

# Calibration of Hydrologic Models: When is a Model Calibrated?

<sup>1</sup>Abbaspour, K.C.

<sup>1</sup>Swiss Federal Institute for Environmental Science and Technology, E-Mail: [abbaspour@eawag.ch](mailto:abbaspour@eawag.ch)

**Keywords:** Inverse modeling, Calibration, Parameter uncertainty, Prediction uncertainty.

## EXTENDED ABSTRACT

Parameter calibration and prediction uncertainty of a model are intimately related. Although much literature is devoted to calibration and uncertainty analysis in hydrologic modelling, a clear definition of when a model is calibrated is generally lacking. In recent years, inverse modelling has become a popular alternative to direct parameter measurement. However, in contrast to obtaining model outputs (*model variables*) from model inputs (*model parameters*), which is unique, obtaining *model parameters* from *model variables* is non-unique by nature. Inversely obtained hydrologic parameters, therefore, are always uncertain (non-unique) because of parameter correlations, errors associated with the measurements, and Level of details and simplifications in the models, among other factors. Quantification of the uncertainty in a calibrated model is vital for a meaningful application of the model. Optimization of a goal (objective) function in a problem with many parameters is often difficult because of the complexities in the structure of the goal function. One complexity is the large number of local minima associated with any given goal function. To draw an analogy, the space of the goal function,  $g$ , could be likened to a block of “Swiss cheese” (if simplified to two parameters only) (Fig. 1) with many holes. Each hole represents a local minimum, with the size of the hole in any direction representing the range of uncertainty.



Figure 1. A block of Swiss cheese.

Figure 2 shows the “Swiss cheese” effect in the response surface of a goal function. In this Figure, for a better visualization, the goal function was inverted so that the minima are represented as peaks. It shows that for any given goal function, there exists many parameter sets for which the goal functions are not significantly different from each other, i.e., there are many potential solutions based on quite different parameter sets. Recently,

emphasis is being placed on identifying all such solutions. This is referred to as uncertainty analysis. As there are many potential solutions, each parameter optimization routine finds one such minimum for a given goal function. Hence, the search for one absolute global minimum in hydrologic problems, where the parameters are generally lumped, is not very meaningful. As the problem of parameter optimization is not unique, it is important that we define when a model is calibrated and what the magnitude of the prediction uncertainty is.

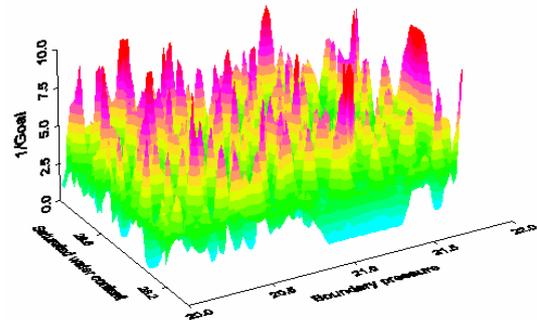


Figure 2. Response surface for an inverted goal function. Local minima are represented as peaks.

In the combined parameter estimation-uncertainty analysis routine, SUFI-2 (Sequential Uncertainty Fitting, ver. 2), we define model calibration as follows: A model is considered calibrated if 1) upon propagation of parameter uncertainties the 95% prediction uncertainty (95PPU) between the 2.5<sup>th</sup> and 97.5<sup>th</sup> percentiles covers more than  $x\%$  of the measured data (i.e.,  $(100-x)\%$  of the data is treated as outliers), and 2) the average distance between the 2.5<sup>th</sup> and 97.5<sup>th</sup> prediction percentiles is less than the standard deviation of the measured data. If, the above two criteria are reached and a significant  $R^2$  (and or Nash-Sutcliffe coefficient) exists between the best simulation and the measured data for a calibration and a test data set, then the model can be considered calibrated. The parameter uncertainties are obtained by first assuming large uncertainty intervals and then iterating until the above conditions are reached. This requires a parameter updating procedure that is explained in this paper. If calibration cannot be attained with the above criteria, then the invoked conceptual model(s) should be re-examined.

