Using spatially distributed parameters and multi-response objective functions to solve parameterization of complex applications of semi-distributed hydrological models

Rafael Marcé,[1] Carlos E. Ruiz,[2] and Joan Armengol[1]

Received 30 November 2006; revised 30 July 2007; accepted 8 November 2007; published 27 February 2008.

[1] Application of semi-distributed hydrological models to large, heterogeneous watersheds deals with several problems. On one hand, the spatial and temporal variability in catchment features should be adequately represented in the model parameterization, while maintaining the model complexity in an acceptable level to take advantage of state-of-the-art calibration techniques. On the other hand, model complexity enhances uncertainty in adjusted model parameter values, therefore increasing uncertainty in the water routing across the watershed. This is critical for water quality applications, where not only streamflow, but also a reliable estimation of the surface versus subsurface contributions to the runoff is needed. In this study, we show how a regularized inversion procedure combined with a multiobjective function calibration strategy successfully solves the parameterization of a complex application of a water quality-oriented hydrological model. The final value of several optimized parameters showed significant and consistent differences across geological and landscape features. Although the number of optimized parameters was significantly increased by the spatial and temporal discretization of adjustable parameters, the uncertainty in water routing results remained at reasonable values. In addition, a stepwise numerical analysis showed that the effects on calibration performance due to inclusion of different data types in the objective function could be inextricably linked. Thus caution should be taken when adding or removing data from an aggregated objective function.

Citation: Marcé, R., C. E. Ruiz, and J. Armengol (2008), Using spatially distributed parameters and multi-response objective functions to solve parameterization of complex applications of semi-distributed hydrological models, *Water Resour. Res.*, *44*, W02436, doi:10.1029/2006WR005785.

1. Introduction

[2] Watershed-scale hydrological models are the usual choice in assessment and prediction of water quality in rivers when the origin of the water quality constituents (e.g., point versus diffuse sources) is a major target [*Singh and Woolhiser*, 2002]. Although good empirical alternatives exist if the working timescale is large enough (e.g., *Alexander et al.* [2002] for annual time steps), dynamic models should be considered at shorter timescales. Researchers involved in water-quality oriented hydrological modeling frequently have detailed data for some watershed features (e.g., topography, land uses), but calibration of adjustable model parameters is often done on the basis of a system integrated response (e.g., a single record of streamflow). As a consequence, a fully distributed model will pose the calibration process in a very demanding situation considering the scarce field data available, while a totally lumped model will obviate many details on hand. Thus the semi-distributed approach usually constitutes a good trade-off between fully distributed applications and system integrated models for water-quality oriented hydrological models.

[3] In this modeling framework, the quality and reliability of a water quality constituent calculation depend on a correct assessment or prediction of the quantity of water flowing through the river. To correctly assess the origin of the water quality constituents (and consequently of the water arriving at the river reach), not only should the water flowing through the river channel be modeled, but a good estimation of the surface versus subsurface contributions to the runoff is needed. Unfortunately, in most water quality applications at the watershed scale, surface flows in the channel of the rivers are the only hydrological data available to help the whole parameterization of the hydrological modules. Even if more data exist, problems related to incompleteness and inaccuracy usually restrict the amount of useful data [*Singh and Woolhiser*, 2002]. In spite of this, few works focused on water quality issues check the uncertainty associated with the different components of the water inflow to the channel [e.g., *Kannan et al.*, 2007]. This could be problematic for water quality applications, especially if outcomes should support management decisions. The same flow record in the river channel can be obtained from very different combinations of subsurface inputs and surface runoff which can have dramatic differences in water quality characteristics [*Butturini and Sabater*, 2000].

[1] Fluvial Dynamics and Hydrologic Engineering (FLUMEN), Department of Ecology, University of Barcelona, Diagonal, Barcelona, Spain.
[2] CEERD-EP-W, U.S. Army Engineer Research and Development Center, Vicksburg, Mississippi, USA.

0043-1397/08/2006WR005785$09.00

[4] This common lack of high quality field measurements of water flows across the watershed, summed to the high number of adjustable parameters usually present in semi-distributed models, ends with hydrological simulations that reasonably fit observed river runoff, but with a high degree of uncertainty in parameter values, and consequently in the separation of the modeled response in surface and sub-surface inputs to the river. In complex modeling applications, this situation is even worse. If model-driving characteristics of the watershed (meteorology, land uses, soil properties, geomorphology) have high variability across the basin or over time, the number of parameters needed to account for this variability should be even higher. In such a situation, the uncertainty in adjustable parameter values and routing calculations could be severely enhanced.

[5] During the last years, a considerable effort has been devoted to combine automatic calibration methods with proper data management during the calibration process in an attempt to minimize the uncertainty in the modeled hydrological response [*Liu and Gupta*, 2007]. One approach is the use of multiple objective functions during the calibration processes in order to constrain the parameter uncertainty forcing the system to fit different data measured in the field [*Hunt et al.*, 2006] or different features of a single data series [*Madsen*, 2000]. This can be solved combining the different objective functions in a single one [*Madsen*, 2003], searching for the entire Pareto set of parameters that fits the data [*Gupta et al.*, 1998; *Vrugt et al.*, 2003], or a region of it [*Khu and Madsen*, 2005]. On the other hand, the use of proper regularization and prior information methodologies allow the inclusion of all the information available in our observed data and expertise into the calibration process [*Doherty and Johnston*, 2003]. Regularization refers to supplementing the observation data set with information pertaining directly to parameters [*Tonkin and Doherty*, 2005], which usually implies the use of previous knowledge about parameter behavior (i.e., prior information).

[6] In this study we show how a complex, water-quality oriented watershed hydrological model application can be successfully solved combining several of the recent tools developed for hydrological model calibration. It is a typical hydrological simulation for water quality purposes, in the sense that while the system under study is a relatively big watershed showing strong spatial heterogeneity in both atmospheric inputs and terrain features, only limited river water flow information is available to parameterize the flow routing equations. Thus although some degree of spatial discretization of adjustable parameters seemed desirable a priori, we discarded the use of a fully distributed model. Instead, we calibrated a semi-distributed model considering only a moderate amount of spatial resolution for the adjustable parameters, while maintaining a considerable spatial heterogeneity for fixed parameters estimated from reliable prior information (e.g., topography, land cover).

[7] During calibration we combined: (1) a multiobjective function (OF) built mainly from different features of a single data series; (2) a regularized fitting methodology to allow the inclusion of heterogeneity in model parameters during the calibration processes without converting the problem into an ill, under-determined one; and (3) some prior information on adjustable parameter values.

[8] Additionally, we performed several numerical experiments. First, we estimated the significance of the differences between spatially distributed parameters after calibration in order to assess the usefulness of the spatial discretization of adjustable parameters. Second, we investigated the effect of the inclusion of different data in the objective function on the calibration process. This was assessed with a numerical analysis in which different kind of data were serially added to the OF. Finally, because the final objective of this model will be the simulation of water quality constituents, we paid special attention to the effect of the inclusion of different data types in the OF on the uncertainty of the modeled subsurface input to the river, calculating the predictive uncertainty of this response for a set of different calibration runs.

2. Model Description

[9] The Hydrological Simulation Program-Fortran (HSPF) is a semi-distributed hydrological and water quality model that lumps calculations at level of sub-basin [*Bicknell et al.*, 2001]. Thus although not being a fully distributed model, it allows the inclusion of some spatial heterogeneity of watershed features and meteorological inputs in the modeling framework. Nevertheless, HSPF is a robust, relatively user-friendly application that includes many useful tools and sources of help for modelers, including databases with parameter values from many previous applications [*USEPA*, 1999]. In addition, the USEPA supported BASINS expert system offers a number of preprocessing tools compatible with HSPF [*USEPA*, 2001].

[10] The basic modeling unit in HSPF is the Hydrological Response Unit (HRU), a piece of terrain that is intended to have homogeneous watershed and meteorological characteristics, and thus a unique hydrological response. A network of sub-basins and river reaches is superimposed to the HRUs. The hydrological model is solved inside each HRU, and the different flow components are then diverted to the stream reach corresponding to the sub-basin where the HRU is located. Flow is routed downstream from reach to reach by storage routing (kinematic wave) methods.

[11] HSPF is based upon the original Stanford Watershed Model IV [*Crawford and Linsley*, 1966]. Following is a brief description of the main hydrological routines. Full mathematical formulations are given by *Bicknell et al.* [2001]. HSPF is primarily an infiltration-excess model (Figure 1) that separates water inputs into infiltrating and non-infiltrating fractions according to three parameters: a surface storage capacity value (UZSN), an interflow index (INTFW), and an infiltration index (INFILT) that decreases as soil moisture increases. Saturation-excess overland flow can be simulated by adjusting the exponent used in the infiltration equation (parameter INFEXP). Overland flow is simulated by the Chezy-Manning equation. Interflow is calculated on the basis of a linear relation between the conceptual interflow-storage volume and lateral flow as a function of the interflow recession coefficient (IRC). Water that infiltrates or percolates from the upper soil zone may enter the lower zone storage as determined by the parameter LZSN, the inactive groundwater storage, or the active groundwater storage (Figure 1). The fraction of groundwater that enters inactive groundwater is controlled by the parameter DEEPFR; the remainder enters active groundwater storage and is available for discharge to reach. HSPF

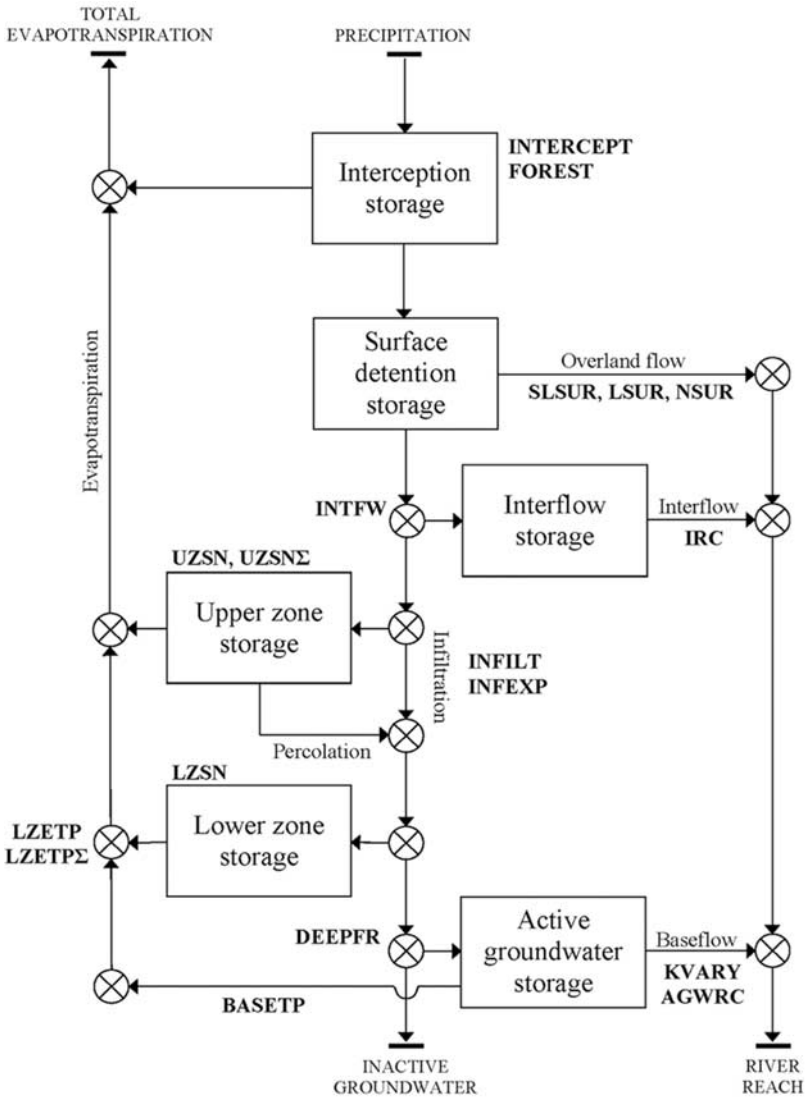

Figure 1. HSPF model structure. Main parameters involved in the different simulated processes are indicated in bold upper case (see Table 1 for explanation).

estimates groundwater outflow from active groundwater storage as a function of three parameters: active groundwater storage (AGWS), the active groundwater recession coefficient (AGWRC), and the active groundwater outflow modifier (KVARY), which governs the extent to which aquifer recharge affects aquifer discharge to the stream. Finally, HSPF computes evapotranspiration as a function of moisture storage and measured potential evapotranspiration, which is adjusted for vegetation cover. It estimates actual evapotranspiration from the potential demand from four sources: interception storage, upper zone storage, vegetation demand, which is satisfied from lower zone storage through the parameter LZETP, and riparian vegetation demand, which is satisfied by active groundwater outflow as stream base flow through the parameter BASETP.

