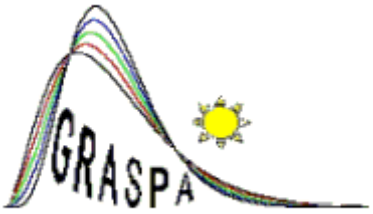
Web Working Papers
by

The Italian Group of Environmental Statistics




Gruppo di Ricerca per le Applicazione della Statistica
ai Problemi Ambientali

**www.graspa.org**

# Statistical Sensitivity Analysis and Water Quality

Alessandro Fassò

GRASPA Working paper n.24, March 2006

# Statistical sensitivity analysis and water quality

Alessandro Fassò*

June 30, 2005

**Abstract**

In this paper, concepts and methods of statistical sensitivity analysis ($SA$) of computer models are reviewed and discussed in relation to water quality analysis and modelling.

The starting point of this approach is based on modelling the uncertainty of the computer code by probability distributions. Despite the fact that computer models are generally speaking non-stochastic, in the sense that if we rerun the code we get the same result, the stochastic approach turns out to be useful to understand how the input uncertainty is propagated through the computer code into the output uncertainty.

We follow the standard approach to $SA$ which is based on variance decomposition and consider three levels of $SA$. At the first or preliminary level, we discuss $DOE$ and response surface methodologies in order to get a first estimate of the input influences on the model output.

At the second level, going further into modelling the relationship between computer model inputs and outputs, we assume that different computer runs are independent. We then discuss techniques derived from Monte Carlo input simulations and regression analysis.

At the third level, recognizing that, since the computer model is actually non-stochastic, the errors are often smoother than independent errors, we consider the geostatistical $SA$ which is based on assuming that the error of the computer code emulator is a stochastic process with positive correlation which gets higher as two inputs get closer.

**Keywords:** Computer models, uncertainty analysis, water quality, Monte Carlo, Geostatistics

*University of Bergamo, Dept. IGI, Via Marconi 5, 24044 Dalmine BG, Italy. Email: alessandro.fasso@unibg.it. Work partially supported by Italian MIUR-Cofin 2004 grants.

# 1 Introduction

## 1.1 Computer models and recreational water quality

Computer models are widely used in hydrology and water quality studies in general. In recreational water quality modelling and assessment, the use of both conceptual and management models are increasingly important.

As a first example, consider real time forecasting E. coli concentrations, which is useful for management beach closure strategies and may be approached by both mechanistic and statistical models. In this frame, Olyphant and Whitman (2004) applied dynamical regression models including hydrological, meteorological and water quality predictors to swimming beaches of Lake Michigan.

Moreover, Vinten et al. (2004) compared soil transport models, multiple regression models and distributed catchment models in the catchment of the River Irvine, Scotland. In deep ocean outfall plumes off Sydney, Miller et al. (1996) used finite element modelling, to assess both long and short term effects.

Reynolds (1999) reviews various modelling strategies for understanding phytoplankton dynamics in water quality and lake management. For river and lake water quality, computer models ($CM$) may be used in integrated analyses at the catchment scale where various dimensions are usually taken into account.

Jamieson et al. (2004), in order to assess microbial pollution of rural surface water considered liquid and solid waste generated from industry, zootechny and domestic sources. They review some approaches to modelling both surface and subsurface transport of the associated microorganisms and their flow through stream networks.

Norton et al. (2004) considered the hydrologic, economic and stream sediment sources of uncertainty in a calibrated computer model applied to the Ben Chefley Dam, Australia.

Hydrological models are important here because they are often used as submodels of water quality models. For example, Whitehead et al. (1997) considered a combined flow and process based river quality model including nitrate, dissolved oxygen, biochemical oxygen demand, ammonium ion, temperature, pH and a *conservative* water quality determinand. In general, mechanistic models have been extensively studied in hydrology, in particular flow models and rainfall-runoff models, see e.g. Beven (2001) . In dry areas, e.g. in the Mediterranean area, water quality may be severely influenced by reduction in flow. Becciu et al. (2002) studied a calibrated conceptual model for minimum instream flow in Central Alps catchments by means of regression modelling and outlier analysis.