## 1. INTRODUCTION

Inverse modeling has brought new opportunities, but also new challenges for model calibration. Some of the advantages are savings in time and cost of laboratory and/or field experiments generally needed to obtain the unknown parameters, and attainment of a better fit with available data. Another advantage is the usefulness of inverse modeling in the analysis of model structure (the invoked conceptual model), boundary conditions, and prevailing subsurface flow and contaminant transport processes. One of the limitations of inverse modeling is that the fitted parameters are conditioned on the experimental setup and the set of measured variables, which is usually limited in both time and space. Other conditioning factors include the choice of the parameter estimation routine, the form of objective function, and the weights associated with the different components of the objective function.

Another problem with inverse modeling is the non-uniqueness of the estimated parameters. While direct modeling with known parameters results in unique model variables, using the variables to obtain model parameters is by nature non-unique and results in uncertain parameters. Quantification of this uncertainty in model parameters has received special attention in recent years (Yapo et al., 1998; Beven, et al., 1992; Duan, et al., 2003; Abbaspour et al., 2004).

Our investigations of many time series hydrologic data showed that there may exist a very large number of parameter combinations that can produce acceptable model outputs. We also found that as the number of variables in a goal function increases, the number of acceptable simulations decreases. Furthermore, as a goal function is subjected to constraints, the number of acceptable simulations also decreases. Therefore, a very restrictive definition of the goal function can help to decrease the non-uniqueness problem. This, however, requires that a large number of variables are measured.

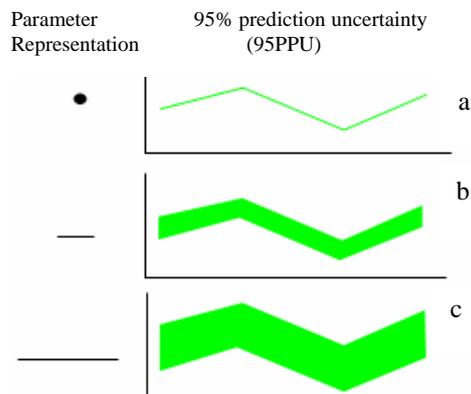
The objective of this paper is to describe a procedure for a combined parameter estimation and uncertainty analysis algorithm referred to as SUFI-2 (Sequential uncertainty fitting, ver. 2). SUFI-2 identifies a range for each parameter in such a way that upon propagation: 1) the 95% prediction uncertainty (95PPU) between the 2.5<sup>th</sup> and 97.5<sup>th</sup> percentiles contains (brackets) a predefined percentage of the measured data, and 2) the average distance between the 2.5<sup>th</sup> and 97.5<sup>th</sup> prediction percentiles is less than the standard deviation of the measured data. If, the above two criteria are reached and a significant  $R^2$  exists between the best simulation and the measured data

for a calibration and a test (validation) data set, then the model can be considered calibrated, and the parameter range is defined as the parameter uncertainty.

SUFI-2 was used for the calibration of several hydrologic problems including two bottom ash landfills using the program MACRO (Jarvis, 1994), transport of Cd from an agricultural field using HYDRUS-1D (Simunek, et al., 1998), and watershed modeling using the program SWAT (Arnold et al., 1998). SUFI-2 performs a combined optimization and uncertainty analysis using a global search procedure, and can deal with a large number of parameters through Latin Hypercube Sampling. This paper explains the above concepts using an example in which two municipal solid waste incinerator bottom ash monofills were successfully calibrated and tested for flow, and one monofill also for transport; and an example were a 1700 km<sup>2</sup> watershed in Switzerland, the Thur watershed, was calibrated and tested for discharge, sediment, phosphate, and nitrate loads at the outlet of the watershed.