3. Study Watershed

[12] The Ter River watershed is a 1680 km^2 populated area, mainly covered by woodland (78%) and agricultural land (16%). Headwaters are located in the north end of the basin (Figure 2), at 2500 m altitude in the Spanish Pyrenees ranges, and for this study the watershed ends at the Sau Reservoir dam, at 360 m altitude. Land cover, geomorphology, and geological features in the watershed show high spatial heterogeneity across the basin (Figure 2). The headwaters flow over hard materials (igneous and metamorphic rocks) covered by a mixture of high altitude shrublands and conifer and deciduous forests over a steeped terrain (mean slope 0.38). Downstream, around the meeting point of the two main headwater courses, the forested land accounts for nearly all the terrain, the slopes have moderated (mean slope 0.28), and the river flows over sedimentary rocks. The river then enters the populated agricultural plain (most of the slopes under 0.03) where the main human settlements are located. Here the alluvial deposits are abundant, and unirrigated crops cover most part of the land. Finally, the Southeast end of the basin is a mountainous terrain similar to that located in the north top section. Thus the watershed includes a complex mixture of geomorphologic, lithologic, and land cover features. In addition, empirical evidence from twelve meteorological stations located inside the

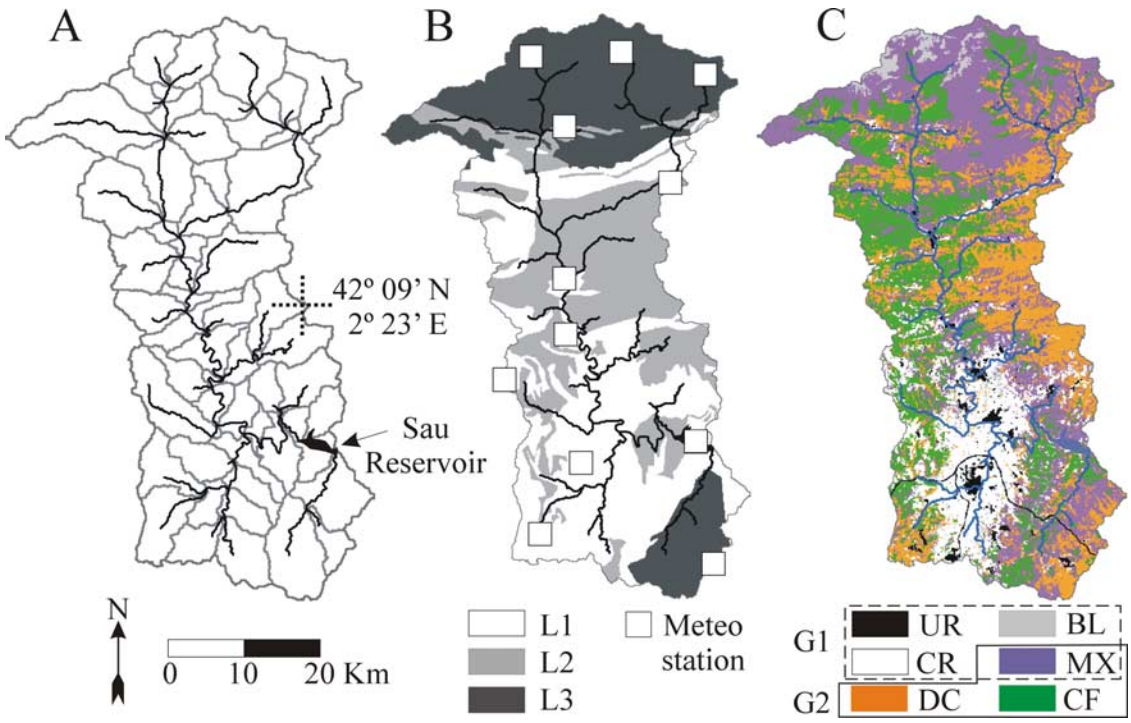

Figure 2. (a) Ter watershed location and sub-basins delineated for HSPF simulation. (b) Lithological zones (L1: alluvial deposits and soft sedimentary rocks; L2: consolidated sedimentary rocks; L3: igneous and metamorphic rocks) and location of meteorological stations. (c) Land uses in the catchment (UR: urban; CR: unirrigated crops; DC: deciduous forest; BL: barren land; MX: for clarity, meadows, shrublands, and few portions of oak forest are included here; CF: conifers forest). G1 groups the non-forested uses, while G2 includes the different forestlands. These are used to distribute the parameter AGWRC (see Table 1).

watershed (Figure 2) shows that rain and other meteorological variables are also highly spatially variable, especially in the North-South direction, and at any timescale. All in all, the watershed under study constitutes a demanding problem for any hydrological model.

3.1. Hydrological and Meteorological Data

[13] Hydrological data for this study come from the daily water budget in Sau Reservoir (Figure 2), which was supplied by the local water agency (Agència Catalana de l'Aigua, ACA). This is a long record that spans from 1964 to the present, but considering the available meteorological information only data from 1 January 1998 to 31 July 2004 were used during modeling exercises. Additionally, several hourly water budgets in Sau Reservoir calculated during storm events were available for the period 1998–2004. Also, we had a daily streamflow record collected at a gauge station just upstream of Sau Reservoir, which worked during the 1980s. Thus we had a direct measure of streamflow to compare with the daily water budget in Sau Reservoir.

[14] Meteorological data consisted of hourly records of temperature, precipitation, relative humidity, wind velocity, solar radiation, atmospheric pressure, and evapotranspiration, collected at 12 automatic meteorological stations (Figure 2) run by the Servei de Meteorologia de Catalunya (METEOCAT), and normally used for regional meteorological characterization.

3.2. Model Input Data Descriptions

[15] Terrain elevation and slope were calculated using a LANDSAT-TM 1997–1998 digital elevation model (DEM) by the Institut Cartogràfic de Catalunya (ICC). Lithological maps were digitalized from ICONA (1990), and land use in the Ter watershed was obtained from a LANDSAT-TM 1997 digital raster map by ICC. Spatial resolution of digital maps were unified to the coarsest scale, which was 30 × 30 m. Information on river channel geometry was found in *ACA* [2001], consisting of 446 morphometric river cross-sections measured for the Ter watershed.

4. Model Implementation

[16] For this study, the HRUs were defined overlaying three watershed features (terrain slope, lithology, and land use) and the meteorological input. The terrain slope was divided into three categories (S1: 0–0.06; S2: 0.06–0.15; S3: >0.15) following other HSPF applications (SWM, 2002). The lithological layer was defined as a three items map, and the land uses layer included eight categories (Figure 2). The meteorological input was spatially distributed assigning each meteorological station to one or more sub-basins, defining twelve meteorological zones. We discarded the use of Voronoi polygons or similar methodologies to draw the meteorological zones because the spatial extent of the meteorological events in this watershed is clearly limited by topography. The combination of the four layers gave a theoretical maximum of 864 HRUs, but due to the spatial covariance between layers and the assimilation of HRUs of less than 5 km^2 by HRUs with similar characteristics, the actual value was 109 HRUs. The 5 km^2 threshold was arbitrary, chosen as a compromise between spatial

resolution and complexity of the model input files. This process of map overlaying resulted in HRUs with very irregular shapes and areas ranging from 5 km^2 to 70 km^2.

[17] The sub-basin delineation was automatically performed with the ArcView algorithm implemented in the BASINS package [*USEPA*, 2001], using the Ter watershed DEM as input. The process resulted in 58 sub-basins (Figure 2). Note that this was a rather subjective step, since an arbitrary threshold area must be supplied to the algorithm, and the average area and final number of sub-basins were very sensitive to this parameter. In fact, the final setting was a compromise between spatial distribution of results and computational time for the modeling effort. Tables summarizing geomorphological traits of river reaches (Table 1) were filled using information from the measured river cross-sections.

[18] Since each of the 109 HRUs generated acts as an independent hydrological model, and HSPF has a considerable amount of parameters (Table 1), we are facing a very complex problem. Fortunately, some of the parameter values could be estimated from GIS based information, inventories, or bibliographical research (e.g., SLSUR, see Table 1). Thus a considerable degree of spatial heterogeneity (i.e., specific parameter values for each HRU) could be preserved. Many parameters represent abstract properties of the basin, or simply there was no field information available, and consequently a calibration process must optimize their values. Considering the field data available for this study and the number of adjustable parameters in HSPF (Table 1), we discarded the possibility to fit different parameter values for each of the 109 HRUs, since this would raise the total amount of adjustable parameters to 1417. Instead, we tried to preserve some heterogeneity by considering different adjustable parameters in relation to some watershed feature (lithology, terrain slope, or land use) considered as a key factor influencing the parameter behavior (Table 1). Information in *USEPA* [1978, 1999, 2000] was used to define the watershed feature relevant to each adjustable parameter. For instance, we defined three adjustable parameters for the model parameter UZSN in relation to mean terrain slope in the HRU. Then, if mean terrain slope in a HRU falls in the S1 category, the adjustable parameter assigned to this HRU will be $UZSN^{S1}$.

[19] In addition, temporal heterogeneity was considered relevant for two parameters: UZSN and LZETP. These parameters could exert a significant control on actual evapotranspiration, which is markedly seasonal in the Mediterranean region. Seasonal variations of UZSN were considered only for HRUs showing terrain slope type 1 (S1), because in these HRUs crops are abundant and UZSN could experience seasonal variation due to agricultural practices. LZETP temporal variability was considered for all the vegetation land uses (see Table 1).

[20] HSPF is a flexible modeling environment that allows the definition of monthly parameter values. Since defining twelve values for these two parameters would almost double the number of adjustable parameters, we introduced two new parameters, UZSNΣ and LZETPΣ (Table 1), which represent the amplitude of variation of the corresponding quantity about its average value. The twelve monthly values were computed using:

$$UZSN_n = UZSN + (UZSN \cdot UZSN\Sigma) \cdot \sin\left((x_n - 270) \cdot \frac{2\pi}{365.25}\right) \quad (1)$$

and,

$$LZETP_n = LZETP + (LZETP \cdot LZETP\Sigma) \cdot \sin\left((x_n - 91) \cdot \frac{2\pi}{365.25}\right) \quad (2)$$

where n = (1, 2,. . .,12) stands for the month (January to December), and x = (1, 32, 60,. . ., 334) represents the day of the year corresponding to the first day of the corresponding month. Thus the two equations define a sinus curve centered at a different day of the year, if the Σ parameter has a non-zero value.

[21] All in all, considering parameters associated with terrain features and seasonal variation, the total amount of adjustable parameters was 34 (see Table 1).

5. Calibration Strategy

5.1. Multiobjective Function

[22] *Gupta et al.* [1998] proposed a new calibration paradigm based on the multiobjective function approach. Following *Madsen* [2003], in a multiOF context calibration can be performed on the basis of multivariable measurements (e.g., river runoff and groundwater levels), multisite measurements (i.e., a variable measured in several sites within the watershed), and multiresponse modes (i.e., OFs that includes several aspects of the hydrological response, but on the basis of the same measured variable). In a typical hydrological simulation for water quality purposes, the multivariable approach is rarely accessible because only river runoff is usually available. Thus in these applications the multiresponse approach is the most frequently applied, supplying multisite measurements of river flow in the best situations. This study represents an extreme situation, where only the multiresponse option was available. Consequently, we applied our efforts in defining a good multiresponse OF. The responses included in the OF follow.