An other issue relevant for recreational water is waste water management. For example in heavy metal biofilter modelling, Fassò et al. (2003) used a conceptual model based on the advection dispersion reaction equation and modelled the multivariate response using a multivariate heteroskedastic statistical approximation.

From the above examples, the CM outputs may be the stream discharge or the concentration of chemicals and/or pollutants or time to next health hazard event; and the CM relates these to anthropic and environmental parameters, initial and boundary conditions, global climate and dynamics of meteorology.

## 1.2 Sensitivity analysis and paper structure

Uncertainty may be related to measurement errors, both at model output ($MO$) and parameter level. Moreover it may be due to the fact that the $CM$ is only an approximation of the *real system*. Such sources of uncertainty will be discussed in some detail in section 2 where, we extend the taxonomy of Kennedy and O'Hagan (2001) .

In some cases, the $CM$ needs to be calibrated on some observational data. It is then interesting to assess the estimation or calibration uncertainty and the sensitivity of the $MO$ to the calibration parameters. In other cases, calibration is not explicitly considered, but once again $SA$ is aimed at understanding to what extent the various parameters affect the $MO$.

Sensitivity analysis ($SA$) is then intended to assess these individual sensitivities and to rank various inputs with respect to certain *sensitivity indexes*. If we avoid uncertainty concepts, the simplest idea for doing $SA$ is to consider first order local expansion at some *internal point* and use the analytical or numerical partial derivatives to carry out this *local* $SA$.

In section 3, we discuss the approach known as *global* $SA$. The aim is to define the global influence of each input to the uncertainty of the $MO$. Then, using an appropriate global performance measure, e.g. variance, squared or absolute fitting error or likelihood, we show how to assess and rank the sensitivity to each parameter. We first review and comment the case considered extensively in Saltelli et al. (2000) and in Fassò and Perri (2002) where the $CM$ is taken for granted or, equivalently, no calibration data are available so we assess the sensitivity of the $MO$ without reference to observed data.

In section 4, we discuss the preliminary $SA$, generally based on a reduced number of computer runs and little statistical modelling. In such a case, design of experiments ($DOE$) and response surface methodology techniques are of interest. At a subsequent step, when computer runs are cheap, Monte Carlo $SA$ is useful. This technique and modified sampling strategies (Latin Hypercube and importance sampling) are discussed in section 5. In section 6, model based $SA$ is discussed and the variance based $SA$ is extended to multivariate and heteroskedastic $CM$'s; in the latter case, the residual model uncertainty is not constant over the input domain.

In section 7, we discuss some $SA$ techniques related to the case where calibration data are available and $CM$ validation may be performed also with $SA$. In section 8, we discuss the case where the uncertainty of the $MO$, prior to running the $CM$, is assumed to be a stochastic process indexed by the computer model input $x$. In the previous sections, the Monte Carlo approach was based on independent computer runs. Here, recognizing that the

original computer code is non-stochastic, the error smoothness is described by a geostatistical approach.

# 2 Model uncertainty setup

In order to introduce uncertainty concepts, we first suppose that the *true* environmental phenomenon of interest, say $\zeta$, is related to some observable multidimensional inputs $x = (x_1, ..., x_k)$ in some input domain, say $D$, and some other non-observable or unknown inputs $x^*$, that is

$$\zeta = \zeta(x, x^*).$$

The computer model ($CM$) or code is a computable function, say $f(x)$, which for given inputs $x$ gives an output

$$z = f(x).$$

Usually it is a complex function and its analytical properties are difficult to derive. In some cases, it may be a stochastic function including for example some Monte Carlo or other simulation based components. In this paper, we consider mainly deterministic $CM$'s, in the sense that, if we re-run the code, we get the same result. In the simple ideal case the $CM$ is a perfect model so that

$$f(x) = \zeta(x, x^*) \tag{1}$$

for every $x^*$.