## 2. THEORY

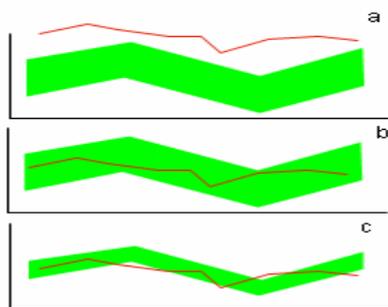
The concept behind the algorithm of SUFI-2 is presented graphically in Figure 3. If a model is provided with a single parameter value then a single simulation results (3a). If the parameter is uncertain, and this uncertainty is expressed as a distribution, then propagating this uncertainty results in range of possible problem solutions. One way of expressing the model result is through the 95% prediction uncertainty (95PPU) measured between the 2.5<sup>th</sup> and 97.5<sup>th</sup> percentiles. If the parameter uncertainty is small, then the 95PPU has a narrow band (3b), and if the parameter uncertainty is large then the 95PPU has a wide band as shown in Figure 3c.



**Figure 3.** The relationship between parameter uncertainty and prediction uncertainty.

Therefore, a procedure for parameter calibration (optimization) would be to start with a wide parameter uncertainty and then narrow this uncertainty in steps until a satisfactory 95PPU is reached. Note that if the initial parameter uncertainty distribution is set as wide as it is physically meaningful, then propagation of this distribution results in two possibilities. First, the measured data falls outside the 95PPU (Fig. 4a), and second, the measured data is inside the 95PPU (Fig. 4b). In the first case, we can conclude that the problem is not one of parameter calibration but rather one of conceptual model problem. In this case the model or the boundary conditions must be re-examined. In the second case, however, the initial parameter uncertainties can be calibrated to a narrower distribution, hence, a smaller 95PPU (Fig. 4c). During the calibration, however, some of the measurements will fall out of the 95PPU and hence, are not respected by the uncertainty in the parameters. Therefore, a balance must be reached by the size of the parameter uncertainty (and consequently the 95PPU) and the amount of data bracketed (respected) by the 95PPU. To obtain this balance we propose two conditions. 1) The 95PPU should bracket x% of the data, where x would depend on the nature of the project and the measured data. Normally, x should be around 80-90%. 2) To ensure that we have the narrowest 95PPU (and hence parameter uncertainty) we require that the ratio of the average distance between the upper and the lower 95PPU bonds and the standard deviation of the measured data should be less than 1.

Upon reaching the above criteria, if there exists a significant  $R^2$  and/or Nash-Sutcliffe coefficient between the best simulation and the measured data for a calibration and a test (validation) data set, then the model can be considered calibrated.



**Figure 4.** Relationship between measured data (red line) and the 95PPU. If parameter(s) distribution is set to the maximum physical limit, then a) is not a parameter calibration problem, b) calibration can obtain smaller uncertainty distribution, c) it can be expected that some measured data can fall outside the 95PPU.

## 2.1 Calibration Procedure

The calibration proceeds as follows:

1. Absolute, physically meaningful parameter distributions are assigned to each parameter.
2. Smaller initial uncertainty ranges are assigned to each parameter.
3. The number of simulations  $n$  is chosen and Latin Hypercube sampling is used to sample from the initial parameter distribution.
4. An objective function is formulated based on the types of measured data.

5. Model simulations are performed and the objective function is calculated for each of the  $n$  simulations. The following measures are then calculated:

- Jacobian matrix:  $J_{ij} = \frac{\Delta g_i}{\Delta b_j} \quad i=1, \dots, C_2 \quad j=1, \dots, m$

- Hessian  $H = J^T J$  and covariance  $C = s_g^2 (J^T J)^{-1}$

- Standard deviation of parameter  $b_j$ :  $s_j = \sqrt{C_{jj}}$

- 95% confidence interval of parameter  $b_j$ :

$$b_{j,lower} = b_j^* - t_{v,0.025} s_j \quad \text{and} \quad b_{j,upper} = b_j^* + t_{v,0.025} s_j$$