[23] WEEK. This response includes weekly mean flows in the Ter River during the period 1999–2002. Since the only recent flow data within this watershed comes from the daily water budgets in Sau Reservoir, we investigated the relationship at different timescales between the water budget estimation and the river flow measurements from the gauging station that operated just upstream the reservoir during the 1980s. We found that the best correlation was computed with mean weekly data (Pearson's $r^2 = 0.89$, sample size = 313, p-level < 0.0001). The period from January 2003 to July 2004 was reserved as a validation data set.

[24] HOUR. Since the recommended HSPF working time step is the hour, and some of the driving parameters work at this time step [*Bicknell et al.*, 2001], we considered it important to include a series of hourly river flow into the OF. Again, the source was the water balance in Sau Reservoir. Since an hourly record of this kind is prompted to include a high level of measurement noise, we looked for a strong storm event with dramatic short-term changes in the river flow. Only one such event could be found in the database pertaining to the studied period (from 10 May 2002 to 14 May 2002). Thus no validation data were available for this response.

[25] TRIM. We aggregated the flow response in 91 day periods. Although this response cannot be considered independent of the WEEK series, the inclusion of series at coarse time resolution is common during calibration of

Table 1. Parameters of the Hydrological Simulation Program-Fortran (HSPF) Model

Parameters Optimized	Description	Zonation	Prior Value	Range[a]
	Storages			
LZSN	Lower zone nominal storage (in)	3 by lithology		0.01–15
UZSN	Upper zone nominal storage (in)	3 by slope		0.01–10
UZSNΣ	Monthly variability for UZSN	applies only to S1	0	0–1
	Infiltration			
INFILT	Index to the infiltration capacity (in hr^{-1})	3 by lithology	0.08[b]	0.001–0.5
INFEXP	Exponent in the infiltration equation	3 by lithology		0–10
	Recession			
KVARY	Non-exponential groundwater recession (in^{-1})	3 by lithology	0	0–5
AGWRC	Base groundwater recession (day^{-1})	2 by land groups		0.85–1
IRC	Interflow recession (day^{-1})	3 by slope	0.7[c]	0.3–0.85
	Routing			
INTFW	Interflow inflow parameter	3 by lithology		1–10
LZETP	Lower zone evapotranspiration parameter	5 by land use		0.1–0.9
LZETPΣ	Monthly variability for LZETP	applies to all	0	0–1
BASETP	Base flow evapotranspiration by deciduous forest	applies only to DC	0	0–0.2
DEEPFR	Groundwater fraction to inactive groundwater	3 by lithology	0	0–0.5

Parameters Estimated	Description	Source	Reference
	Terrain Features		
COVIND	Snowpack that covers the entire HRU (in)	function of slope	this study[d]
SLSUR	Terrain slope	GIS	[e]
LSUR	Length of the overland flow plane (ft)	function of slope	this study[f]
NSUR	Manning's *n* for the overland flow plane	function of land use	[c]
	Vegetal Cover		
FOREST	Fraction of land covered by vegetation	inventories	[g]
INTERCEPT	Monthly interception storage capacity (in)	inventories + Literature	[h]
	River Reaches		
FTABLE	Tables summarizing reaches morphometry	measured	[i]

[a]Modified from *USEPA*, 2000.
[b]*USEPA*, 1978.
[c]*USEPA*, 2000.
[d]COVIND = 1+10·SLSUR.
[e]LANDSAT-TM 1997–1998 30m grid by Institut Cartogràfic de Catalunya (Spain).
[f]LSUR = 100+6358·exp(−0.099·SLSUR).
[g]*CREAF*, 2001.
[h]*van der Leeden et al.*, 1991; *Llorens*, 1997; *Llorens et al.*, 1997; *Bellot and Escarré*, 1998; *Rodà et al.*, 1999; *Gallart et al.*, 2002.
[i]*ACA*, 2001.

semi-distributed models [*Shamir et al.*, 2005]. The calibration and validation periods were as in WEEK.

[26] BASE. Although the knowledge of subsurface inputs to the river flow are of capital importance both in hydrological and water quality applications, direct measures of this flow are very difficult to obtain. Since one of the aims of this study was to test the effect of different kinds of data on the uncertainty in the modeled subsurface inputs to the river, we considered the inclusion of some estimation of the base flow in the calibration process. Thus we used the *Arnold and Allen* [1999] empirical numerical filter to estimate the base flow hydrograph from the daily flow record in the Ter River. The recharge algorithm developed in this study is an automated derivation of the hydrograph recession curve displacement method [*Rorabaugh*, 1964] that utilizes daily streamflow. As a numerical estimate, this method carries some uncertainty, expressed in the fact that the result is not a single base flow series, but a collection of three passes of the numerical filter (forward, backward, and forward). Although the most probable result is the first pass [*Arnold and Allen*, 1999], we used all the available information defining a specific residual for the BASE data. Whereas for the other responses the residuals *r* between observed and simulated data were just the difference series, for BASE we used:

$$\begin{Bmatrix} if & third\_pass_i < simulated_i < first\_pass_i & r_i = 0 \\ if & simulated_i > first\_pass_i & r_i = simulated_i - first\_pass_i \\ if & simulated_i\ third\_pass_i & r_i = simulated_i - third\_pass_i \end{Bmatrix} \quad (3)$$

where $simulated_i$ stands for the ith modeled subsurface input to the river, and $first\_pass_i$ and $third\_pass_i$ are the ith estimated groundwater flux into the river calculated by the first and third base flow algorithm pass, where most probably the actual base flow value is located. Following

the reasoning detailed for WEEK, we aggregated the series taking weekly mean values, and the calibration and validation periods also were as in WEEK.

[27] DURC. The inclusion of duration curves (i.e., the percent of time in which the river flow equaled or exceeded some given values) is a current practice in HSPF applications [*Doherty and Johnston*, 2003]. The definition of several streamflow values which calculate exceedence times includes a frequency descriptor into the calibration process, which is relatively independent of the original flow series and can be compared to a nonparametric streamflow signature, i.e., its calculation does not depend on statistical assumptions [*Shamir et al.*, 2005]. To calculate DURC we used the daily flow record, because for this nonparametric descriptor the small-scale noise should not have a deleterious effect. We assigned streamflow values which calculate exceedence times (1.7, 3.5, 7, 14, 28, 56, and 85 $m^3 s^{-1}$) considering the flushing nature of streamflow in the Ter River. Consequently, most exceedence points accumulated in the low flow region to avoid having empty or poorly represented intervals in the high flow region. Calibration and validation periods were as in WEEK.

5.2. Regularization and Prior Information

[28] The model structure presented in section 4 determines that we had to calibrate 34 parameters, which is a considerable amount considering the limited field information available. In this context, the risk of over-fitting and poor convergence due to numerical instability problems during the calibration process is elevated. Since the reliability of final parameters values and the resultant spatial and temporal heterogeneity is highly degraded if such over-fitting exists [*Moore and Doherty*, 2005], we should introduce some regularization constraint into the calibration process.

[29] Regularized inversion is founded upon supplementing the observation data set with information pertaining directly to parameters. This takes the form of a series of regularization equations based upon the difference of a parameter value from a preferred value, or the difference between parameters [*Tonkin and Doherty*, 2005]. Since the regularization equations could incorporate bibliographical or previous information on parameter values, regularization and prior information approaches can be inextricably linked. While regularization prevents over-fitting and numerical instability, there are shortcomings associated with its use. Regularization could decrease the ability of the calibration process to identify small-scale parameter variations [*Vogel*, 2002], it could force the calibration process to obviate fine-details of the observed data set which could be important for specific predictions, and its inclusion could lead to estimate parameter values that are in accordance with a wrong prior conceptualization.

[30] We applied three different regularization/prior information strategies. First, we imposed a smoothing constraint on parameter values. This was achieved by taking differences between the values of parameters pertaining to the same model parameter (e.g., between the three LZSN parameters defined by lithology), and requesting that each such difference be zero if possible. The second strategy was to impose an economy criterion. For this, we defined a preferred state for certain parameter values (Table 1), in order to include the effect caused by these parameters only if necessary. For instance, UZSNΣ and LZETPΣ were defined with a preferred value of 0, i.e., no intraannual variability for UZSN and LZETP parameters. Thus any departure from zero will be penalized during the calibration process. In a similar way, a zero preferred prior value was assigned to KVARY, BASETP, and DEEPFR parameters. Finally, the third strategy was to assign a prior value for INFEXP and IRC, using bibliographical information consistent with our watershed characteristics (Table 1). All these regularization constraints and prior information were implemented in the form of prior information equations [*Doherty*, 2004], which constitutes an additional observation group (REGU) added to the OF. This implies that the residuals pertaining to REGU will depart from zero during the calibration process only if this is compensated by the residuals for the other OF components going to zero.

5.3. Calibration Algorithm

[31] The next step is to define the way in which the parameter estimation method will use the above responses during the calibration process. Although the most powerful way would be to search for the entire Pareto set of parameter values that simultaneously fits the different responses (i.e., considering each response as a separate OF), previous runs with standard methods to apply these methodologies [*Vrugt et al.*, 2003] in our problem revealed that the convergence of results was extremely slow (very poor convergence after 100,000 simulations of 25 s each one), therefore unsuitable for all practical purposes. We turned to the PEST package [*Doherty*, 2004] because it implements a very efficient and stable algorithm for OF minimization and because its predictive uncertainty analysis capabilities were the cornerstone of this work.

[32] PEST implements a particularly robust variant of the Gauss-Marquardt-Levenberg (GML) method of nonlinear parameter estimation. While this method requires that a continuous relationship exist between model parameters and model outputs to calculate finite difference derivatives, it can normally find the minimum in the objective function in fewer model runs than any other parameter estimation method [*Doherty*, 2004]. However, the chief disadvantage of the GML method is its propensity to find local minima rather than the global minimum, depending on the user-supplied set of parameters from which the iterative solution process was started. To avoid this important limitation, *Skahill and Doherty* [2006] presented an improvement of the GML method, based on what they called "trajectory repulsion" scheme (hereafter TR-GML). First, a population of initial parameter sets is randomly generated from the hypercube defined by parameter bounds. Model results are generated using these parameter sets, and then these sets are ranked in order of increasing objective function value. Next, it initiates a PEST calibration run, with initial values for this run being equal to the random parameter sample for which the objective function was lowest. After completion of the first PEST run, another PEST run is initiated. For this run, a starting point from the initial population is chosen that is maximally distant from any point on the parameter trajectory taken by the initial PEST run. After the next PEST run is complete, another parameter set is selected from potential starting points. The parameter set selected is that which is maximally distant from all previous points on all previous trajectories. The process is then repeated until no improvement in the OF is detected.

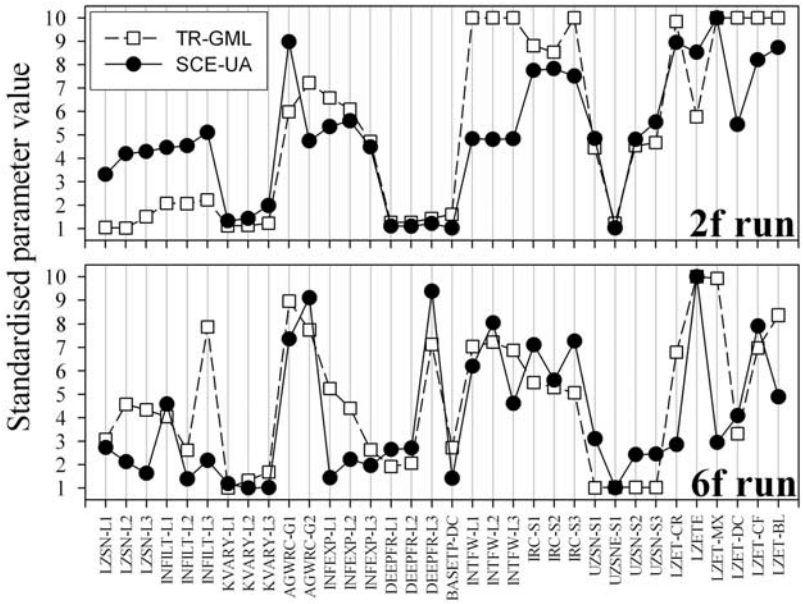

Figure 3. Standardised parameter values for TR-GML and SCE-UA calibration runs. In the upper panel, calibration results considering 2 responses in the OF (2f). In the lower panel, results considering 6 responses in the OF (6f). More details of these calibration settings in the text.

