular system/model. In fact, if the correlation structure of the parameters was an intrinsic property of the model, the same correlation pattern should be detected by both procedures. The results found in this study by the two procedures differ significantly in this respect. SCE provides multiple solutions with lower correlation than the local method, thus proving that a global optimisation procedure can better deal with the problem of high correlation of the parameter estimates. The larger correlation found for the parameters estimated by PEST might be the effect of particular features of the local procedure itself, rather than a property of the models. One of these is the already mentioned problem of the procedure being attracted towards particular regions in the objective function space.

Efficiency and effectiveness of the procedures

The calibration results have so far been considered in the parameter space. In this part the solutions found by PEST and SCE are compared in terms of effectiveness and efficiency. The optimal parameter set is always unknown when calibrating hydrological models on real systems and it is never exactly known whether the calibration procedure has reached the global optimum. In this analysis, the global minimum of the aggregated objective function of each experiment is assumed to be the smallest value found among all the calibration runs by SCE and PEST. In reality, due to the presence of model and data error, the location of the real global optimum in the proximity of the best solution can only be supposed, depending on the degree of belief of the modeller on the applied calibration procedure. In this study, all the calibrations that have converged nearby the assumed global optimum are considered successful. According to this characterisation of convergence, effectiveness is defined as the probability of a calibration procedure of converging to the region close to the global optimum, estimated as the percentage of successful calibrations out of the total of N_0 provided by the two procedures.

The optimal region is defined in terms of objective function change by a maximum deviation of 0.5% from the estimated global optimum. Efficiency is defined accordingly as the number of model runs needed by the calibration procedure to converge, when convergence occurs in the optimal region. For SCE, the computational cost is represented by the interval of number of model runs required for the first and the last point of the population to converge to the optimal region. For PEST, efficiency is given by the interval between the minimum and the maximum number of model runs needed by a single execution to locate an optimal solution. Table 4 summarises the results obtained by the two procedures in terms of effectiveness and efficiency. It can be noticed that SCE is more effective for the steady-state GW model and for the integrated model, while PEST is superior for the transient GW model. In the case of the steady-state GW model, the effectiveness of SCE in converging to the region where the global optimum is located is 100% for 5 out of the 6 objective functions considered, while PEST reaches the optimal region only for objective function no. 2 where only the well elevation heads are optimised. For the transient GW model, PEST always finds optimal solutions for the considered objective functions, but the effectiveness is low, ranging between 12.9 and 58.1%. SCE converges to the optimal region only in two cases, also with a low convergence probability. For the integrated model 79.7% of SCE solutions converge to

Table 4 Convergence results. Effectiveness: percentage of solutions of each procedure converging to optimal points; the asterisk indicates the procedure finding the global optimum. Efficiency: range between the minimum and maximum number of model runs necessary by PEST and SCE to converge to the optimal region

			Objective function					
	No.		**1**	**2**	**3**	**4**	**5**	**6**
	Weights	w^w	**1**	**1**	**0**	**0.06**	**0.24**	**0.03**
Model		w^r	**1**	**0**	**1**	**1**	**1**	**1**
Steady-state GW model	Effectiveness	PEST	0%	93.9%*	0%	0%	0%	0%
		SCE	100%*	100%	100%*	100%*	78.8%*	100%*
	Efficiency	PEST	–	51–231	–	–	–	–
		SCE	549–700	527–699	234–356	543–700	620–700	583–699
Transient GW model	Effectiveness	PEST	29.0%*	58.1%*	12.9%*	38.7%*	32.3%*	19.4%*
		SCE	0%	18.2%	0%	6.1%	0%	0%
	Efficiency	PEST	52–127	53–161	53–159	61–186	64–169	52–166
		SCE	–	624–668	–	643–676	–	–
Integrated model	Effectiveness	PEST	–	–	–	–	3.6%	–
		SCE	–	–	–	–	79.7%*	–
	Efficiency	PEST	–	–	–	–	187–684	–
		SCE	–	–	–	–	1219–1395	–

the optimal region, while only 3.6% converge in the case of PEST, illustrating the great difficulties of the local procedure to reach the optimal region for the most complex model. Table 4 also shows the number of model runs necessary for the procedures to locate an optimal solution, thus illustrating the efficiency of the procedures. The efficiency of SCE for the GW models ranges between 234 and 700 model runs, while about twice as many model runs are required for the integrated model, ranging between 1219 and 1395. The number of model runs required by PEST to find optimal solutions is between 51 and 231 for the GW models and increases to between 187 and 684 for the integrated model. The average number of model runs required by one PEST execution to converge to an optimal solution is lower than that needed by SCE. On the other hand, it must be noticed that the overall number of model runs required by the local method to find the solutions of these experiments is much larger compared to that of SCE, as it includes also the unsuccessful iterations of the N_0 trials (see also Table 5). Therefore, the advantages of the fast objective function reduction of PEST are depleted by the necessity of starting the procedure at several points to overcome suboptimal convergence. As a result of this, we can say that, considering the trade-off between effectiveness and efficiency, the global method generally performs better than the local one.