## 2.1 Input uncertainty

In environmental $CM$'s it is common to have two kinds of input parameters, that is fixed and variable parameter vectors denoted by $\theta = (\theta_1, ..., \theta_h)$ and $x = (x_1, ..., x_k)$ respectively, giving the $CM$ equation

$$z = f(x, \theta). \tag{2}$$

The vector $\theta$ is often referred to as the "calibration parameter" to be *estimated* on observational data. For example, in a hydrological model applied to a certain watershed, the parameter set $\theta$ may be related to geomorphological and/or evapotranspiration parameters of that watershed, while $x = (t, y_1, y_2)$ may be the time index $t = 1, 2, ...$ and meteorological conditions $y_1$ and discharges $y_2$ at time $t$.

We are often interested in the global behaviour of the true system $\zeta$ without fixing the input $x$. Or in a risk analysis, we are interested in right tail behaviour of risk-related $MO's$. So, in practice, the $k - dimensional$ input $x = (x_1, ..., x_k)$ is uncertain and it may be useful to describe such uncertainty by an appropriate $k - variate$ probability distribution with joint probability density function given by $p(x)$ and cumulative distribution $P(x)$.

The simplest example of input distribution is given by independent rectangular marginals. We will see in the following sections that when inputs are independent the sensitivity indexes satisfy certain additivity properties.

In some cases this simple setup has to replaced by other multivariate distribution. For example in Fassò et al. (2002*b*), section 4.2, considering the $SA$ of a heavy metal biofilter $CM$, the maximum uptake constant ($q_{max}$) and the Langmuir constant ($b$) are supposed bivariate Normal with moderate positive correlation, $\rho = 0.30$, to reflect the calibration uncertainty source of these parameters.

## 2.2 Simulation and residual uncertainty

Except the simplistic case of equation (1), since $x^*$ is unknown, the $CM$ or simulator $z = f()$ is, at best, an approximation of the averaged values of $\zeta$, say $\mu$. This is given by

$$\mu(x) = E_{x^*}(\zeta(x, x^*)|x)$$

where $E_{x^*}(.|x)$ is the conditional expectation operator with respect to some conditional distribution $p(x^*|x)$.

Hence, if $\zeta$ is observed without error the *residual uncertainty* is given by the probability distribution of

$$e_0 = \zeta - \mu \tag{3}$$

and the $CM$ inadequacy or simulator uncertainty is given by

$$e_1 = f - \zeta = \bar{e}_1 + e_0 \tag{4}$$

where $\bar{e}_1 = (f - \mu)$ is the *partial simulation uncertainty* while $e_1$ is the *total simulation uncertainty*. If observational data, say $Z$, are available about $\zeta$ then measurement errors are possible and

$$Z = \zeta + \varepsilon_\zeta.$$

This case may be handled in the Bayesian framework of section 8 or, under Markovian assumption on the unobserved $\zeta$, by the dynamical system setup and the Kalman filter, see e.g. Fassò and Nicolis (2005) for an application of this approach to air quality.

## 2.3 Emulation

The next step is to suppose that we have a simplified model, say $g(x, \beta)$, where $\beta$ is a "*regression type*" parameter to be estimated in order to give a *good* approximation of the CM $f(x)$.

Of course we have partial and total emulation uncertainty given, respectively, by

$$\bar{e}_2 = g - f \tag{5}$$

and

$$e_2 = g - \zeta = \bar{e}_2 + e_1.$$

The fixed but unknown parameter $\beta$ may be interpreted, for example, as the minimum mean square error parameter which minimizes

$$E_x \left( g\left(x, \beta\right) - f\left(x\right) \right)^2.$$

## 2.4 Estimated emulator

In practice we may get an estimate $\hat{\beta}$ using simulated data from the CM

$$\left(x_i, f\left(x_i\right)\right), i = 1, ..., n.$$

This gives the estimated emulator

$$\hat{g}\left(x\right) = g\left(x, \hat{\beta}\right)$$

and we have another two sources of uncertainty, say partial and total estimation uncertainty, given respectively by

$$\bar{e}_3 = \hat{g} - g$$

and

$$e_3 = \hat{g} - f. \tag{6}$$

In some cases $\hat{\beta}$ is a statistical estimate, e.g. maximum likelihood estimates, and the uncertainty on $\beta$ and the errors $e_3$ and $\bar{e}_3$ may be assessed using some standard approximate normality and confidence intervals. In other cases $\hat{\beta}$ is calibrated using e.g. hydrological techniques giving $GLUE$ methodology which is discussed in section 7.1.