[33] The TR-GML method as implemented in PEST has proven to be a very efficient and robust modification of the GML method to find global OF minima in the context of hydrological modeling. Moreover, comparisons between TR-GML and the SCE-UA algorithm [*Duan et al.*, 1992] showed that TR-GML performs equally or even better than SCE-UA, but demanding much less computational time [*Skahill and Doherty*, 2006; *Gallagher and Doherty*, 2007]. Additionally, we compared results from two calibration runs performed in this study (runs 2f and 6f, see explanations below) obtained using TR-GML and SCE-UA. OF minimum values found during the calibration run 2f were 47.01 using TR-GML (after 7000 model runs) and 51.4 using SCE-UA (after 40,000 model runs). For the run 6f, values were 227.7 using TR-GML (after 15,000 model runs) and 250.0 using SCE-UA (after 96,000 model runs). The optimized parameter values were comparable in both calibration scenarios (Figure 3). Thus the TR-GML methodology is a good choice for the calibration runs performed in this study. We acknowledge that user expertise in SCE-UA may have a large bearing on its efficiency, and although we did our best to maximize SCE-UA efficiency, we recognize that others could have done better. However, this is not the first record of a poorer performance of SCE-UA when compared to other global optimization algorithms calibrating watershed models with tens of adjustable parameters [*Tolson and Shoemaker*, 2007].

5.4. Aggregation of Components

[34] The aggregation of the OF components was implemented in two steps. First, to avoid a bias toward high observed values during the calibration step, we assigned weights for each residual to guarantee that all data pertaining to one specific response had the same potential effect on the objective function during calibration (this weighting will be called W1). Thus residuals were weighted squared differences. The weight for each residual was calculated as the inverse of the observed value. Obviously, other error functions or weighting strategies could be used if certain aspects of the observed series (low flows, storm peaks, etc) are of especial interest, in order to tailor the calibration process to reduce the predictive error of these particular predictions [*Moore and Doherty*, 2005]. However, this was not the case in our study.

[35] Second, since PEST is not able to manage several OFs at the same time, we aggregated the different responses into a single OF, using the Euclidian distance function by *Madsen* [2003]. This transformation function compensates for differences in the magnitudes of the different responses, giving to all the same influence on the aggregated OF near the optimum. This is equivalent to search a balanced optimum Pareto solution [*Khu and Madsen*, 2005]. The original procedure proposed by *Madsen* [2003] uses the best parameterization from the initial sample of SCE-UA as the reference for the transformation function. Although this could be applied in our case simply using the best parameter set from the initial sampling of the TR-GML algorithm, preliminary calibration runs (i.e., without Madsen's transformation) suggested that this "best" parameter set gave a poor approximation to the optimum OF, probably because of the high number of adjustable parameters included in our model. Thus we used results from the best preliminary calibration run to apply Madsen's transformation to the calibration runs in this study. However, note that results from this work should not be used to support the outperformance of our procedure over Madsen's one, since no systematic comparison was done.

[36] The objective function was defined as the sum of all weighted squared residuals between observed and modeled data, including REGU. This weighting strategy in two steps (first to promote equal importance inside each response (W1), and then to search a balanced Pareto solution)

implied that every observed figure roughly had the same influence on the calibration process.

6. Numerical Analyses

6.1. General Performance

[37] We first solved a calibration run including all the available information, i.e., defining an OF with all the responses detailed in section 5.1, plus REGU as a regularization constraint. This run will be referred hereafter as 6f (for six responses in the OF). Results from this calibration run were used for evaluating the performance of the calibration strategy and for assessing the spatial distribution of adjustable parameters.

[38] We evaluated model performance calculating the Nash-Sutcliffe Coefficient [*Legates and McCabe*, 1999] for each observed response and the corresponding model results. For the calculation of NS, we used the W1 weights applied during calibration, in order to avoid sensitivity to extreme values [*Madsen*, 2000]:

$$NS = 1 - \frac{\sum_{i=1}^{N} w_i^2 (O_i - S_i)^2}{\sum_{i=1}^{N} w_i^2 (O_i - \bar{O})^2} \quad (4)$$

where NS is the Nash-Sutcliffe Coefficient for a specific response, N is the total number of observations for this response, O_i is the *i*th observed value, S_i is the *i*th simulated value, $\bar{O}$ is the average value of the N observations, and w_i is the *i*th weight assigned following W1 weighting function.

[39] In addition, we visually inspected the fit between model results and observed data, plotting observed time series and the corresponding 95% predictive confidence intervals. Schefeé non-linear predictive confidence intervals were defined implicitly by [*Gallagher and Doherty*, 2007]:

$$\Phi(p) - \Phi(p^o) \le m\sigma_r^2 F_\alpha(m, n-m) \quad (5)$$

where $\Phi(p)$ is the objective function calculated for a parameter set p, $\Phi(p^o)$ is the objective function calculated for optimized parameter values, m stands for the number of adjustable parameters, and n for the calibration sample size. $F_\alpha(m,n\text{-}m)$ is the F(m, n-m) distribution at the α level, and $\sigma_r^2 = \Phi(p^o)/(n\text{-}m)$. For a particular model prediction (e.g., WEEK volume during the calibration period), a confidence interval can be calculated by minimizing/maximizing that prediction while constraining parameters to lie within the join confidence region defined implicitly by equation (5). For a non-linear model, this minimization/maximization is an iterative process (and thus computationally intensive), which PEST solves using the method based on Lagrange multipliers developed by *Vecchia and Cooley* [1987].

[40] Finally, parameter sensitivity to the different responses included in the OF was calculated as the magnitude of the column of the Jacobian matrix (calculated by PEST) pertaining to that parameter, and normalized with respect to the number of observations. Composite sensitivity of a parameter was calculated as the average of the sensitivities calculated for the 5 OF components based on field data. The Jacobian matrix composed of derivatives of observations with respect to each adjustable parameter and its calculation is central during TR-GML optimization. As implemented in PEST, the elements of the Jacobian matrix associated with different parameters are independent, since derivatives for each parameter are calculated in turn.

6.2. Spatial Distribution of Adjustable Parameters

[41] In order to elucidate if the final values of related parameters were different enough to justify the proposed zonation, we tested paired differences between parameters pertaining to the same model parameter. First, we estimated the distribution of the final parameters values taking advantage of the parameter confidence interval calculation implemented in PEST [*Doherty*, 2004]. At the end of the calibration run, PEST gives the optimized value (i.e., the 50 percentile), and the 2.5 and 97.5 percentiles for each parameter. These percentiles are calculated as linear individual parameter intervals using the relationship [*Gallagher and Doherty*, 2007]:

$$\mathrm{Prob}\left\{p_i^o - t_{\alpha/2}(n-m)\sigma_i \le p_i \le p_i^o + t_{\alpha/2}(n-m)\sigma_i\right\} = 1 - \alpha. \quad (6)$$

where p_i is the true (unknown) value of the *i*th parameter, p_i^o is the adjusted value, $t_{\alpha/2}(n\text{-}m)$ is the upper $\alpha/2$ confidence point of the Student's t(n-m) distribution, and σ_i is the *i*th diagonal element of the parameter covariance matrix. Confidence intervals were truncated at parameter limits (Table 1). It is worthy to note that these confidence limits are calculated on the basis of the same linearity assumption used during the GML optimization. Thus their calculation is very efficient in terms of computational demand, but if the confidence limits are large they will extend further into parameter space than the linearity assumption itself, compromising the reliability of calculations. However, the use of linear parameter confidence limits during calibration of HSPF applications has proven to be a good alternative for parameter uncertainty assessment [*Gallagher and Doherty*, 2007].

[42] After calculating the 2.5, 50, and 97.5 percentiles, we adjusted a Beta cumulative distribution function (CDF) to these three points in order to obtain an approximation of the posterior parameter distribution. We chose the Beta CDF for this purpose because it is an extremely flexible function with only two adjustable parameters and is capable of generating very different probability density functions (symmetric unimodals, skewed unimodals, strictly increasing and decreasing functions, and more). Thus by using a Beta CDF the only assumption was that the posterior parameter distribution was smoothed between estimated percentile points. However, in few cases the deviation from a Beta function was too high to prevent a Beta CDF to fit one of the extreme percentiles. In those cases we adjusted an additional Beta CDF using the problematic extreme percentile and the 0.50 percentile. Then, we built a chimera CDF using the CDF sections that adequately fitted the extreme percentiles at both sides of the 0.50 percentile.

[43] Once the empirical distributions were obtained, we randomly sampled 10,000 points from each of the fitted CDFs. These synthetic samples were then used to test paired differences between the values of the parameters pertaining to the same model parameter (e.g., between the three LZSN

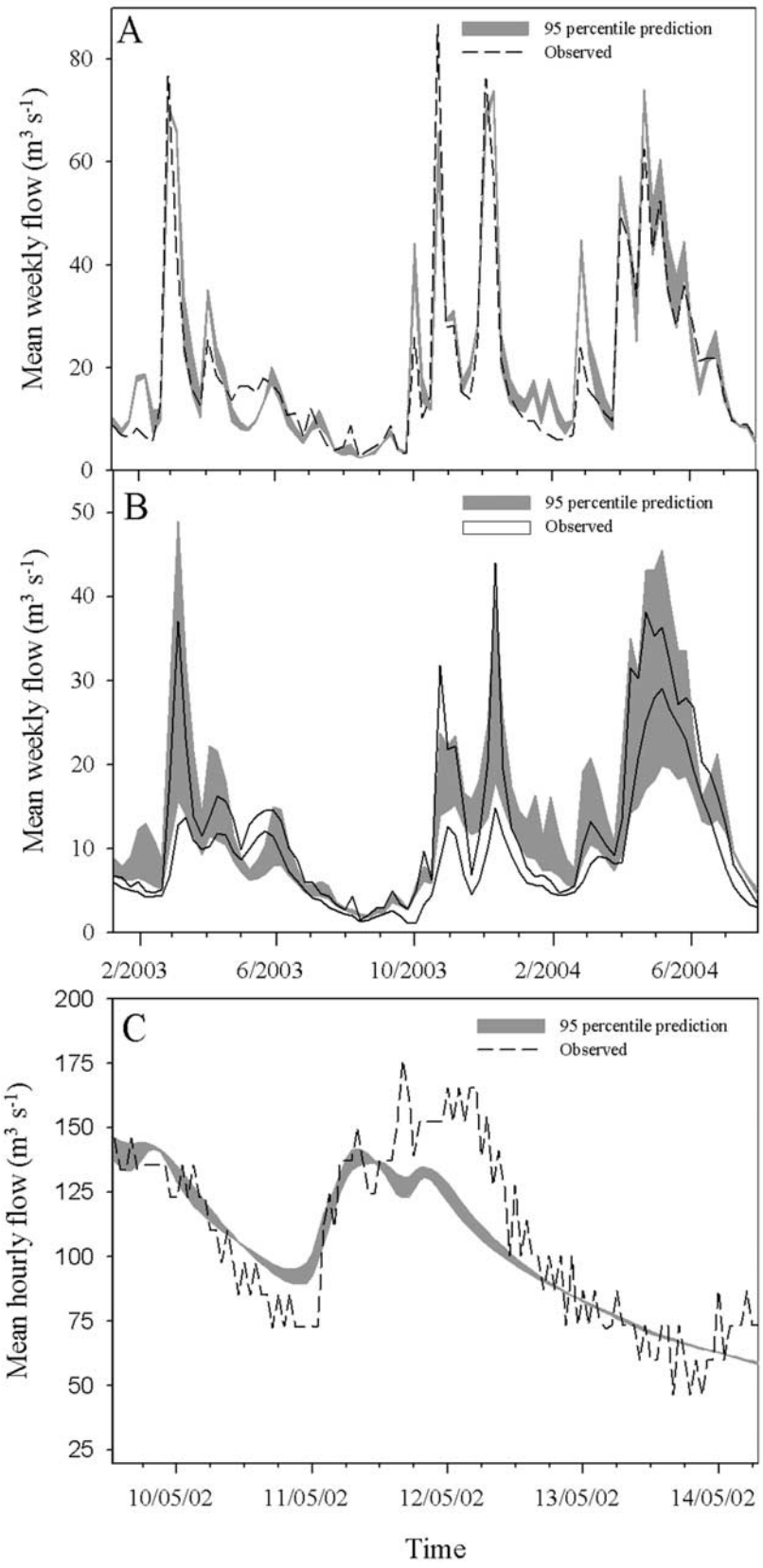

Figure 4. (a) WEEK and (b) BASE observed values and prediction during the validation period. (c) HOUR observed values and prediction during the calibration period.

parameters defined by lithology). This was achieved by evaluating the distributions overlapping with a 1000 bootstrap resampling procedure, in which the probability of a point from one distribution lying inside the 95 percentile of the other distribution was computed.