Figure 4 illustrates the convergence of the aggregated objective function for the three models. These plots are shown for objective function no. 5, but they represent features common also to the other functions. The results are shown for all the N_0 optimisation experiments by PEST and for the evolution of the best-so-far results from the SCE population of solutions. As can be seen in the plots, the objective function is generally lowered to reasonable values pretty fast by PEST, while SCE needs a higher number of runs to reach similar results. Only for the steady-state GW model the objective function reduction is much faster for SCE than for PEST. For the GW models some PEST executions end up in points that are suboptimal according to the chosen calibration criterion. In the case of the steady-state GW model, convergence occurs in two suboptimal regions (Figure 4(a)). Similarly, three main regions of attraction of the aggregated objective function can be identified for the transient GW model, two of which are suboptimal (Figure 4(b)). Although for this model PEST is the most effective procedure, its probability of failure is quite high, as previously illustrated in Table 4. For the integrated model the PEST final solutions are not sensitive to particular regions of attraction, as they are more spread in the objective function space compared to the previous cases (Figure 4(c)): however, only two results out of $N_0 = 69$ converge to the optimal region.

The effectiveness of the search algorithms can also be assessed in a multi-objective perspective, i.e. by evaluating the calibration results according to both the discharge and the groundwater head observations. The trade-off between calibration objectives can be described by representing the results from all the evaluations made by each search procedure

Table 5 Total number of solutions contained in the estimated Pareto sets and total number of model runs conducted for each model by PEST and SCE

Model	Procedure	Pareto solutions	Total no. of model runs
Steady-state GW model	PEST	31	19,636
	SCE	541	4200
Transient GW model	PEST	104	18,788
	SCE	187	4145
Integrated model	PEST	15	27,374
	SCE	19	1400

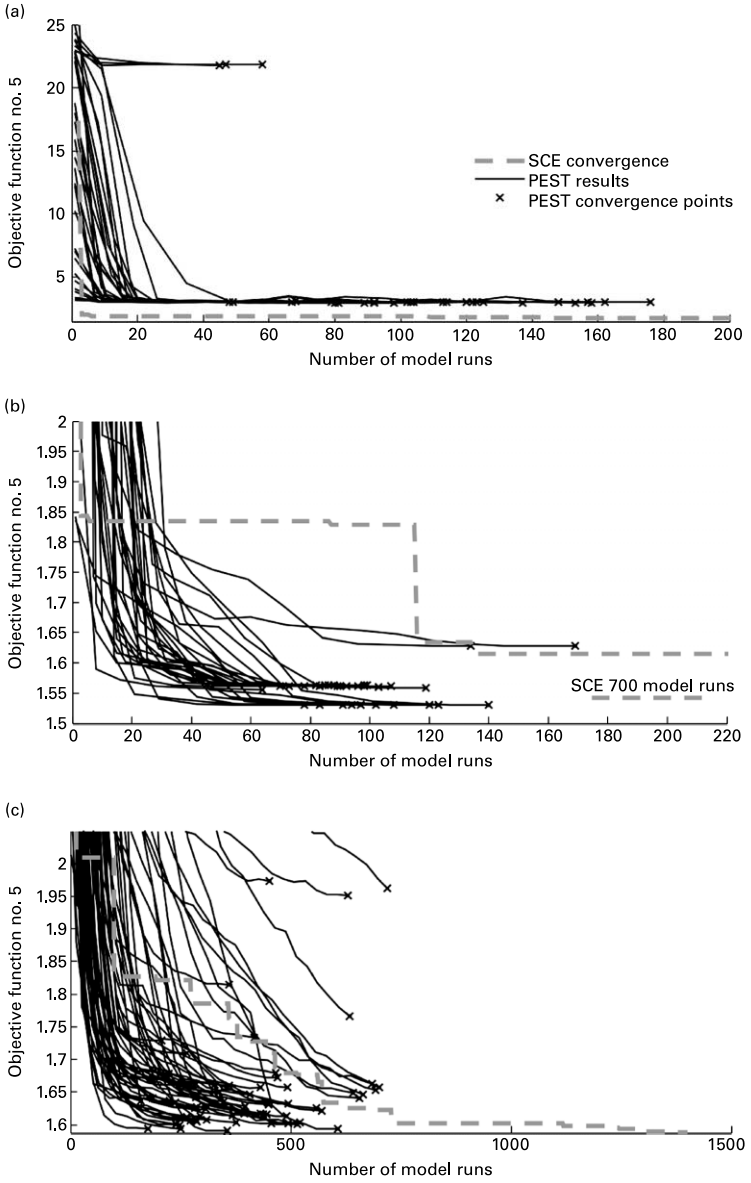

Figure 4 Objective function convergence plots of PEST and SCE calibrations; objective function no.5; (a) steady-state GW model, (b) transient GW model and (c) integrated model

in the objective function space and determining the best solutions according to the Pareto criterion. The optimal Pareto solutions, which are equally good according to the optimisation criteria, are located along a front called the Pareto front. In this way, the trade-off between the optimised objectives is reflected in the multiple solutions of the estimation problem. Moreover, considering the two calibration criteria instead of the aggregated objective function also allows a comparison of the results obtained by applying different aggregation weights to the objective functions, MSE^w and MSE^r. The two estimates of the Pareto front obtained by SCE and PEST for each model are shown in Figure 5. These fronts are generated by combining the solutions found for the six different objective functions for the GW models, whereas only one combination of objective functions is available for the integrated model.