In the following sections 4 and 5, the quantities of main interest are the emulated values $\hat{g}$ and the corresponding errors given by equation (6) .

## 2.5 Output Uncertainty

The uncertainty on the input $x$ propagates to the output $z$ via the $CM$, $f\left(x\right)$, so that, as long as $x$ is a random vector with distribution $p\left(x\right)$, we are interested in the output uncertainty distribution, $p\left(z\right)$ say, which is related to $p\left(x\right)$ via the code $f\left(x\right)$. For example in risk analysis we are interested in the cumulative output distribution $P\left(z\right)$ and its right tail quantiles.

A typical quantity for assessing the squared uncertainty is the output variance, which may be computed using the input uncertainty distribution $p\left(x\right)$:

$$Var\left(z\right) = \sigma_z^2 = \int_D \left(f\left(x\right) - f_0\right)^2 p\left(x\right) dx$$

where $f_0 = E\left(z\right)$.

Moreover, the $MO'$s may be compared with observational data of the *true system.* Let $e$ be the *forecasting error* according to one of the setups

from sub-sections 2.1-2.4. For example, for an exactly observed system with known $CM$, we have $e = f - \mu$ and, for an emulated model, the forecasting error is $e = \hat{g} - \mu$.

As can be seen from these last two quantities, such error accounts also for model inadequacy. Therefore, the output uncertainty is generally given by the error cumulative distribution, $P(e)$ say. If we have replicated input values, for example a random sample $x_1, ..., x_n$ from $p(x)$ as discussed in section 5, we can use standard statistical inference to estimate $P(e)$ its mean, variance, confidence intervals etc.

# 3 Variance based $SA$

Most of the remaining part of this paper is based on data coming exclusively from the $CM$. Hence, except in section 7, we will not consider in detail either the residual uncertainty (3) or the simulator uncertainty (4).

In principle the sensitivity of the $MO$'s, $z$, to each component of $x = (x_1, ..., x_k)$ may be based on the local approach by the partial derivatives

$$\frac{\partial f}{\partial x_j}$$

which can be computed either analytically or numerically around a "central point" $x^0 = (x_1^0, ..., x_k^0)$. Whenever this approach has been used for a long time and is still being used, it is rather simplistic for complex nonlinear $CM$.

Extending the local $SA$ to "many" $x^0 \in D$ would give more information but, of course, would rebuild the complexity and the multidimensionality of $f()$ itself. So we need a "global" approach that is able to give information for every $x$ but is also a synthesis which reduces the original complexity. Moreover, we search for quantities that can be "estimated" on a reduced set of $CM$ runs.

The basic idea of global $SA$ is to study the overall influence of each input component $x_j$ to the uncertainty of the $MO$. In variance based $SA$, we assess the uncertainty by the variance and we are naturally lead to $SA$ measures based on variance decomposition, for example using a main-effect model

$$z = f_0 + \sum_{j=1}^{k} f_j + \varepsilon \tag{7}$$

with $f_0 = E(z)$ as above and

$$f_j = E(z|x_j) - f_0.$$

Note that the error $\varepsilon$ here is non-stochastic as it is a pure model-inadequacy quantity. Whenever the standard statistical interpretation does not hold, in many situations such an error, being *a complicated* function of many independent inputs $x$ behaves close to a stochastic error.