6.3. Relevance of Data Types on the Calibration Process

[44] The fact that TR-GML calibration runs demanded only a moderate amount of HSPF runs (around 10,000) allowed us to apply a stepwise procedure to assess the influence of the inclusion of different data on the calibration process. We simply performed independent calibration runs with different number of responses included in the OF. Since we had 6 different responses, showing every possible sequence of OF construction (from one response to six responses) would imply 720 calibration runs (or 10^6 model runs, or almost one year of computation time). Therefore we applied a rather arbitrary criterion to reduce the number of calibration runs tested to an affordable one and to define just one sequence of OF construction.

[45] Taking the most simple regularized calibration run as a reference (i.e., including only the WEEK series and REGU in the OF. This run will be referred hereafter as 2f), we performed independent TR-GML calibration runs including one additional response at a time (e.g., 2f + TRIM, 2f + BASE, etc.). It is worthy to mention that these calibration runs were totally independent of results generated from calibration run 2f, i.e., every calibration run was started independently following the methodology described in section 5.3, without regard to the optimum found in the previous step. Once the calibration runs were completed, we calculated the NS coefficient for each component of the OF for each calibration run, and then we standardized the results obtained with 2f:

$$SNS_c^r = \frac{NS_c^r - NS_c^{2f}}{s_c} \tag{7}$$

where *SNS* is the standardized NS calculated with the *c* component of the OF (c corresponds to WEEK, HOUR, etc) for the calibration run *r* (*r* corresponds to 2f + TRIM, 2f + HOUR, etc), and *s* is the standard deviation of all NSs calculated for the *c* component of the OF. We standardized the NSs because the variability across runs was very different between OF components, and we were interested in an aggregate measure of performance. Next, for each calibration run the mean of the standardized NSs was calculated (excluding REGU and the response included during the run). Therefore this mean expresses the effect of the inclusion of a new response in the OF on the fit of the other modeled responses. The run with the highest mean was taken as the new starting point for the next step (this run was called 3f). Then, we performed calibration runs including one response at a time (e.g., 3f + TRIM, 3f + BASE, etc), and solved equation (6) (always taking run 2f as the reference) to define run 4f. The process was repeated until all responses were included in the OF (i.e., the run 6f).

6.4. Relevance of Data Types on the Uncertainty of Base Flow Calculations

[46] Finally, the effect of the inclusion of different data in the OF on the uncertainty in the modeled base flow was assessed with the predictive analysis capability of the PEST package, using results from calibration runs 2f, 3f, 4f, 5f, and 6f obtained in the preceding section. For each run, we calculated 95% Schefeé non-linear predictive confidence intervals associated to the modeled base flow solving equation (5). In this case, the prediction to maximize/minimize was the base flow volume during the validation period.

7. Results

7.1. General Performance

[47] Table 2 shows the NSs calculated for all the calibration runs and the NS for the prediction of 6f compared to validation data. Focusing on 6f results, both for calibration and validation experiments, the NSs are in the range observed in many other studies. Only the NS for HOUR shows a value considerably lower than for the other groups (except the regularization group REGU). However, observing Figures 4 and 5, it is clear that the major features present in the observed data are represented in the model with enough detail, including the HOUR response. Modeled values uncertainty did not bracket many observed values, suggesting that error sources other than parameter uncertainty are present, as is usual in any complex hydrological application [*Vrugt et al.*, 2005].

7.2. Spatial Distribution of Adjustable Parameters

[48] Taking again the results obtained from the 6f calibration run, we tabulated the sensitivities of all the adjustable parameters with respect to the different components of the OF and the composite, the final optimized values, and their uncertainties (Table 3). From the composite sensitivity, we can observe that the most sensitive parameters are those related to the nominal storages (LZSN and UZSN), the losses of the system via different pathways (LZETP and DEEPFR), and the shape of the recession curves (AGWRC). By contrast, parameters related to the interflow had no effect on the OF. More interestingly, there were some parameters that showed a significant spatial variability. Whereas the storage parameters did not show any significant variation across the watershed feature considered (i.e., the distribution of final parameter values was highly overlapped between parameters pertaining to the same model parameter), for the infiltration driving parameters (INFILT and INFEXP), recession parameters, evapotranspiration, and deep percolation values significant spatial variability was evident across different basin attributes. Remarkably, all but one of the sensitive parameters assigned to different lithological types showed significant heterogeneity, especially L3 compared to the other two lithologies. Similarly, parameters defined by land use classes (LZETP) or groups (AGWRC) showed significant differences between them.

[49] Monthly variation in parameter values (expressed by UZSNΣ and LZETPΣ) was only supported for the evapotranspiration parameters, which showed a strong seasonal cycle (Table 3). Percentiles for UZSNΣ included zero, therefore no seasonal variation in this nominal storage is supported by the calibration data.

7.3. Relevance of Data Types on the Calibration Process

[50] Following the stepwise procedure detailed in section 6.3, the first response added to the 2f calibration

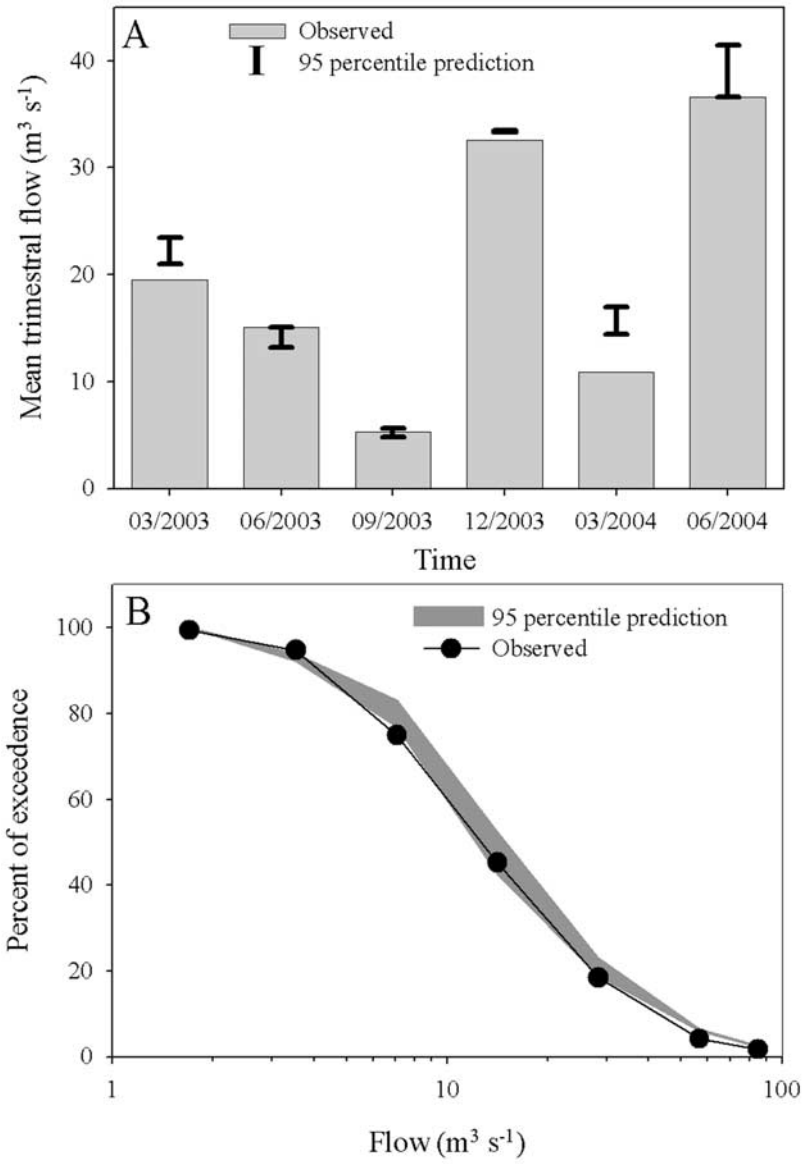

Figure 5. (a) TRIM and (b) DURC observed values and prediction during the validation period.

setting was HOUR, followed by TRIM, BASE, and finally DURC (see Table 2 for detailed calculations). Note that the order of response inclusion was in close relationship with the trade-offs showed by the different responses (Figure 6). Since when evaluating the inclusion of a response we considered the standardized NS values for the other responses, responses showing considerable trade-offs (e.g., BASE, DURC) were the last signals included in the OF.

[51] Considering the standardized NSs for the 6f run, the most important differences with respect to the 2f run were those showed by HOUR and DURC, and only the fit with BASE was marginally inferior. The final solution was in fact in the region of a balanced optimum Pareto solution (see Figure 6). As expected, only REGU final results were far from the optimum, showing substantial trade-offs with most of the other responses (Figure 6). Thus it can be stated that the inclusion of different data types into the calibration process exerted a positive effect on the performance of the model. However, although the inclusion of a response in the OF motivated a positive effect in the model performance for the same response, no clear pattern can be depicted about the effect on the other responses. For instance, the effect of the inclusion of DURC into 2f over the fit to TRIM was negative (see Table 2) despite the relatively high correlation between these two responses (Figure 6). The same response added to 3f (when HOUR was already included in the OF) exerted a positive effect on TRIM. Similarly, the effect of including additional responses into 3f was notoriously negative (see the Mean field in Table 2), but once TRIM was added to the OF, the inclusion of BASE exerted a very positive effect on the fit of the other responses. Again, once BASE was inside the OF, the inclusion of DURC was positive for the rest of OF components compared to the 2f run, despite the inclusion of this response being consistently negative before BASE was in the OF. All these indicate that the inclusion of additional responses in an aggregated, single OF has a profound effect in its multidimensional shape, and consequently in the whole calibration process.