- Parameter sensitivity  $S_j = \bar{b}_j \frac{1}{C_2} \sum_{i=1}^{C_2} \left| \frac{\Delta g_i}{\Delta b_j} \right|$

- Parameter correlation  $A_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}} \sqrt{C_{jj}}}$

- Updated parameter uncertainties (assumed here to have uniform distributions) are calculated from:

$$b'_{j,max} = b_{j,upper} + \text{Max} \left( \frac{(b_{j,lower} - b_{j,min})}{2}, \frac{(b_{j,max} - b_{j,upper})}{2} \right)$$

$$b'_{j,min} = b_{j,lower} - \text{Max} \left( \frac{(b_{j,lower} - b_{j,min})}{2}, \frac{(b_{j,max} - b_{j,upper})}{2} \right)$$

In the above equations  $b$  is a parameter value,  $b^*$  is the best value (smallest goal function) in an iteration, and if the updated parameters go beyond the absolute values set in step 1, then they are adjusted back to the absolute values. This updating is centred on the best value in each iteration and the range is always smaller than the previous iteration. Also, if a region of the parameter space goes out of calculation in an iteration, it may come back again in a subsequent iteration.

- And finally, the 95PPU is calculated as the 2.5th and 97.5th percentiles of the cumulative distributions of every simulated point. If the condition of Figure 4b exists, then the parameters are updated and the procedure repeated from step 5 above.

When the two criteria mentioned above are reached, the  $R^2$  and or other required statistics are calculated. If these statistics are significant then it can be concluded that the model is calibrated. A similar procedure with a validation data test can validate the model.

The final parameter distribution calculated as such are considered to be the parameter uncertainties and the final 95PPU the model simulation uncertainty. This concept of uncertainty, in author's view, is more physically based and more meaningful than what is normally calculated through linear regression analysis, which is often too small.

### 3. APPLICATION OF SUFI-2 TO TWO BOTTOM ASH LANDFILLS

Bottom ash landfills are typically composed of equal amounts of fine ash material and melted components of which half have crystallised, and small quantities of metallic components, ceramics and stones (Kirby and Rimstidt, 1993). The Lostorf landfill, a municipal solid waste incinerator (MSWI) bottom ash monofill in Switzerland, was studied in detail by Johnson et al. (1999). Chemical analyses of leachate from this landfill at discrete intervals between 1994 and 1996 determined average total concentrations of Na, Cl, K, Mg, Ca, and  $SO_4$  to be 44.5, 47.1, 11.8, 0.63, 8.2, and 12.4 mM, respectively. A host of other metals such as Cu, Zn, Sb and Cr, Cd, Mo, V, Mn and Pb were also detected. While the leachate composition was found to be relatively constant during dry periods, considerable dilution occurred during rain events. The relatively good reproducibility of the experimental observations in response to rain events made us believe that transport modelling would be possible. Since the MSWI bottom ash contains high concentrations of heavy metal, monitoring and modelling of such landfills is important from an environmental point of view.

Different models have been used to simulate flow through MSWI bottom ash landfills (Guyonnet et al., 1998; Hartmann et al., 2001; Johnson et al., 2001). In the study of Johnson et al. (2001), flow through the Lostorf landfill was modelled using several approaches. They found that flow was dominated by preferential paths, reason why the variably-saturated dual-permeability model MACRO of Jarvis (1994) yielded the best simulation results.