The number of Pareto optimal solutions is reported in Table 5, together with the total number of model runs required by SCE and PEST for each model. Table 5 illustrates the big gap between SCE and PEST results when the number of solutions located along the Pareto front is compared to the number of total model runs required by the two procedures. This trend can be noticed for all the models. As previously illustrated, PEST is more efficient for a single optimisation run, even if convergence to the global optimum is not guaranteed. When considering Pareto optimality, i.e. the ability of locating solutions belonging to the Pareto front, SCE is both more efficient and effective. In fact, for all the models a lower number of model runs is sufficient for SCE to locate more solutions belonging to the Pareto front. SCE is also more effective in exploring the trade-offs between the two objectives because more variability is present in the solutions from the global method. The effect of this feature is also evident in Figure 5: the Pareto fronts estimated by SCE are generally more extended than those from PEST results. The main reason why fewer PEST solutions belong to the fronts is due to the sensitivity of the local procedure to particular regions of attraction of the objective function, as already mentioned and illustrated in Figure 4. This results in an incomplete estimation of the Pareto fronts, leaving some areas uncovered where SCE is able to locate solutions, as is particularly evident in Figures 5(a) and (b). Despite all the drawbacks of the local optimisation method, it can be noticed for the transient GW model (Figure 5(b)) that PEST can locate in some cases better balanced solutions than SCE, i.e. solutions which are good according to both the optimisation criteria employed.

Merging SCE and PEST

From the results of the previous analysis, it is seen that both the SCE and the GML algorithms have specific advantages in calibrating hydrological models. In particular, PEST ensures a fast reduction of the objective function, but convergence to the global optimum is not guaranteed. On the other hand, SCE is capable of globally searching the solution space and of exploring a larger variety of solutions, when trade-offs between optimised objectives are present, without the problem of being trapped in local optima. However, a full SCE execution needs more model runs than a single PEST calibration. Therefore, the potential of combining the SCE and PEST procedures to develop a more efficient and effective methodology is also explored in this work. In this respect, two attempts at merging SCE and PEST are made. The first one, similarly to what was done by Kuczera (1997), consists in fine-tuning the SCE estimate of the parameters by means of a single PEST run. The second approach is based on the awareness that the results of the GML method are strongly dependent on the starting point of the search and that multiple PEST runs started from different points enhance the chance to converge to the optimal region. Therefore, in the second trial, several PEST runs are launched from a given SCE population at an initial convergence stage.

In the first attempt of hybrid optimisation a PEST execution is started from the best point found by one SCE optimisation. This result is compared to that of a new SCE calibration, run with a larger population size (i.e. with the number of complexes increased from 3 to 5) to decrease the probability of converging into a suboptimal region. This study is conducted using the transient GW model, as it has an average degree of complexity among those used and, at the same time, it is not too expensive to run. Moreover, as the solutions previously found by SCE for this model are not as good as those by PEST, it is also tested whether better results can be found by tuning some of the SCE algorithmic parameters. The objective function no. 5 is the one chosen for the test, as it provides more balanced solutions according to the two optimisation criteria.

The results of the combined methods and of the enhanced SCE execution are compared in terms of efficiency and effectiveness. Figure 6 shows the estimates of the Pareto front obtained by running SCE with 3 and 5 complexes and the refinement of the SCE solution