Table 2. Response-Specific Nash-Sutcliffe Coefficients for the Different Calibration Runs

Objective Function Components	WEEK	HOUR	TRIM	BASE	DURC	REGU	
			Nash-Sutcliffe Coefficient (Calibration)				
WEEK + REGU (2f)	0.974	0.223	0.959	0.979	0.9998254	0.931	
2f + TRIM	0.972	0.261	0.979	0.978	0.9999784	0.733	
2f + DURC	0.968	0.641	0.957	0.968	0.9999997	0.739	
2f + BASE	0.972	0.139	0.939	0.993	0.9997346	0.803	
2f + HOUR (3f)	0.975	0.886	0.958	0.976	0.9999183	0.674	
3f + DURC	0.968	0.880	0.961	0.963	0.9999994	0.573	
3f + BASE	0.971	0.857	0.942	0.991	0.9999407	0.682	
3f + TRIM (4f)	0.972	0.872	0.980	0.950	0.9999674	0.711	
4f + DURC	0.971	0.852	0.975	0.938	0.9999989	0.072	
4f + BASE (5f)	0.976	0.866	0.975	0.981	0.9999760	0.539	
5f + DURC (6f)	0.975	0.835	0.976	0.977	0.9999989	0.081	
			Standard Deviation of Nash-Sutcliffe Coefficient				
	0.003	0.302	0.015	0.017	0.0000856		
			Nash-Sutcliffe Coefficient Standardized by 2f (equation (6))				*Mean*[a]
WEEK + REGU (2f)	0	0	0	0	0		
2f + TRIM	−0.50	0.12	1.33	−0.07	1.79		0.34
2f + DURC	−1.98	1.38	−0.13	−0.68	2.04		−0.35
2f + BASE	−0.75	−0.28	−1.42	0.85	−1.06		−0.88
2f + HOUR (3f)	0.64	2.19	−0.10	−0.17	1.08		0.36
3f + DURC	−2.19	2.17	0.11	−0.97	2.03		−1.02
3f + BASE	−0.94	2.09	−1.19	0.74	1.35		−0.26
3f + TRIM (4f)	−0.66	2.15	1.45	−1.74	1.66		−0.25
4f + DURC	−0.92	2.08	1.06	−2.46	2.03		−0.43
4f + BASE (5f)	0.81	2.12	1.09	0.13	1.76		1.56
5f + DURC (6f)	0.60	2.02	1.17	−0.11	2.03		0.60
			Nash-Sutcliffe Coefficient (Validation)				
6f	0.942	na	0.973	0.968	0.99977	0.081	

[a]Mean of the Nash-Sutcliffe Coefficients for the corresponding calibration run, excluding the coefficient corresponding to the response included during the run. na = not applicable.

[52] The inclusion of data types in the OF also affected the regularization component of the OF (REGU). Table 2 shows that NS for REGU descended as more responses were added to the OF, which means that a moderate additional amount of spatial heterogeneity in model parameters was possible thanks to the new information added to the calibration process. However, the most striking effect was that shown by the inclusion of DURC, once HOUR and TRIM were already included in the OF. Then, the fit to REGU descended by an order of magnitude, implying that a huge amount of spatial and temporal heterogeneity in model parameters was necessary to maintain an acceptable level of fit. This again stresses the fact that the shape of the aggregated OF in the hypercube defined by the number of parameters undergoes very profound changes as new data are included into the calibration process.

[53] The central role played by DURC in the calibration process was also highlighted from the response-specific parameter sensitivities during run 6f (Table 3). This response was sensible to variations of several parameters, including nominal storages, infiltration drivers, recession curves, evapotranspiration losses, and deep percolation. TRIM was also sensibly affected by the same parameters, excluding those related to infiltration processes. As expected, parameters related to recession and groundwater routing (AGWRC, LZSN, LZETP, and DEEPFR) showed high sensitivity for BASE. The short-term surface responses (HOUR and WEEK) were sensitive to several parameters.

7.4. Relevance of Data Types on the Uncertainty of Base Flow Calculations

[54] The inclusion of several responses into the OF is intended to constrain the possible values that the adjustable parameters can take. Thus it was expected that the range of the estimated parameter uncertainty, as implemented in PEST, would reduce as more data were included in the OF. Consequently, we calculated the range of uncertainty for the 25 composite-sensitive parameters (Table 3), taking the 95% percentile limits calculated by PEST and normalizing by the predefined parameter limits (see Table 1). Figure 7 shows the results for five calibration runs expressed as box-plots that describe the distribution of the normalized ranges for the 25 parameters. The results during the first inclusions (3f to 4f) did not support the expectation, because indeed parameter ranges globally rose. Only after the inclusion of DURC (i.e., 6f) were the uncertainties globally below the values obtained during run 2f.

[55] However, the uncertainty of the modeled mean base flow obtained during the validation period supports the idea that the inclusion of several responses in the OF had beneficial consequences (Figure 8). The effect on base flow uncertainty was conspicuous in the minimum values, especially in the steps that defined 3f and 6f. Also interesting, the mean value for 2f is probably biased, if it is compared to the mean values for the other runs, established at higher values. Thus the inclusion of responses in the OF not only

Table 3. Parameter Sensitivity and Uncertainty for the Run 6f

	Sensitivities[a]							Percentiles	
Parameter	WEEK	HOUR	TRIM	BASE	DURC	Composite	Optimized Value[b]	2.5	97.5
$LZSN^{L1}$	0.04	0.03	**0.13**	0.03	**0.20**	0.087	3.46	2.13	5.26
$LZSN^{L2}$	0.03	0.03	**0.11**	0.02	**0.19**	0.077	5.96	3.59	9.40
$LZSN^{L3}$	0.04	0.03	**0.14**	**0.05**	**0.20**	0.093	5.57	3.83	7.85
$UZSN^{S1}$	0.05	0.04	**0.11**	0.02	**0.36**	**0.115**	0.018	0.01[d]	0.172
$UZSN^{S2}$	0.03	0.02	**0.09**	···	**0.19**	**0.068**	0.040	0.01[d]	0.460
$UZSN^{S3}$	**0.32**	**0.11**	**0.71**	0.08	**1.84**	**0.612**	0.032	0.01[d]	0.061
$UZSN\Sigma^{S1}$	···	···	···	···	···	···	0.004	0 [d]	0.432
$INFILT^{L1}$	0.01	0.02	0.02	0.01	**0.09**	0.033	0.17	0.09	0.30
$INFILT^{L2}$	0.02	0.03	0.03	0.02	**0.13**	0.047	0.09	0.05	0.16
$INFILT^{L3}$	0.01	0.02	0.03	0.01	**0.09**	0.033	$*^{(L1)}$ $***^{(L2)}$ 0.38	0.23	0.50[d]
$INFEXP^{L1}$	···	0.02	0.01	···	0.03	0.016	4.72	2.63	7.98
$INFEXP^{L2}$	0.01	0.01	0.02	···	0.04	0.019	3.78	2.28	5.95
$INFEXP^{L3}$	···	0.02	0.02	···	0.03	0.016	$**^{(L1)}$ $\S^{(L2)}$ 1.81	0.81	3.34
$KVARY^{L1}$	0.01	0.01	0.02	0.02	**0.05**	0.022	0[c]	0 [d]	0.29
$KVARY^{L2}$	0.01	0.01	0.02	0.02	0.04	0.019	0.18	0 [d]	0.75
$KVARY^{L3}$	0.03	0.04	**0.06**	**0.05**	**0.10**	**0.056**	$\S^{(L1)}$ 0.38	0.10	0.77
$AGWRC^{G1}$	**0.31**	**0.16**	**0.61**	**0.55**	**0.63**	**0.452**	$**^{(all)}$ 0.980	0.975	0.986
$AGWRC^{G2}$	**0.11**	**0.08**	**0.20**	**0.18**	**0.34**	**0.182**	** 0.959	0.945	0.976
IRC^{S1}	···	···	···	···	···	···	0.53	0.39	0.85[d]
IRC^{S2}	···	···	···	···	···	···	0.51	0.39	0.82
IRC^{S3}	0.01	0.04	···	···	0.03	0.016	0.50	0.43	0.60
$INTFW^{L1}$	···	···	···	···	···	···	7.04	1[d]	10[d]
$INTFW^{L2}$	···	···	···	···	···	···	7.23	1[d]	10[d]
$INTFW^{L3}$	···	···	···	···	···	···	6.88	1[d]	10[d]
$LZETP^{CR}$	0.02	0.01	**0.06**	0.03	**0.09**	0.041	0.61	0.43	0.89
$LZETP^{MX}$	0.04	0.03	**0.14**	**0.06**	**0.18**	**0.090**	$\S^{(CR,\ CF)}$ 0.89	0.68	0.9[d]
$LZETP^{DC}$	**0.05**	0.02	**0.15**	0.04	**0.18**	**0.088**	$***^{(all)}$ 0.31	0.24	0.39
$LZETP^{CF}$	0.04	0.02	**0.11**	0.04	**0.19**	**0.080**	0.63	0.48	0.83
$LZETP^{BL}$	···	···	0.01	···	···	···	0.75	0.46	0.9[d]
$LZETP\Sigma$	**0.05**	0.01	**0.13**	**0.05**	**0.07**	**0.063**	1.0[c]	0.83	1[d]
$BASETP^{DC}$	···	···	0.02	0.01	0.03	0.014	0.04	0[d]	0.17
$DEEPFR^{L1}$	···	···	0.01	···	0.02	···	0.05	0[d]	0.16
$DEEPFR^{L2}$	···	···	0.01	···	0.02	···	0.06	0.001	0.18
$DEEPFR^{L3}$	0.04	0.04	**0.12**	**0.05**	**0.21**	**0.092**	$***^{(all)}$ 0.34	0.24	0.47

[a]Sensitivities < 0.01 are not shown, those above 0.05 are in bold case. Composite sensitivities above 0.1 are underlined.
[b]Significant paired differences between spatially distributed parameters are denoted by: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, and § $p < 0.1$. Within brackets, the index for the significantly different parameter.
[c]Parameter optimized on a limit (see Table 1).
[d]Percentile at the limit of the predefined range for the parameter (see Table 1).

reduced the uncertainty of calculations, but also added accuracy to the results.

8. Discussion and Conclusions

[56] The final distribution of optimized parameters clearly showed that maintaining some spatial heterogeneity in the model parameterization was justified. Several parameters distributed by both lithology and land uses showed significant differences, although those linked to the slope classification did not show any spatial heterogeneity. This poses the question whether an a priori definition of intended homogeneous zones is the best way to impose some spatial resolution to parameter values. Despite the fact that this approach worked well in this study, a more convenient method will involve the definition of as many parameters as HRUs the model has. Of course, the application of such a method in a complex watershed-scale model demands appropriate numerical tools to deal with the thousands of parameters that will be present. Genetic algorithms [*Wang*, 1991] could be a good option to deal with such a calibration exercise, but considering the time one HSPF run takes, its applicability in our problem is dubious. An alternative is the hybrid regularized inversion by *Tonkin and Doherty* [2005], in which after a time consuming initial calibration run the eigenvectors representing principal orthogonal directions in parameter space are used to define a feasible number of super parameters. Then, these super parameters can be optimized using GML. In principle, we could use this method to parameterize the system and also in searching for the relationships between parameter values and watershed features. This could be a powerful way to assess the feasibility of the final parameter values and to understand the basin's hydrological cycle. However, the hybrid regularized inversion is based on parameter sensitivities, which in a complex hydrological model can be dependent on the value around which sensitivities are calculated. Thus a good procedure could be using an a priori definition of homogeneous zones to search for a suitable initial parameters set to feed the hybrid inversion. Although the results from this study could be a good starting point to test this possibility, this is beyond the scope of this work.