If the inputs are independent we can decompose the total uncertainty as

$$Var\left(z\right)=\sum_{j=1}^{k}Var\left(f_{j}\right)+Var\left(\varepsilon\right) \tag{8}$$

and the Pearson's correlation ratio

$$S_{j}=\frac{Var\left(f_{j}\right)}{\sigma_{z}^{2}} \tag{9}$$

is the *natural* first order sensitivity index for $x_j$. As a matter of fact, $Var\left(f_{j}\right)$ may be interpreted as that part of the uncertainty of the output which can be reduced by fixing the $j^{th}$ input parameter and, correspondingly, the sensitivity $S_j$ may be interpreted as the fraction of (squared) uncertainty of $z$ due to the uncertainty on $x_j$

In principle, we can assess interactions of any order starting from the full interaction model

$$z=f_{0}+\sum_{j}f_{j}+\sum_{i<j}f_{i,j}+...+f_{j_{1},...,j_{k}} \tag{10}$$

where $f_{i,j}=E\left(z|x_{i},x_{j}\right)-f_{i}-f_{j}-f_{0}$ and so on.

In this case, in order to cover the effect of the interactions between $x_j$ and the other inputs, the sensitivity index $S_j$ may be increased to get the total effect. To see this, let $x_{(j)}$ be the $(k-1)-dimensional$ vector corresponding to $x$ without the $j^{th}$ component, and consider the following decomposition

$$z=f_{0}+f_{j}+f_{(j)}+f_{j,(j)}.$$

Now, using the input independence, we have

$$Var\left(z\right)=Var\left(f_{j}\right)+Var\left(f_{(j)}\right)+Var\left(f_{j,(j)}\right)$$

and, following Homma and Saltelli (1996), the total sensitivity index for $x_j$ is given by

$$S_{T_{j}}=\frac{Var\left(f_{j}\right)+Var\left(f_{j,(j)}\right)}{\sigma_{z}^{2}}=1-S_{(j)}.$$

## 3.1 Further details

Let $D_j$ and $D_{(j)}$ be the input domains of $x_j$ and $x_{(j)}$ respectively. Then the output response to $x_j$ is given by the $(k-1)-$dimensional integral

$$E\left(z|x_{j}\right)=\int_{D_{(j)}}f\left(x\right)p\left(x_{(j)}\right)dx_{(j)} \tag{11}$$

and its variance is given by the one-dimensional integral

$$Var\left(E\left(z|x_{j}\right)\right)=\int_{D_{j}}E\left(z|x_{j}\right)^{2}p\left(x_{j}\right)dx_{j}-f_{0}^{2}. \tag{12}$$

Moreover, $Var\left(f_{(j)}\right)$, which enters the total sensitivity $S_{T_j}$, is given by

$$Var\left(E\left(z|x_{(j)}\right)\right) = \int_{D_{(j)}} E\left(z|x_{(j)}\right)^2 p\left(x_{(j)}\right) dx_{(j)} - f_0^2.$$

Shortcuts for estimating $S_{(j)}$ are discussed e.g. in Chan et al. (2000) and their efficiency may be assessed in practice using equation (27) of Fassò and Perri (2002).

# 4 Exploratory analysis

At the early stages of the $CM$ analysis, especially if the computer runs are *expensive*, it may be worth considering a simplified emulator, based on a reduced set of values for each input $x_j$. For example, consider just binary inputs which assume the values High/Low giving the new input domain, say $D^*$, with $2^k$ different values. This is known as a $2^k$ factorial design which requires running the code $2^k$ times and allows to identify the zero error full interaction model (10).

When the figure $2^k$ is too large and/or "all the interactions" are too many, we need *smaller* designs such as the fractional designs $2^{k-h}$ which allows the estimation of a reduced version of model (10) with high order interactions being encompassed in the error component as in model (7).

The related techniques known as design of experiment ($DOE$) and response surface methodology ($RSM$) are well established for stochastic experiments, see e.g. the classic Box et al. (1978) and the more recent Wu and Hamada (2000). Sacks et al. (1978) considered the so-called *design of computer experiments* and its optimization is also discussed in section 8.1 below. Whenever at first sight, standard $DOE$ seems to work also in this case, it has to be recognized that, due to the non-stochastic nature of computer experiments, now replications, blocking and randomization loose their usual meaning.