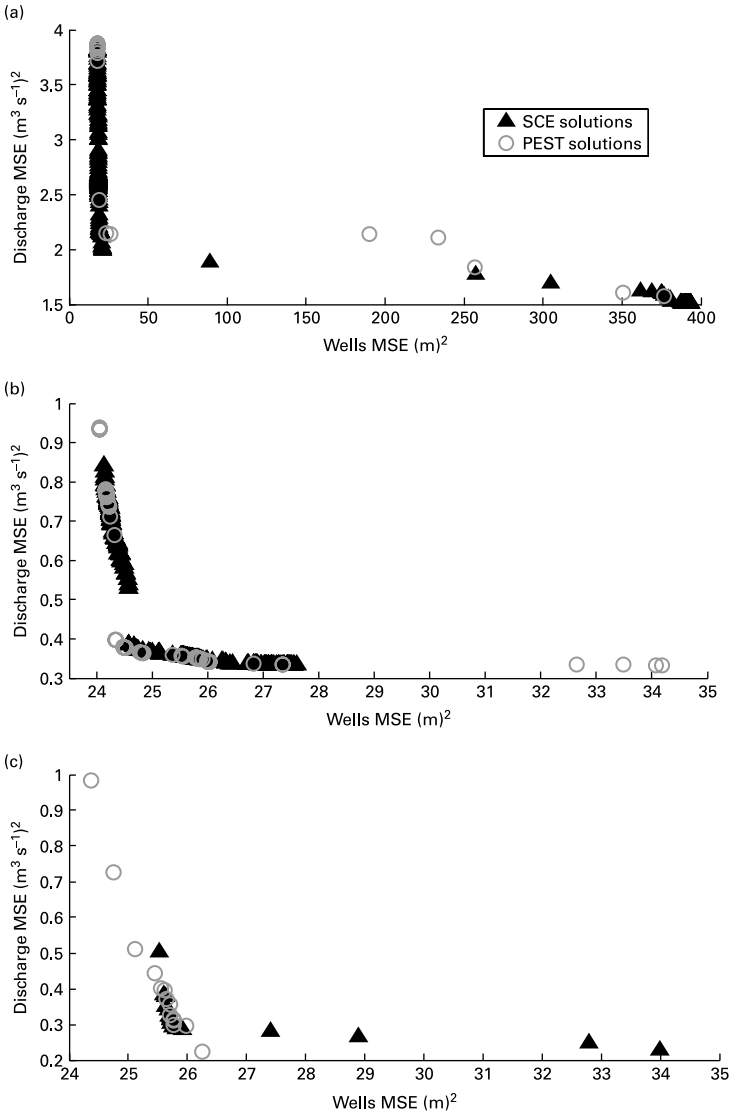

Figure 5 Pareto fronts found by PEST and SCE; (a) steady-state GW model, (b) transient GW model and (c) integrated model (results only for objective function no. 5)

found by starting PEST from the best solution found by SCE with a lower number of complexes. As can be noticed in the figure, the Pareto front of the SCE execution with 5 complexes dominates the front found by running SCE with 3 complexes. This indicates that the results found by the global procedure in the previous calibration experiments can be further improved, if the SCE procedure is adequately tuned. This also explains the poorer performance of SCE than PEST for the transient GW model. The best solution by SCE with 5 complexes can be compared to the one resulting by coupling SCE with PEST. As for the effectiveness of the two approaches, by coupling the global and local methods the aggregated objective function is reduced from the initial value of 1.542 to 1.531, while SCE with 5 complexes reaches 1.535. A further reduction of the objective function by SCE up to the estimate refined by PEST might be possible, but at the cost of an increased number of complexes and, thus, at an increased computational cost. As for the efficiency of the

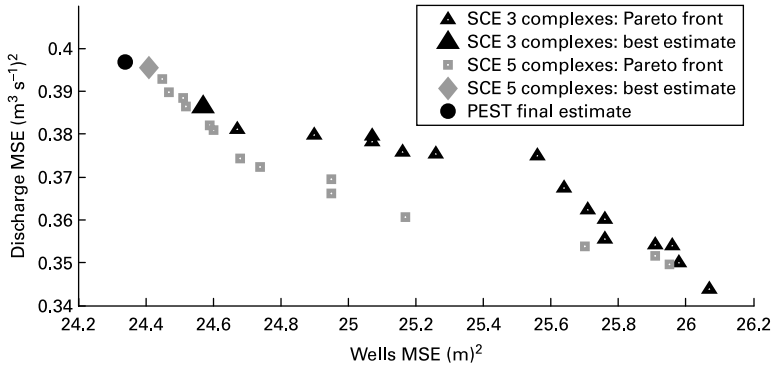

Figure 6 Improvement of SCE results. Pareto fronts estimated by running SCE with 3 and 5 complexes and result of a PEST execution started from the best solution by SCE with 3 complexes. Transient GW model, objective function no. 5

two approaches, coupling the global and local procedures requires a total of 751 model runs (668 by SCE plus 47 by PEST), against the 1059 required by running SCE with 5 complexes. Therefore, the hybrid approach is more efficient for a fixed trade-off between objectives. If more variety of solutions is of interest, running the global method with a larger population size can provide additional solutions at a slightly increased computational cost.

In the second experiment a SCE calibration is initially run and PEST is subsequently launched from different points from the SCE optimisation. Following this approach, PEST is still used as a fine-tuning procedure, but in this case the local method is started at an initial convergence stage of SCE and from more than one point. The steady-state GW model is chosen to demonstrate this method and to further investigate the poor performance demonstrated by PEST for this model. In fact, by launching PEST from new points, it is also tested whether the bad performance of the local method on the simpler model is caused by the particular initial LHS selection of the N_0 starting points of the search. As in the previous experiment, objective function no. 5 is used in this hybrid-optimisation approach. Six points from the SCE results within the first 150 runs are selected as points from where to start the PEST executions.