In this study we extend the work of Johnson et al. (2001) and apply inverse modelling to study water flow in the Lostorf and Seckenberg landfills, both located in Switzerland. Solute transport modelling, through simulation of the electrical conductivity

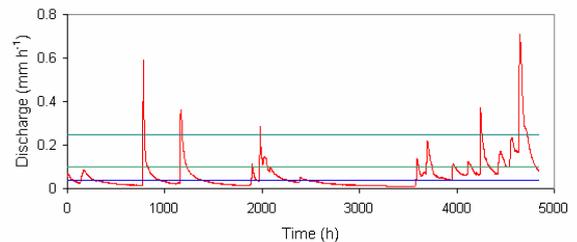
(EC), was performed only for the Lostorf landfill. To perform these analyses, we linked SUFI-2 with MACRO. Our objectives were to calibrate the model using hourly discharge, and its EC, from the landfills, and to test the calibrated models.

To formulate the objective function in this project, the time series of discharge was divided into four sections representing base-flow, recession, intermediate flows and peak flows. In this manner the calibration was forced to find equally good solutions to all sections of the flow as shown in Figure 5. The objective function was formulated as

$$g = \sum_{i=1}^I w_i \text{RMSE}_i, \text{ where } w_i = \frac{\text{avg}(\text{RMSE})_1}{\text{avg}(\text{RMSE})_i}$$

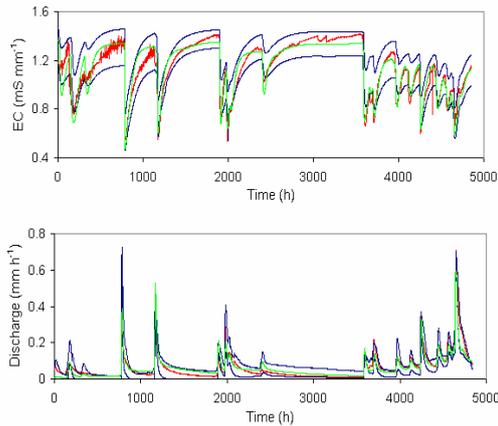
where  $I = 4$  if only discharge is considered and  $=8$  if both discharge and EC are considered. In the above equations

$\text{RMSE} = \sqrt{\frac{1}{k} \sum_{j=1}^k (q^o - q^s)_j^2}$ , where  $q$  is a measured variable,  $k$  is the number of observations in the  $i$ th section, and superscripts  $o$  and  $s$  refer to observed and simulated, respectively.



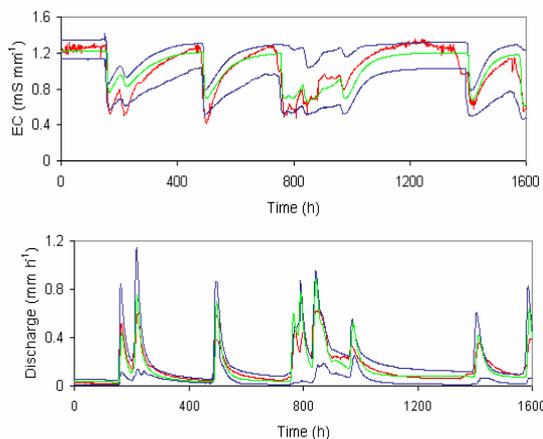
**Figure 5.** An example showing the division of a response signal into different regions. The weights are calculated such that each region contributes equally to the objective function.

In Figure 6 the calibration of EC and discharge for the Lostorf landfill is shown. The outer (blue) lines are the 95PPU, the red line is the measured discharge and the green line is the best simulation. The 95PPU contained 90% of the measured discharge and EC data, while the ratio of the average difference between the upper and the lower 95PPU over the standard deviation of the measured data was less than 1 for both discharge and EC. With  $R^2$  values between the best simulation and the observed data of 0.86 and 0.88, respectively, for discharge and EC all calibration requirements were met for Lostorf.