Moreover, for output uncertainty estimation, considering only binary inputs may be of limited value. One can extend to the $n$ level factorial design with $n^k$ components or fractionally reduced, but as $n$ and $k$ are not very small it does not work in practice.

# 5 Monte Carlo and other sampling techniques

In the previous section "optimal" systematic sampling has been considered for the case where the response surface is fixed in advance and certain cardinality reduction assumptions are in order. Using this approach for complex emulators $g(x)$ and/or high dimensional and high cardinality inputs is not feasible in practice because of computational complexity. Then in this section, we get some techniques which are not optimal but are informative for any particular emulator.

To do this, the idea of section 2.1 which describes the input uncertainty by a certain probability distribution, say $P(x)$, is accomplished with the assumption that different runs are independent. This gives a natural way to get information about the $CM$, that is simple random sampling from $P(x)$. This means that we need (pseudo) random numbers from $P(x)$ and this is easily done with standard software. Using this approach we get a (possibly large) sample from the $CM$, namely $(x_1, z_1), ..., (x_n, z_n)$,which is informative about the code $f()$ and may be used for empirical modelling, estimating and validating the emulator $g()$. Moreover, it is useful for estimating the indexes of section 3 and, as $z_1, ..., z_n$ is a random sample from the unknown distribution $P(z)$, it may be used to get the estimated output uncertainty distribution, say $\hat{P}(z)$.

Of course this approach is especially appropriate when computer runs are cheap and getting "a large Monte Carlo sample" is a feasible task in terms of computing resources.

## 5.1 Importance sampling

Suppose we are interested in estimating the average of the positive output function $h(x) > 0$

$$\mu = E(h(x)) = \int_D h(x) p(x) dx.$$

For example, we may be interested in computing the output mean, with $h(x) = |f(x)|$ or the variance with $h(x) = (f(x) - f_0)^2$.

Using the standard Monte Carlo approach, we would estimate $\mu$ by means of a random sample $x_1, .., x_n$ from $p(x)$ and its sample average

$$m = \frac{1}{n} \sum h(x_i).$$

The idea of importance sampling is to use a stratified sample from a cumulative distribution $Q(x) \neq P(x)$ which gives higher probability to those inputs $x$, where $h(x)$ is large. In practice the $i^{th}$ stratified importance sample, $x_i'$ say, is given by

$$x_i' = Q^{-1}\left(\frac{i - 1 + R_i}{n}\right) \tag{13}$$

where $R_i$ is a uniform random number and the unknown $\mu$ is now unbiasedly estimated by the weighted estimator

$$m' = \sum \frac{h(x_i')}{q(x_i')/p(x_i')}. \tag{14}$$

It is easily seen that if $q(x) = \frac{h(x)p(x)}{\mu}$ then $m'$ is zero variance and, hence, optimal. On the one hand, the sampling strategy, which increases the sampling size where the $CM$ uncertainty is large, is more efficient than standard

Monte Carlo sampling. On the other hand, application of this method requires approximate knowledge of the $CM$ itself. Moreover, in equation (14), weighting is essential to avoid bias. Finally, if $x$ is multivariate then stratified sampling gives the course of dimensionality of the previous section and $LHS$ of the next section should be taken into consideration.

## 5.2 Latin hypercube sampling

This sampling method, acronimized by $LHS$, is a multidimensional generalization of the stratified sampling which assigns each scalar sample $x_i$, $i = 1, ..., n$ to a different equi-probability interval or cell, $c_i$ say, using equation (13) with $P_i$ instead of $Q$. In the $k-$dimensional case, we have a $k-$dimensional grid of $n^k$ cells $c_i$ given by the Cartesian product of the marginal intervals $c_{i,j}$, that is $c_i = c_{i,1} \times ... \times c_{i,k}$.