The results of this experiment are shown in Figure 7, which illustrates the convergence pattern of the aggregated function optimised by SCE, together with the refinements by PEST started at different points. As for the effectiveness of this method, Figure 7 clearly illustrates that the convergence of the local method to suboptimal solutions cannot be avoided (similar results were found by Heidari and Ranjithan (1998)). This confirms the previous results: PEST still has convergence problems for the steady-state model, demonstrated, also in this case, by the very low probability of convergence to the global optimum. Only one of the 6 PEST calibration runs reaches a value of the aggregated objective function (1.66), which is contained in the optimal region and which is also lower than the optimum obtained by the SCE calibration (1.67).

We must point out that we cannot directly compare the calibration results of applying these two different combined methods, as they have been obtained using different models. However, it seems that the first approach is the more effective of the two, since it allows a reduction of the objective function, which would have been impossible (or obtainable with a very low probability) by using PEST and achievable at a very high computational cost by SCE. Starting PEST at the optimal point found by SCE will significantly reduce the risk of the local procedure being trapped in suboptimal regions of attraction, while this drawback

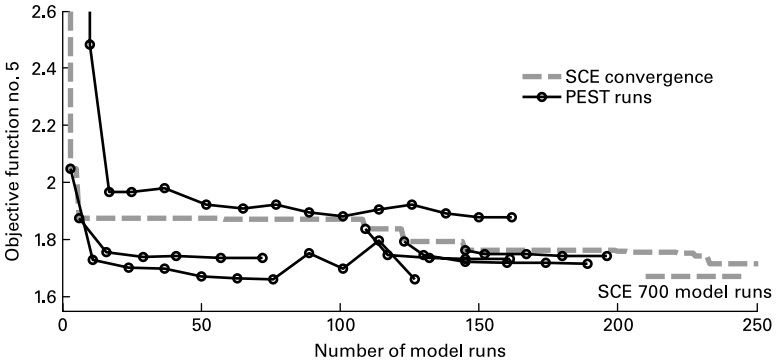

Figure 7 Convergence of SCE and refinements by PEST starting at different points from the SCE search. Steady-state GW model, objective function no. 5

is still present when PEST is started from initial points of the SCE convergence. As for the computational time, the first method requires a total number of model simulations comparable to that required by the global method. Instead, the second approach, similarly to the application of the pure local procedure, ensures a quick reduction of the objective function for a single PEST execution, but more executions have to be run, since convergence to the region where the global optimum is located is not guaranteed. Moreover, both approaches demonstrate that, when PEST is started from SCE suboptimal points, a further refinement of the solution is always found by the local procedure. As calibration of real case studies is conducted on a response surface with an unknown optimum value, the criteria to terminate the search must be fixed *a priori* by the modeller on a relative change of the parameters or the objective function value. Starting a local procedure as PEST from the optimal solution obtained by a global optimisation method can be used as a cheap test to verify if the global method has converged to the global optimum or if a further improvement of the results is still possible.

We are aware that these are only preliminary results of merging the global and local calibration procedure, but some conclusions can be drawn from these findings and used to address the future research in this field.

Discussion and conclusion

This paper compares the results obtained by applying two different automatic calibration algorithms to hydrological models of different complexity. The methodologies used are the SCE global calibration procedure and the PEST local optimisation method. The hydrological models are three distributed physically based models: a steady-state and a transient GW model and an integrated model applied to a Danish catchment. Two optimisation criteria, defined from measurements of the river discharge and the groundwater elevations, are included in the calibration process by means of six different aggregated objective functions. The efficiency and effectiveness of the two procedures are compared in terms of objective function reduction and parameter estimates.

The main finding about parameter estimates is that SCE identifies more well-determined parameters than PEST. It is also found that the specific objective function used affects the results of the two procedures, mainly in terms of parameter identifiability, while the model formulation has an effect on the parameter estimates, in particular by shifting their calibrated values. The SCE results are more sensitive to these variations, while the PEST solutions are generally quite spread in the parameter space for all the case studies considered. The correlation among the parameters is low for all the cases considered, in

particular for the objective functions that include both types of measurements. Different correlation patterns among the parameters are found by the two procedures in terms of pairs of parameters with a significant correlation and correlation coefficient values. In all the cases, PEST provides higher correlated parameter estimates than SCE. We do not know which calibration method gives more realistic results for our case studies. We can only acknowledge the facts that different conclusions about parameter sensitivity, identifiability and correlation can be drawn based on applying different calibration procedures and that parameter estimates are more identifiable when a global calibration method is applied than a local one. This can be the effect of SCE over-fitting the best solutions and PEST being trapped in suboptimal regions of attraction far from the optimum. These findings reveal that parameter identifiability and correlation, as well as model sensitivity to the parameters, cannot be objectively evaluated based on a calibrated set of results, as they are deeply affected by the calibration method employed. Therefore care must be taken in judging these properties of the system/model based on the solutions obtained by an optimisation procedure.