[57] Independently of the parameter zonation methodology applied, a proper prior information scheme is of a paramount importance for the success of a regularized calibration process [*Tonkin and Doherty*, 2005]. Despite using an extremely simple regularization methodology, the calibration runs were numerically stable and the final parameter values and ranges (Table 3) were reasonable. For instance, although the prior information scheme imposed homogeneity and economy criteria, the inversion

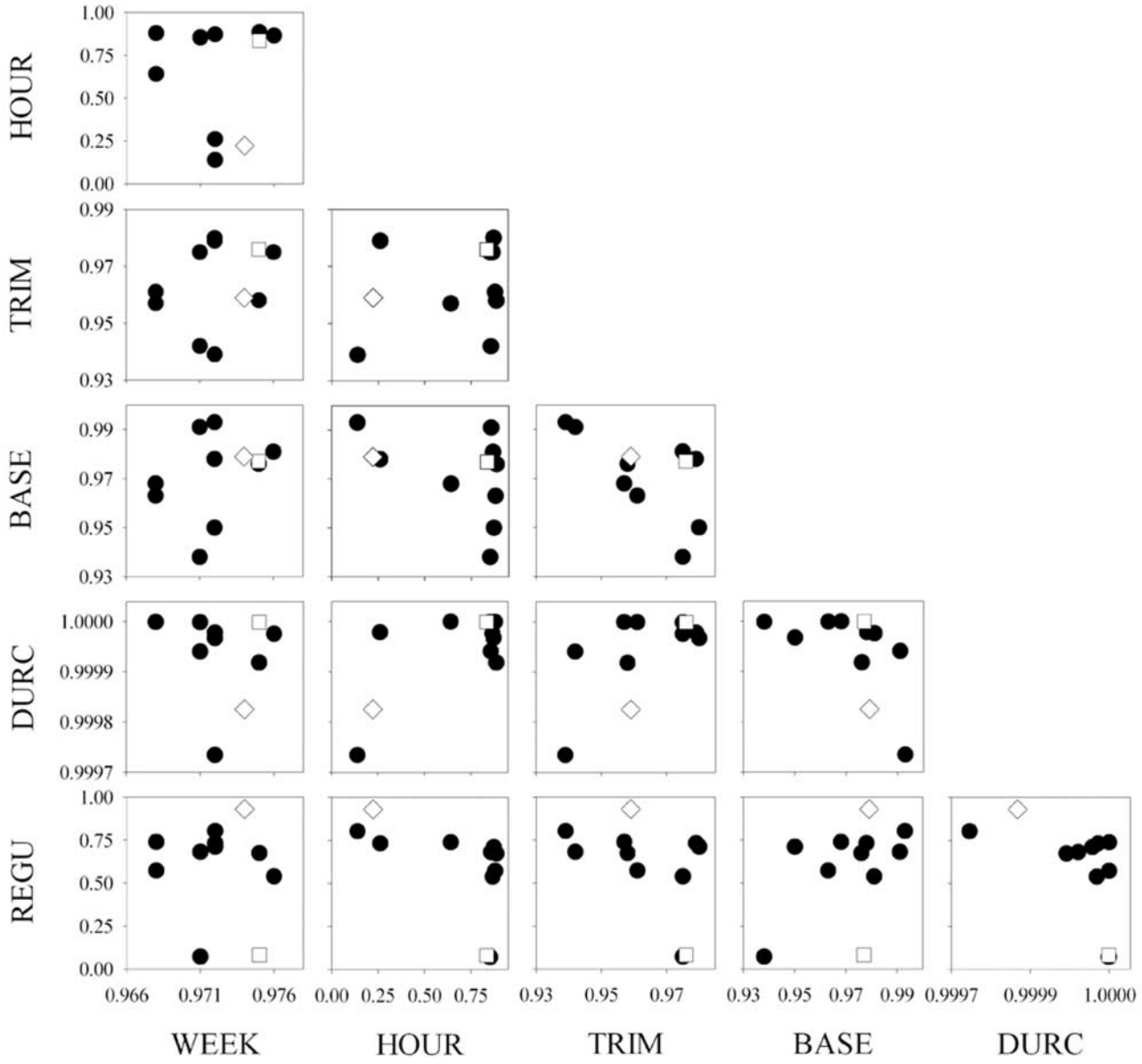

Figure 6. Trade-off plots for the different responses included in the OF, expressed as NS coefficients. Each dot corresponds to a calibration run detailed in Table 2. Diamonds are run 2f, and white squares correspond to run 6f.

process assigned a positive DEEPFR value for the hard rock, fractured area, while maintaining the other areas not significantly different from zero. Similarly, the LZETP value for the deciduous forest was significantly lower than that for other forestlands, as is expected [*Swift et al.*, 1975]. The seasonality of this parameter, expressed by LZETPΣ, was also very pronounced, indicating a strong seasonality in evapotranspiration losses, a usual result for Mediterranean areas [*Bernal et al.*, 2004]. It is also worthy to mention the significant differences found in the INFEXP parameters (Table 3). Since a value significantly greater than 2 prompts HSPF to switch from an infiltration-excess runoff generation to a saturation-excess one [*Berris*, 1995], we can state that there is a very different hydrological behavior in the different lithological zones of our watershed. The high altitude mountainous zone is governed by a Hortonian overland flow, while zones dominated by sedimentary rocks and alluvial deposits show saturation overland flow. This result agrees with the view of *Johnson et al.* [2003], who argued against the common practice of HSPF modelers to maintain the default INFEXP value (i.e., 2) during simulations. All the above significant differences between spatially discretized parameters clearly shows that a rich variety of hydrological behaviors can be present in a complex watershed and that a proper regularized inversion is a powerful methodology to help solving the calibration step.

[58] Although the aggregation of several responses in the OF was helpful during the calibration process, it is certainly difficult to predict what kind of data will be most worthy to include in the OF. Apart from the main target (e.g., WEEK in this study), the inclusion of nonparametric descriptors

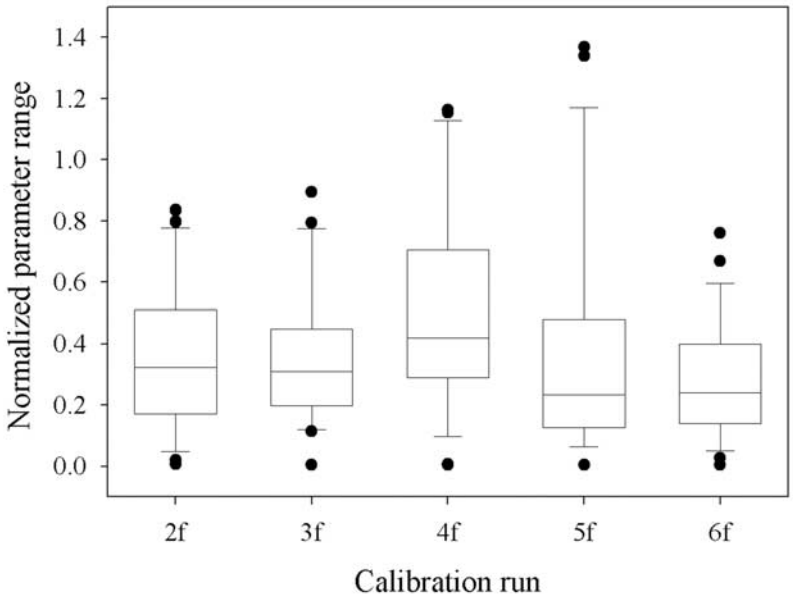

Figure 7. Box-whisker plots of the normalized range of the 25 sensitive parameters after completion of five calibration runs (middle line, median; lower box line, first quartile; upper box line, third quartile; lower whisker line, 2.5 percentile; upper whisker line, 97.5 percentile; dots denote individual outliers).

like DURC is judged conveniently [*Shamir et al.*, 2005; *Montanari and Toth*, 2007]. However, as we demonstrated in this study, the beneficial effects of a response can arise only after the inclusion of other data. Thus insertion of a response in the OF does not simply add observations to the calibration problem, but can also change the shape of the OF and facilitate finding of the overall OF minimum. Of course, the importance of this effect will depend on the searching capabilities of the calibration algorithm used and the correlation of the different responses included in the OF (i.e., a perfect calibration algorithm should find the overall OF minimum for a set of responses included in the OF, irrespective of other data available and not already included in the OF). Considering the good performance shown by TR-GML in this study, this should be interpreted in the way that, although state-of-the-art calibration methods are powerful tools to calibrate hydrological models, as the dimension of the problem increases a proper data management strategy is of fundamental importance to help the calibration process. This is because the multidimensional shape of the OF will be a challenge even for evolutionary algorithms, especially if model runs take more than few seconds and the

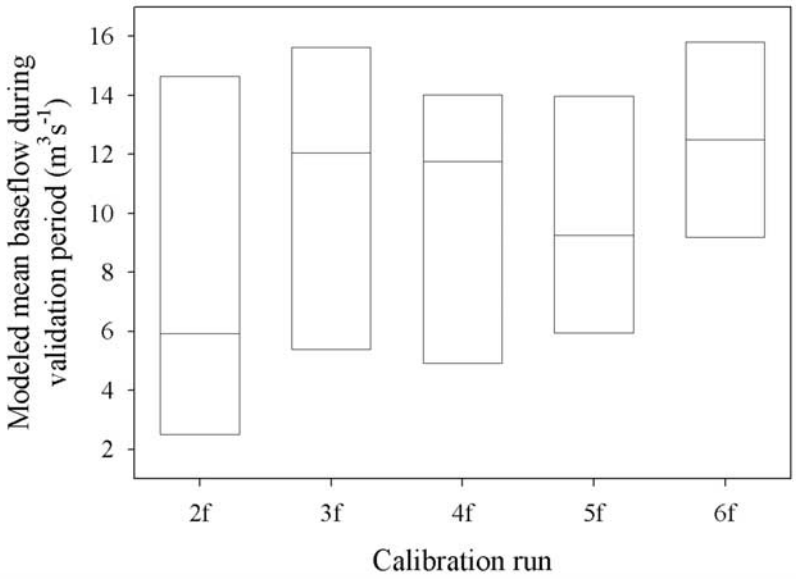

Figure 8. Modeled mean base flow and associated uncertainty during the validation period after completion of five calibration runs (middle line, median; lower box line, 2.5 percentile; upper box line, 97.5 percentile).

available computing power is limited. Results obtained using SCE-UA in this study appear to point in this direction.

[59] Thus we recommend introducing in the OF as much data as available, including by-products of sampled data (e.g., TRIM and BASE). However, since it's not guaranteed that the topological complexity of an aggregated OF will decrease as more responses are included, algorithms that search for the Pareto set of parameters [*Gupta et al.*, 1998; *Vrugt et al.*, 2003; *Khu and Madsen*, 2005] without aggregating responses should be preferred a priori. The applicability of these methodologies will depend on the computational burden necessary to reach convergence, which can be very demanding for complex watershed-scale hydrological models. In addition, the use of data mathematically derived from observed data is not free of risk. In our case, BASE was built with an algorithm that assumes a constant dynamics for the recession curve. Considering that HSPF includes a parameter (KVARY) to model time-varying recession curves, the use of a series like BASE could artificially impose zero values for KVARY. Although it seems that this was not the case for our calibration, since $KVARY^{L3}$ was optimized to a non-zero value (Table 3), the presence of this significant non-zero value for KVARY is not in accordance with the assumptions used to build the BASE series. Thus caution should be taken when including non-observational data into the OF. In principle, temporal aggregation (e.g., TRIM) and non-parametric values (e.g., DURC) are best options [*Shamir et al.*, 2005; *Montanari and Toth*, 2007].

[60] Obviously, the criteria used in this paper only suggested one of the 720 possibilities of sequential construction of the OF, and probably the order of responses inclusion would change in a different scenario (another calibration period, watershed, or model structure).

[61] Thus we were only interested in demonstrating the profound effect of the inclusion of different data types on the calibration process, with no aim to extrapolate our results to propose general rules for the inclusion of data in the OF.

[62] In any case, calibrating a complex watershed-scale model using only one response is a very dangerous procedure, even if good prior information is available. As depicted in Figure 8, the uncertainty associated with water routing was very high for the 2f run, and the estimated values were probably biased. For a water quality application this would be critical, especially if management decisions lay on the modeling outcomes, because this lack of precision and accuracy produces useless results despite the goodness of fit attained with river runoff.

[63] In conclusion, combining regularized inversion and a multiresponse OF useful results can be obtained even with complex applications including spatial and temporal heterogeneity in model parameters. Since good numerical methods appropriate for these purposes are now available [e.g., *Tonkin and Doherty*, 2005], maintaining the observed basin heterogeneity in the model abstraction should always be considered. In fact, the inclusion of the inherent heterogeneity present in real catchments into hydrological models is an old concern that still is a matter of debate [e.g., *Jakeman and Hornberger*, 1993; *Beven*, 1996; *Boyle et al.*, 2001; *Schulz et al.*, 2006], and this concern should be transferred to water quality models even though hydrology is not the main target in these applications. The management of field observations has also captured the attention of hydrologists [*Kuczera*, 1982; *Raat et al.*, 2004], and the results from this work emphasize that when an OF aggregating different data is the only practical option, caution should be taken when including and excluding responses. It must be emphasized that regularized inversion and the other calibration strategies used throughout this study could also be combined with OFs consisting on multisite and multivariable data. If the model structure is complex, the approximations outlined in this paper, combined with other tools developed to calibrate complex models, like prior sensitivity analysis [*Zheng and Keller*, 2006] and inclusion of water quality constituents as chemical tracers [*Bernal et al.*, 2004], can help obtaining reliable model outcomes even in complex, real world problems.