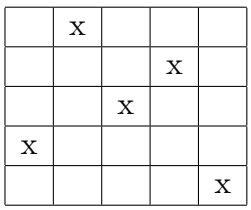



Figure 1: *Example of a two-dimensional Latin hypercube assignment with* $n = 5$ *and rectangular marginals*

The $n^k$ factorial design of previous sections would simply give one element $x_i$ for each cell $c_i$. Now in $LHS$, as shown by Figure 1, the cells are chosen so that each marginal has just one observation in each of the $n$ equi-probability intervals and it may be seen as a highly fractionalized factorial design. As a matter of fact, the term comes from Latin Squares where there is an array of symbols and each occurs just once.

### 5.2.1 Algorithm

To do this, note that the cells $c_i$ are identified by $k$ integers ranging in $1, ..., n$ hence the $n \times k$ matrix $C$ of such integers has columns which are given by random permutations of the integers $1, ..., n$.

After choosing the cell $c_i$ the value $x_i$ is chosen from $P(x|c_i)$ using equation (13) thanks to independence. The extension to certain correlation structures is considered by Stein (1987).

### 5.2.2 Optimality

It is known that if code $f(x)$ is monotonic in each component $x_j$, then $LHS$ improves on random sampling for estimating the output mean, variance and cumulative distribution function (McKey et al., 1979). Nevertheless, due

to the high degree of fractionalization, this technique requires some caution when used for high order interaction models (e.g. Hungtington et al., 1998).

# 6 Model based $SA$

In section 4, we considered response surface methodology as a way to understand how the inputs affect the computer code. In this section, we are more deeply concerned with the model emulator and its capability to give further insight into the $CM$ in general and in its sensitivity indexes in particular.

Let us start by considering a linear regression emulator

$$g(x, \beta) = \beta_0 + \sum_{j=1}^{k} x_j \beta_j \tag{15}$$

with errors (5) close to independent, homoskedastic Gaussian errors. If the input components are uncorrelated as in section 3, we get the sensitivity indexes $S_j$ from the variance decomposition.

$$\sigma_z^2 = \sum_{j=1}^{k} \sigma_{x_j}^2 \beta_j^2 + \sigma_\varepsilon^2. \tag{16}$$

To do this, using a large enough Monte Carlo sample, we can use the least square estimates of $\beta$ to get the estimated sensitivity indexes

$$\hat{S}_j = \frac{\hat{\sigma}_{x_j}^2 \hat{\beta}_j^2}{\hat{\sigma}_y^2}$$

and from (16) we have

$$\sum \hat{S}_j = R^2 = 1 - \frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_y^2}. \tag{17}$$

This approach easily extends to interactions, polynomial components and transformed inputs, using e.g. the following generalized linear model

$$g(x, \beta) = h(x)' \beta. \tag{18}$$

Some caution is required for high dimensional input sets and high order interactions. For example, Helton et al. (2005) , doing $SA$ of a waste isolation plant with more then thirty inputs, found that step-wise regression was unstable and they preferred separated analyses.

## 6.1 Nonlinear and multivariate $SA$

Often the code output is a vector and we are interested in assessing the sensitivity of the $CM$ as a whole. For example, considering a waste water biofilter model, Fassò et al. (2003) were interested in performance outputs

given by the length of unused biofilter bed as well as the breakthrough time which is the working time over which it is necessary to regenerate the fixed bed. In this case, using the covariance decomposition which extends equation (16) to the multivariate case, they proposed both the trace sensitivity indexes which retain additivity as in equation (17) and determinantal sensitivity indexes which consider also the output correlations.

Nonlinear extensions of the linear model (15) follow two main approaches. Keeping homoskedastic independent errors, the first path focusses on generalizing the parametric emulator into nonparametric models. In the case of additive models and independent inputs, the decomposition (8) and the sensitivity indexes (9) may be still used.

The second nonlinearity approach arises when the emulator errors (6) are heteroskedastic and the output uncertainty depends on certain input parameters. For example, going on with the above biofilter example, it has been found that the emulator errors for the length of unused biofilter bed may be modeled as

$$e_3 = \varepsilon\sqrt{\alpha_0 + \alpha_1 u + \alpha_2 u^2} \tag{19}$$

where $\varepsilon$ is a standardized error with unit variance and $u$ is the input parameter given by adsorption particle diameter. Equation (19) shows that the model uncertainty is not constant over the input domain $D$ and the model predictions are more reliable for certain input values.