**Figure 6.** Calibration results for Landfill Lostorf.

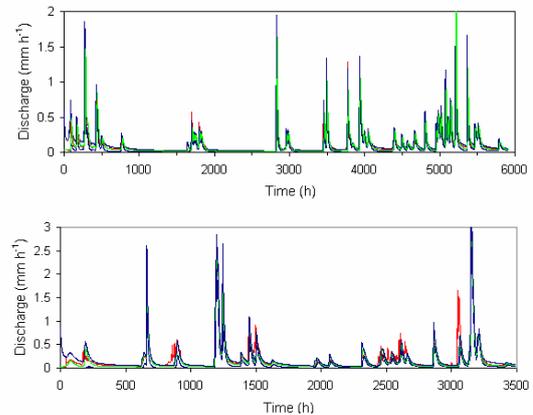
The validation results for Lostorf landfill are shown in Figure 7. The 95PPU contained more than 90% of the measured discharge and EC data, while the ratio of the average difference between the upper and the lower 95PPU over the standard deviation of the measured data was about 1 for discharge and 1.7 for EC. With  $R^2$  values of 0.85 and 0.82, respectively, for discharge and EC all except the ratio requirement for EC were again met. As it is discussed in more detail in Abbaspour et al. (2004), this indicates that the calibrated parameter ranges had produced a large number of relatively poor simulations for EC.



**Figure 7.** Validation results for landfill Lostorf.

Figure 8 presents the calibration and test results for flow through the Seckenberg landfill. The plots show both the best simulation results and the 95PPU. The 95PPU contained 93% and 90% of the measured discharge data for the calibration and validation data sets, respectively. The ratios of the average difference between the upper and the lower 95PPU over the standard deviation of the measured data were much less than one for both

calibration and validation data sets. The  $R^2$  requirement was also met with highly significant values of 0.94 and 0.88 for calibration and validation data, respectively. One reason for the better calibration results for the Seckenberg case is the fact that the objective function contained only the discharge data. For more detail see Abbaspour et al. (2004).



**Figure 8.** Calibration (upper) and validation (lower) results for landfill Seckenberg.

#### 4. APPLICATION OF SUFI-2 TO THUR WATERSHED

In a national effort, since 1972, the Swiss Government started the “National Long-term Monitoring of Swiss Rivers” (NADUF) program aimed at evaluating the chemical and physical states of major rivers leaving Swiss political boundaries. The established monitoring network of 19 sampling stations included locations on all major rivers of Switzerland. This study complements the monitoring program and aims to model one of the program’s catchments – Thur River basin (area 1,700 km<sup>2</sup>), which is located in the north-east of Switzerland and is a direct tributary to the Rhine.

We used the program SWAT (Soil and Water Assessment Tool) (Arnold et al., 1998) to simulate all related processes affecting water quantity, sediment, and nutrient loads in the catchment. Manual calibration was performed, in the first step, based on observations at the catchment outlet. As this calibration produced unacceptable loads from various landuses, a second calibration was performed with SUFI-2 based on the knowledge of loads from landuses as well as the observations at the catchment outlet. The estimates of loads from different landuses were available from literature and expert judgement. The second calibration produced much more reasonable calibration and validation results as there were good agreements between simulated and observed bi-weekly

discharge, total suspended sediment, and nutrient loads. Agricultural areas delivered the largest sediment and nutrient loads to the streams. This was the case especially after high rainfall events. The present study demonstrates the overall effectiveness of the adopted integrated spatially-distributed modelling approach in investigating the holistic relationships between natural system and landuse.

The objective function formulated for the optimization is referred to as a “constraint objective function”. Its formulation is as follows:

$$\min: g = w_1 \sum_{n_1} (q_m - q_s)^2 + w_2 \sum_{n_2} (S_m - S_s)^2 + w_3 \sum_{n_3} (NO_{3,m} - NO_{3,s})^2 + w_4 \sum_{n_4} (P_m - P_s)^2$$