[64] **Acknowledgments.** The Agència Catalana de l'Aigua (ACA), Aigües Ter Llobregat (ATLL), and Servei Meteorològic de Catalunya (METEOCAT) generously provided most of the data used in this study. R.M. is a pre-doctoral grantee of the Spanish Government. This material is based upon work supported by the European Research Office of the US Army under Contract No. N62558-02-M-6007. We thank Pat Deliman for advice on HSPF modeling. Comments from three anonymous reviewers greatly improved the quality of this paper.

References

Agència Catalana de l'Aigua (ACA) (2001), Delimitació de zones inundables per a la redacció de l'INUNCAT, Departament de Medi Ambient, Generalitat de Catalunya, Barcelona.

Alexander, R. B., A. H. Elliott, U. Shankar, and G. B. McBride (2002), Estimating the sources and transport of nutrients in the Waikato River basin, New Zealand, *Water Resour. Res.*, *38*(12), 1268, doi:10.1029/2001WR000878.

Arnold, J. G., and P. M. Allen (1999), Automated methods for estimating baseflow and ground water recharge from streamflow records, *J. Am. Water Resour. As.*, *35*(2), 411–424.

Bellot, J., and A. Escarré (1998), Stemflow and throughfall determination in a resprouted Mediterranean holm-oak forest, *Ann. Forest Sci.*, *55*(7), 847–865.

Bernal, S., A. Butturini, J. L. Riera, E. Vázquez, and F. Sabater (2004), Calibration of the INCA model in a Mediterranean forested catchment: the effect of hydrological inter-annual variability in an intermittent stream, *Hydrol. Earth Syst. Sci.*, *8*(4), 729–741.

Berris, S. N. (1995), Conceptualization and simulation of runoff generation from rainfall for three basins in Thurston County, Washington, U.S. Geological Survey, Water Resources Investigation Report 94-4038, Prepared in Cooperation with Thurston County Department of Public Works, USGS, Tacoma, WA.

Beven, K. (1996), The limits of splitting: hydrology, *Sci. Total Environ.*, *183*, 89–97.

Bicknell, B. R., J. C. Imhoff, J. L. Kittle, T. H. Jobes, and A. S. Donigian (2001), Hydrological Simulation Program-Fortran (HSPF) user's manual for release 12, U. S. Environmental Protection Agency, National Exposure Research Laboratory, Athens, GA.

Boyle, D. P., H. V. Gupta, S. Sorooshian, V. Koren, Z. Zhang, and M. Smith (2001), Toward improved streamflow forescasts: value of semidistributed modeling, *Water Resour. Res.*, *37*(11), 2749–2759.

Butturini, A., and F. Sabater (2000), Seasonal variability of dissolved organic carbon in a Mediterranean stream, *Biogeochemistry*, *51*(3), 303–321, doi:10.1023/A:1006420229411.

Centre de Recerca Ecològica i Aplicacions Forestals (CREAF) (2001), Inventari Ecològic i Forestal de Catalunya, Centre de Recerca Ecològica i Aplicacions Forestals, Universitat Autònoma de Barcelona, Bellaterra, Spain.

Crawford, H. H., and R. K. Linsley (1966), Digital Simulation in Hydrology: Stanford Watershed Model IV. Technical Report No. 39, Dept. of Civil Eng., Stanford University, Stanford, CA.

Doherty, J. (2004), PEST Model-independent Parameter Estimation, User Manual: 5th edition, Watermark Numerical Computing, Brisbane, Australia.

Doherty, J., and J. M. Johnston (2003), Methodologies for calibration and predictive analysis of a watershed model, *J. Am. Water Resour. As.*, *39*(2), 251–265.

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

Gallagher, M., and J. Doherty (2007), Parameter estimation and uncertainty analysis for a watershed model, *Environmental Modelling and Software*, *22*, 1000–1020.

Gallart, F., P. Llorens, J. Latron, and D. Regues (2002), Hydrological processes and their seasonal controls in a small Mediterranean mountain catchment in the Pyrenees, *Hydrol. Earth Syst. Sci.*, *6*(3), 527–537.

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

Hunt, R. J., D. T. Feinstein, C. D. Pint, and M. P. Anderson (2006), The importance of diverse data types to calibrate a watershed model of the Trout Lake Basin, Northern Wisconsin, USA, *J. Hydrol.*, *321*, 286–296.

Instituto Nacional para la Conservación de la Naturaleza (ICONA) (1990), Mapas de estados erosivos. Cuenca hidrográfica del Pirineo Oriental, Ministerio de Agricultura, Pesca y Alimentación, Madrid, Spain.

Jakeman, A. J., and G. M. Hornberger (1993), How much complexity is warranted in a rainfall-runoff model?, *Water Resour. Res.*, *29*(8), 2637–2649.

Johnson, M. S., W. F. Coon, V. K. Mehta, T. S. Steenhius, E. S. Brooks, and J. Boll (2003), Application of two hydrologic models with different runoff mechanisms to a hillslope dominated watershed in the northeastern US: a comparison of HSPF and SMR, *J. Hydrol.*, *284*, 57–76.

Kannan, N., S. M. White, F. Worrall, and M. J. Whelan (2007), Hydrological modelling of a small catchment using SWAT-2 000. Ensuring correct flow partitioning for contaminant modelling, *J. Hydrol.*, *334*, 64–72.

Khu, S. T., and H. Madsen (2005), Multiobjective calibration with Pareto preference ordering: an application to rainfall-runoff model calibration, *Water Resour. Res.*, *41*, W03004, doi:10.1029/2004WR003041.

Kuczera, G. (1982), On the relationship between the reliability of parameter estimates and hydrologic time series data used in calibration, *Water Resour. Res.*, *18*(1), 146–154.

Legates, D. R., and G. J. McCabe Jr. (1999), Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation, *Water Resour. Res.*, *35*(1), 233–241.

Liu, Y., and H. V. Gupta (2007), Uncertainty in hydrologic modeling: toward an integrated data assimilation framework, *Water Resour. Res.*, *43*, W07401, doi:10.1029/2006WR005756.

Llorens, P. (1997), Rainfall interception by Pinus sylvestris forest patch overgrown in a Mediterranean mountainous abandoned area II. Assessment of the applicability of Gash's analytical model, *J. Hydrol.*, *199*, 346–359.

Llorens, P., R. Poch, J. Latron, and F. Gallart (1997), Rainfall interception by Pinus sylvestris forest patch overgrown in a Mediterranean mountainous abandoned area I. Monitoring design and results down to the event scale, *J. Hydrol.*, *199*, 331–345.

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 modeling using automatic calibration with multiple objectives, *Adv. Water Resour.*, *26*, 205–216.

Montanari, A., and E. Toth (2007), Calibration of hydrological models in the spectral domain: an opportunity for scarcely gauged basins?, *Water Resour. Res.*, *43*, W05434, doi:10.1029/2006WR005184.

Moore, C., and J. Doherty (2005), Role of the calibration process in reducing model predictive error, *Water Resour. Res.*, *41*, W05020, doi:10.1029/2004WR003501.

Raat, K. J., J. A. Vrugt, W. Bouten, and A. Tietema (2004), Towards reduced uncertainty in catchment nitrogen modelling: quantifying the effect of field observation uncertainty on model calibration, *Hydrol. Earth Syst. Sci.*, *8*(4), 751–763.

Rodà, F., J. Retana, C. A. Gracia, and J. Bellot (Eds.) (1999), Ecology of Mediterranean evergreen oak forests, Ecological Studies 137, Springer-Verlag, Berlin.

Rorabaugh, M. I. (1964), Estimating changes in bank storage and groundwater contribution to streamflow, *International Association of Scientific Hydrology*, *63*, 432–441.

Schulz, K., R. Seppelt, E. Zehe, H. J. Vogel, and S. Attinger (2006), Importance of spatial structures in advancing hydrological sciences, *Water Resour. Res.*, *42*, W03S03, doi:10.1029/2005WR004301.

Shamir, E., B. Iman, H. V. Gupta, and S. Sorooshian (2005), Application of temporal descriptors in hydrologic model parameter estimation, *Water Resour. Res.*, *41*, W06021, doi:10.1029/2004WR003409.

Shohomish County Surface Water Management (SWM) (2002), Swamp Creek drainage needs report, Department of Public Works, Surface Water Management Division, Everett, WA.

Singh, V. P., and D. A. Woolhiser (2002), Mathematical modeling of watershed hydrology, *J. Hydrol. Eng.*, *7*(4), 270–292.

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

Swift, L. W., Jr., W. T. Swank, J. B. Mankin, R. J. Luxmoore, and R. A. Goldstein (1975), Simulation of evapotranspiration and drainage from mature and clear-cut deciduous forest and young pine plantation, *Water Resour. Res.*, *11*(5), 667–673.

Tolson, B. A., and C. A. Shoemaker (2007), Dynamically dimensioned search algorithm for computationally efficient watershed model calibration, *Water Resour. Res.*, *43*, W01413, doi:10.1029/2005WR004723.

Tonkin, M. J., and J. Doherty (2005), A hybrid regularized inversion methodology for highly parameterized environmental models, *Water Resour. Res.*, *41*, W10412, doi:10.1029/2005WR003995.

U. S. Environmental Protection Agency (USEPA) (1978), User's manual for the Agricultural Runoff Management (ARM) model, EPA-823/3-78-080, U. S. Environmental Protection Agency, Environmental Research Laboratory, Athens, GA.

U. S. Environmental Protection Agency (USEPA) (1999), HSPFParm: an interactive database of HSPF model parameters, Version 1.0, EPA-823-R-99-004, U. S. Environmental Protection Agency, Office of Water, Washington, D. C.

U. S. Environmental Protection Agency (USEPA) (2000), BASINS Technical Note 6: estimating hydrology and hydraulic parameters for HSPF, EPA-823-R-00-012, U. S. Environmental Protection Agency, Office of Water, Washington, D. C.

U. S. Environmental Protection Agency (USEPA) (2001), Better Assessment Science Integrating point and Non-point Sources, BASINS Version 3.0, EPA-823-H-01-001, U. S. Environmental Protection Agency, Office of Water, Washington, D. C.

van der Leeden, F., F. L. Troise, and D. K. Todd (1991), *The Water Encyclopedia*, Second Edition, Lewis Publishers, Chelsea, MI.

Vecchia, A. V., and R. L. Cooley (1987), Simultaneous confidence and prediction intervals for nonlinear regression models with application to a groundwater flow model, *Water Resour. Res.*, *23*(7), 1237–1250.

Vogel, C. R. (2002), *Computational methods for inverse problems*, Soc. For Ind. And Appl. Math., Philadelphia, Pa.

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

Vrugt, J. A., C. G. H. Diks, H. V. Gupta, W. Bouten, and J. M. Verstraten (2005), Improved treatment of uncertainty in hydrologic modeling: combining the strengths of global optimization and data assimilation, *Water Resour. Res.*, *41*, W01017, doi:10.1029/2004WR003059.

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

Zheng, Y., and A. A. Keller (2006), Understanding parameter sensitivity and its management implications in watershed-scale water quality modeling, *Water Resour. Res.*, *42*, W05402, doi:10.1029/2005WR004539.

J. Armengol and R. Marcé, Fluvial Dynamics and Hydrologic Engineering (FLUMEN), Department of Ecology, University of Barcelona, Diagonal 645, 08028 Barcelona, Spain. (rafamarce@ub.edu)

C. E. Ruiz, CEERD-EP-W, U.S. Army Engineer Research and Development Center, 3909 Halls Ferry Road, Vicksburg, MS 39180-6199, USA.