It is difficult to make a general conclusion of the two procedures with respect to effectiveness and efficiency, as the performance of the two methods varies with the models used. The results by PEST are deteriorated by the frequent convergence to suboptimal regions of attraction in the objective function space. In fact, despite PEST is more effective than SCE in the case of the transient GW model, the probability of a single optimisation run to converge to the optimum is quite low for almost all the cases. In particular, it is 0% for 5 out of 6 cases for the steady-state GW model, the kind of model for which local optimisation techniques have been extensively applied. For SCE the effectiveness is 100% for 5 out of 6 cases for the steady-state GW model and large also in other cases except for the transient GW model. Even that the average number of model runs required by a single PEST execution to locate an optimal solution is lower than that needed by SCE, the low probability of PEST to converge to the optimal region makes multiple runs from different starting points necessary, thus decreasing the computational advantages of the local method. Therefore, considering the trade-off between effectiveness (as probability of converging to the optimal region) and efficiency (as total number of model simulations to run), we believe that the global method is preferable to the local method.

If the solutions are evaluated in the multi-objective space using the Pareto optimality criterion, SCE is both more effective and efficient in exploring the trade-offs between the two objectives. Compared to PEST, SCE can identify more solutions along the Pareto front, by requiring a lower total number of model runs. PEST fails to provide a complete estimate of the front, but in some cases it locates better balanced solutions than SCE, according to the two optimisation criteria.

It is also demonstrated that using PEST to refine the SCE optimum parameter estimates is a promising way of merging the global and local methods. This procedure ensures a relatively fast convergence to the global optimum from an advanced stage of the SCE optimisation, overcoming the high probability of suboptimal convergence of PEST. In this respect the local procedure is also suitable to verify if the results of a global method have converged to the global optimum or if an improvement of the parameter estimates can still be achieved.

It should be emphasised, as is the case with any model application to real world problems, that the formulation and set-up of the model used for testing the two optimisation algorithms are based on specific choices regarding model perameterization, selection of calibration parameters and calibration data, and definition of objective functions. All these choices will have an impact on the calibration. For instance, the convergence problems we observe for the PEST algorithm may be caused by model structural deficiencies, e.g. that the

applied 2D model for the saturated zone is too simple to represent the groundwater flow in the catchment. In addition, especially in relation to calibration of the steady-state GW model, the use of only 17 observation wells may not be sufficient to properly constrain the parameter estimation, resulting in an ill- or poorly-posed optimisation problem. It may be speculated that different results would be obtained if another perameterization of the saturated zone was applied and more data were used in the calibration, and hence one may question the generality of our results. However, we think that the model perameterization and choice of calibration data used in this study reflect a typical real world model application. The results demonstrate strengths and weaknesses of the two parameter estimation algorithms, which are essential to acknowledge when applied to hydrological model calibration.

References

Bentley, L.R. (1993). Least squares solutions and calibration of steady state groundwater flow systems. *Adv. Wat. Res.*, **16**, 137–148.

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

Boyle, D.P., Gupta, H.V. and Sorooshian, S. (2000). Towards improved calibration of hydrologic models: combining the strengths of manual and automatic methods. *Wat. Res. Res.*, **36**(12), 3663–3674.

Carrera, J. and Neuman, S.P. (1986). Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information. *Wat. Res. Res.*, **22**, 199–210.

Christensen, S. and Cooley, R.L. (1999). Evaluation of confidence intervals for a steady-state leaky aquifer model. *Adv. Wat. Res.*, **22**(8), 807–817.

Cooper, V.A., Nguyen, V.T.V. and Nicell, J.A. (1997). Evaluation of global optimization methods for conceptual rainfall-runoff model calibration. *Wat. Sci. Technol.*, **35**(5), 53–60.

Dawdy, D.R. and O'Donnell, T. (1965). Mathematical models of catchment behaviour. *J. Hydraul. Div. ASCE*, **91**(HY4), 123–137.

D'Agnese, F.A., Faunt, C.C., Hill, M.C. and Turner, A.K. (1999). Death valley regional ground-water flow model calibration using optimal parameter estimation methods and geoscientific information systems. *Adv. Wat. Res.*, **22**(8), 777–790.

DHI (2004). *AUTOCAL, Autocalibration Tool, User Guide*, DHI Water & Environment, Denmark.

Doherty, J. (2005). *PEST: Model Independent Parameter Estimation, User Manual* (5th edn), Watermark Numerical Computing, Brisbane, Australia.

Dornes, P.F. (2003). *Calibration Strategy for Parameter Estimation in the MIKE SHE Modelling System*. Master of Science thesis. International Institute for Infrastructural, Hydraulic and Environmental Engineering, IHE, Delft, The Netherlands.