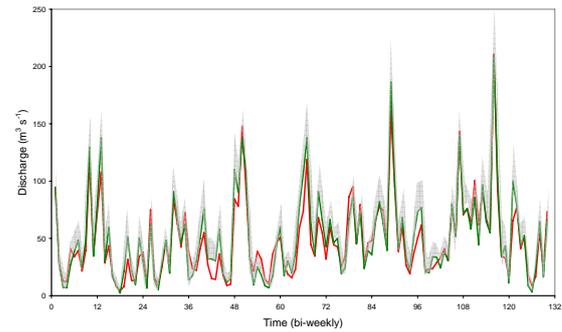
subject to:

$$\begin{aligned} S_{\min,j} &\leq S_{load}(j) \leq S_{\max,j} \\ NO_{3,\min,j} &\leq NO_{3,load}(j) \leq NO_{3,\max,j} \\ P_{\min,j} &\leq P_{load}(j) \leq P_{\max,j} \end{aligned} \quad j=1,\dots,L$$

where  $q$  is discharge,  $S$  is sediment load,  $NO_3$  is nitrate load, and  $P$  is the phosphate load in the river at the watershed outlet.  $w$ 's are the weights chosen such that each component has equal contribution to the objective function, and  $n$ 's are the number of measurement.  $S_{load}(j)$  is the sediment load from landuse  $j$ ,  $NO_{3,load}(j)$  is the nitrate load from the landuse  $j$ , and  $P_{load}(j)$  is the phosphate load from landuse  $j$ . subscripts min and max are the lower and upper range of loads from the landuses determined from previous studies and expert judgement, and  $L$  is the number of landuses. In this study there were six landuses: forest, summer pasture, wheat, alpine pasture, barren land, urban, and orchards. Optimization of the above objective function resulted in parameters that produced the best simulations and at the same time respecting proper loads from various landuses within the watershed. This type of constraint multi-component objective functions helps to decrease the range of solutions dramatically and is a practical solution to the non-uniqueness problem. A downside, however, is that various measurements and constraining conditions must be available. This is usually not the case for most watersheds.

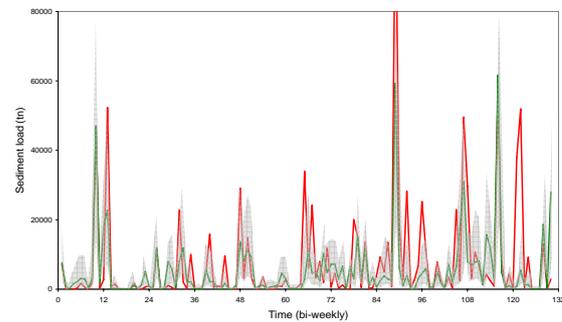
The calibration result for discharge is shown in Figure 9. Shaded region shows the 95PPU while observed discharge is shown in red and best simulation result is shown in green. The  $R^2$  value is

0.91 and about 80% of the measured data is bracketed by the 95PPU.



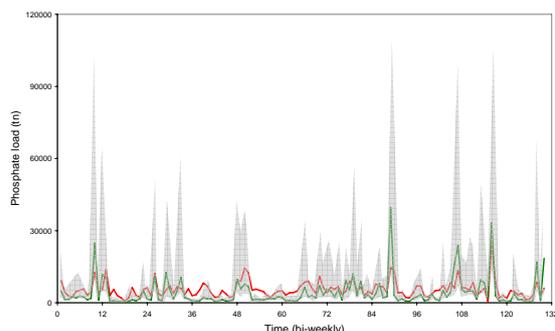
**Figure 9.** Calibration result for discharge at the outlet of Thur watershed.

Figure 10 shows the calibration result for sediment load. Figure explanation is similar to discharge. The  $R^2$  value is 0.55 and around 75% of the measured data is bracketed by the 95PPU. For most times, the simulated sediment underestimates the measured data. In view of a local expert who was involved in sediment measurement, the measuring devices were erroneously located in areas with high local turbulence, and hence, the measurements are overestimated.



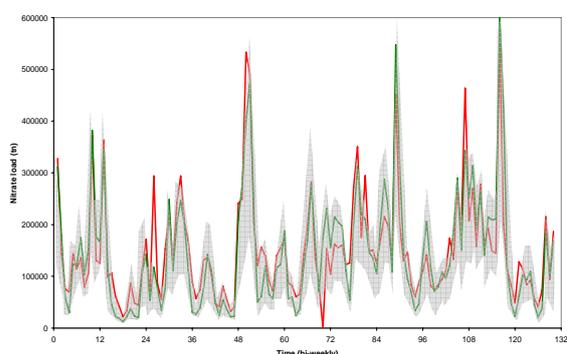
**Figure 10.** Calibration result for sediment at the outlet of Thur watershed.

The calibration result for phosphate is shown in Figure 11. Shaded region shows the 95PPU while observed discharge is shown in red and best simulation result is shown in green. The  $R^2$  value is 0.53 and about 70% of the measured data is bracketed by the 95PPU.



**Figure 11.** Calibration result for phosphate at the outlet of Thur watershed.

value is 0.80 and about 90% of the measured data is bracketed by the 95PPU.



**Figure 12.** Calibration result for nitrate at the outlet of Thur watershed.

The above calibration results for a watershed of 1700-km<sup>2</sup> area is quite acceptable, especially, when nutrient and sediments loads from each landuse is within acceptable limits.

## 5. CONCLUSIONS

The SUFI-2 algorithm for a combined parameter optimization-uncertainty analysis was developed to handle a large number of parameters and was applied to two different situations. In both cases quite good calibration and validation results were obtained. Parameter uncertainties and prediction uncertainties were calculated and statistics such as % measured data in the 95PPU and  $R^2$  gave a good measure of the strength of calibration results.

## 6. REFERENCES

Abbaspour, K.C., A. Johnson, M. Th. van Genuchten (2004). Estimating uncertain flow and transport parameters using a sequential

uncertainty fitting procedure. *Vadose Zone Journal*, 3:1340-1352.

Arnold, J.G., Srinivasan, R., Mutiah, R.S., Williams, J.R. (1998). Large area hydrologic modeling and assessment part I: model development. *Journal of American Water Resources Association*, 34 (1), 73-89.

Beven, K., and A. Binley (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*. 6:279-298.

Duan, Q., S. Sorooshian, V. Gupta, A. N. Rousseau, and R. Turcotte (2003). *Advances in Calibration of Watershed Models*. AGU, Washington, DC.

Guyonnet, D., B. Didier-Guelorget, G. Provost, and C. Feuillet. 1998. Accounting for water storage effects in landfill leachate modeling. *Manag. Res.* 16:285-295.

Hartmann, F., H. P. Bader, R. Scheidegger, and P. Baccini (2001). Water transport in a bottom ash landfill from a municipal solid waste ("MSW") incinerator. *J. Solid Waste Technol. Manag.* 27:76-81.

Jarvis, N. J. (1994). The MACRO model (version 3.1). Technical description and sample simulations. Reports and Dissertations 19, Dept. Soil Sci., Univ. Agric. Sci., Uppsala.

Johnson, C. A., M. G. Schaap, and K. C. Abbaspour (2001). Model comparison of flow through a municipal solid waste incinerator ash landfill. *J. of Hydrol.* 243:55-72.

Johnson, C. A., M. Kaeppli, S. Brandenberger, A. Ulrich, and W. Baumann (1999). Hydrological and geochemical factors affecting leachate composition in municipal solid waste incinerator bottom ash. Part II: The geochemistry of leachate from Landfill Lostorf, Switzerland. *J. of Contam. Hydrol.* 40:239-259.

Kirby, C. S., J. D. Rimstidt (1993). Mineralogy and surface properties of municipal solid waste ash. *Environ Sci. Technol.* 27:652-660.

Simunek, J., Sejna, M., and van Genuchten, M. Th., (1998). The JYDRUS-1D software package for simulating the one-dimensional movement of water, heat, and multiple solutes in variably-saturated media. U. S. Salinity Lab., Riverside, California.

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