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

Duan, Q., Sorooshian, S. and Gupta, V.K. (1994). Optimal use of the SCE-UA global optimization method for calibrating watershed models. *J. Hydrol.*, **158**(3–4), 265–284.

Dubus, I.G., Beulke, S., Brown, C.D., Gottensbüren, B. and Dieses, A. (2004). Inverse modelling for estimating sorption and degradation parameters for pesticides. *Pest Mngmnt. Sci.*, **60**, 859–874.

Eckhardt, K. and Arnold, J.G. (2001). Automatic calibration of distributed catchment model. *J. Hydrol.*, **251**, 103–109.

Franchini, M. and Galeati, G. (1997). Comparing several genetic algorithm schemes for the calibration of conceptual rainfall-runoff models. *Hydrol. Sci. J.*, **42**(3), 357–379.

Franchini, M., Galeati, G. and Berra, S. (1998). Global optimization techniques for the calibration of conceptual rainfall-runoff models. *Hydrol. Sci. J.*, **43**(3), 443–458.

Gan, T.Y. and Biftu, G.F. (1996). Automatic calibration of conceptual rainfallrunoff models: optimization algorithms, catchment conditions, and model structure. *Wat. Res. Res.*, **32**(12), 3513–3524.

Graham, D.N. and Butts, M.B. (2006). Flexible, integrated watershed modelling with MIKE. In in V.P. Singh and D.K. Frevert (Eds.), *Watershed Models*, CRC Press, Boca Raton, pp. 245–272.

Gupta, H.V., Bastidas, L., Vrugt, J.A. and Sorooshian, S. (2003). Multiple criteria global optimization for watershed model calibration. In in Q. Duan, *et al.* (Eds.), *Calibration of Watershed Models*, Water Sci. Appl. Ser. 6. AGU, Washington, DC, pp. 125–132.

Gupta, H.V., Sorooshian, S. and Yapo, P.O. (1998). Toward improved calibration of hydrological models: multiple and noncommensurable measures of information. *Wat. Res. Res.*, **34**(4), 751–763.

Heidari, M. and Ranjithan, S.R. (1998). A hybrid optimization approach to the estimation of distributed parameters in two-dimensional confined aquifers. *J. AWRA*, **34**(4), 909–920.

Henriksen, H.J., Troldborg, L., Nyegaard, P., Sonnenborg, T.O., Refsgaard, J.C. and Madsen, B. (2003). Methodology for construction, calibration and validation of a national hydrological model for Denmark. *J. Hydrol*, **280**(1–4), 52–71.

Hill, M.C. (1998). *Methods and Guidelines for Effective Model Calibration*. U S Geological Survey, Water Resources Investigations Report 98-4005. US Geological Survey, Denver, CO. Available at: http://wwwbrr.cr.usgs.gov/projects/GW_Subsurf/mchill/pubs/method/wr984005.pdf.

Hsieh, L.S. and Wang, R.Y. (1999). Semi-distributed parallel-type linear reservoir rainfall-runoff model and its application in Taiwan. *Hydrol. Process.*, **13**(8), 1247–1268.

Islam, N., Wallender, W.W., Mitchell, J., Wicks, S. and Howitt, R.E. (2006). A comprehensive experimental study with mathematical modeling to investigate the affects of cropping practices on water balance variables. *Agric. Wat. Mngmnt.*, **82**, 129–147.

Johnston, P.R. and Pilgrim, D.H. (1976). Parameter optimization for watershed models. *Wat. Res. Res.*, **12**(3), 477–486.

Kristensen, K.J. and Jensen, S.E. (1975). Model for estimating actual evapotranspiration from potential evapotranspiration. *Nordic Hydrol.*, **6**(3), 170–188.

Kuczera, G. (1997). Efficient subspace probabilistic parameter optimization for catchment models. *Wat. Res. Res.*, **33**, 177–185.

Kuczera, G. and Mroczkowski, M. (1998). Assessment of hydrologic parameter uncertainty and the worth of multiresponse data. *Wat. Res. Res.*, **34**(6), 1481–1489.

Lee, R.H. and Wang, R.Y. (1998). Parameter estimation with colored noise effect for differential hydrological grey model. *J. Hydrol.*, **208**(1–2), 1–15.

Madsen, H. (2000). Automatic calibration of a conceptual rainfall-runoff model using multiple objectives. *J. Hydrol.*, **235**, 276–288.

Madsen, H. (2003). Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. *Adv. Wat. Res.*, **26**, 205–216.

Madsen, H., Wilson, G. and Ammentorp, H.C. (2002). Comparison of different automatic strategies for calibration of rainfall-runoff models. *J. Hydrol.*, **261**, 48–59.

Marshall, L., Nott, D. and Sharma, A. (2004). A comparative study of Markov Chain Monte Carlo methods for conceptual rainfall-runoff modelling. *Wat. Res. Res.*, **40**. W02501, doi: 10.1029/2003WR002378.

McKay, M.D., Conover, W.J. and Beckman, R.J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, **21**, 239–245.

McLaughlin, D. and Townley, L.R. (1996). A reassessment of the groundwater inverse problem. *Wat. Res. Res.*, **32**(5), 1131–1161.

Mertens, J., Madsen, H., Feyen, L., Jacques, D. and Feyen, D. (2004). Including prior information in the estimation of effective soil parameters in unsaturated zone modelling. *J. Hydrol.*, **295**, 251–269.

Moore, C. and Doherty, J. (2006). The cost of uniqueness in groundwater model calibration. *Adv. Wat. Res.*, **29**, 605–623.

Muleta, M.K. and Nicklow, J.W. (2005). Sensitivity and uncertainty analysis coupled with automatic calibration for a distributed watershed model. *J. Hydrol.*, **306**, 127–145.

Nash, J.E. and Sutcliffe, J.V. (1970). River flow forecasting through conceptual models. 1. A discussion of principles. *J. Hydrol.*, **10**(3), 282–290.

Nelder, J.A. and Mead, R. (1965). A simplex method for function minimization. *Comput. J.*, **7**(4), 308–313.

Pareto, V (1906). *Manuale di Economica Politica. Societa' Editrice Libraria, Milano, Italy* (Schwier, A.S (1971) *Manual of Political Economy.* Macmillan, New York, English translation).

Poeter, E.P. and Hill, M.C. (1997). Inverse models: a necessary next step in ground-water modeling. *Ground Water*, **35**(2), 250–260.

Poeter, E.P. and Hill, M.C. (1998). *Documentation of UCODE, a Computer Code for Universal Inverse Modeling*. U S Geological Survey, Water Resources Investigations Report 98-4080. US Geological Survey,

Denver, CO. Available at: http://www.mines.edu/igwmc/freeware/ucode/ucode-PREJUPITER/UCODEmanual_wrir98-4080.pdf.

Refsgaard, J.C. (1997). Parameterisation, calibration and validation of distributed hydrological models. *J. Hydrol.*, **198**, 69–97.

Shigidi, A. and Garcia, L.A. (2003). Parameter estimation in groundwater hydrology using artificial neural networks. *J. Comput. Civil Engng.*, **17**(4), 281–289.

Skahill, B.E. and Doherty, J. (2006). Efficient accommodation of local minima in watershed model calibration. *J. Hydrol.*, **329**, 122–139.

Solomatine, D.P., Dibiki, Y.B. and Kukuric, N. (1999). Automatic calibration of groudwater models using global optimisation techniques. *Hydrol. Sci. J.*, **44**(6), 879–894.

Sonnenborg, T.O., Christensen, B.S.B., Nyegaard, P., Henriksen, H.J. and Refsgaard, J.C. (2003). Transient modeling of regional groundwater flow using parameter estimates from steady-state automatic calibration. *J. Hydrol.*, **273**(1), 188–204.

Sorooshian, S. and Gupta, V.K. (1995). Model calibration. In *Computer Models of Watershed Hydrology*, V.P. Singh (Ed.),, Water Resources Publications, Colorado, pp. 23–68.

Sumner, N.R., Fleming, P.M. and Bates, B.C. (1997). Calibration of a modified SFB model for twenty-five Australian catchments using simulated annealing. *J. Hydrol.*, **197**, 166–188.

Thyer, M., Kuczera, G. and Bates, B.C. (1999). Probabilistic optimization for conceptual rainfall-runoff models: a comparison of the shuffled complex evolution and simulated annealing algorithms. *Wat. Res. Res.*, **35**(3), 767–773.

van Griensven, A. and Bauwens, W. (2003). Multiobjective autocalibration for semidistributed water quality models. *Wat. Res. Res.*, **39**(12), 1348. doi: 10.1029/2003WR002284.

Vrugt, J.A., Gupta, H.V., Bastidas, L.A., Bouten, W. and Sorooshian, S. (2003a). Effective and efficient algorithm for multi-objective optimization of hydrologic models. *Wat. Res. Res.*, **39**(8), 1214. doi: 10.1029/2002WR001746.

Vrugt, J.A., Gupta, H.V., Bouten, W. and Sorooshian, S. (2003b). A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Wat. Res. Res.*, **39**(8), 1201. doi: 10.1029/2002WR001642.

Wang, Q.J. (1991). The genetic algorithm and its application to calibrating conceptual rainfall–runoff models. *Wat. Res. Res.*, **27**(9), 2467–2471.

Yapo, P.O., Gupta, H.V. and Sorooshian, S. (1998). Multi-objective global optimization for hydrologic models. *J. Hydrol.*, **204**, 83–97.

Zyvoloski, G., Kwicklis, E., Eddebbarh, A.A., Arnold, B., Faunt, C. and Robinson, B.A. (2003). The site-scale saturated zone flow model of Yucca Mountain: calibration of different conceptual models and their impact on flow paths. *J. Contam. Hydrol.*, **62-63**, 731–750.