The sensitivity indexes may account easily for heteroskedasticity. In the biofilter case, extending equation (16) for heteroskedasticity, the index for the adsorption particle diameter is given by

$$\hat{S}_u = \frac{\hat{\sigma}_u^2 \hat{\beta}_u^2}{\hat{\sigma}_y^2} + \frac{\left(\alpha_1 \hat{E}\left(u\right) + \alpha_2 \hat{E}\left(u^2\right)\right)}{\hat{\sigma}_y^2}.$$

Note that in the right hand side, the second term is a part of the residual uncertainty $Var\left(e_3\right)$ which, thanks to the heteroskedastic approach, has been attributed to the adsorption particle diameter.

# 7 $SA$ and calibration

Often a $CM$, being in the form of equation (2) requires appropriate calibration and validation on some observational data sets. For example, Sincock et al. (2003) considered a river water quality model under unsteady flow conditions including a flow component and a water quality component. After calibration on historical data they found that the model performance was insensitive to algal activity while nitrification and sedimentation were important.

We will not go much further into validation issues here, we only remark that one of the steps in validation is the understanding of the performance of the $CM$ with respect to variation of fixed parameters. For example, if the model performance is not sensitive to a parameter component $\theta_j$ then

the observational data are inappropriate for that parameter or the $CM$ is over-parametrized for that application.

## 7.1 Equifinality and GLUE

Hydrological modelling often requires some form of calibration so that the fixed $CM$ parameters $\theta$, in equation (2) are adjusted to get a better fit to some observed data. In this section, we consider methods developed in hydrology, but useful beyond that for various instances of $CM$ calibration and validation. For example McIntyre and Wheater (2004) considered the calibration of a simulation model for monthly total phosphorus in Hun River, China.

Using notation and concepts of section 2, we then have a set of observed data

$$(x_1, \zeta_1), ..., (x_N, \zeta_N)$$

and we want to understand the influence of the parameter vector $\theta = (\theta_1, ..., \theta_h)$ on the forecasting performance of the $CM$ with respect to this data. Such performance is traditionally based on the mean of squared errors

$$\hat{\sigma}^2_{e(\theta)} = \frac{1}{N} \sum_{i=1}^{N} (\zeta_i - f(x_i, \theta))^2$$

but other measures may be used, e.g. mean of absolute errors ($MAE$), maximum relative error, etc.

We then have the so-called *likelihood measure*, $L$ say, discussed by Beven (2001), which is constrained to be zero for *non-behavioural* values of $\theta$ and one for the ideal case of perfect forecasts $\zeta_i = f(x_i, \theta)$. The first example is the truncated forecasting efficiency

$$\begin{aligned} L(\theta) = &\quad 1 - \frac{\hat{\sigma}^2_{e(\theta)}}{\hat{\sigma}^2_{\zeta}} \quad \text{if } L > 0 \\ = &\quad 0 \qquad\qquad\quad \text{else} \end{aligned}$$

which is well known to statisticians as the coefficient of determination $R^2$. A second example is the Box and Tiao measure

$$L(\theta) = \left(\hat{\sigma}^2_{e(\theta)}\right)^{-H}$$

where, $H > 0$ is a subjective shaping coefficient.

Equifinality arises here since it is common in environmental applications that $L(\theta)$ is almost the same for many different values of $\theta$. In other words, we have the well known modelling fact that, different $CM$'s give forecasts which are almost the same with respect to a certain likelihood measure $L$.

Hence, a natural choice is to apply output uncertainty to the new $CM$ given by $f(x, \theta)$ weighted by the likelihood measure. To do this, consider $n$ Monte Carlo simulations, $\theta^*_1, ..., \theta^*_n$ of the possibly multivariate parameter $\theta = (\theta_1, ..., \theta_h)$ with rectangular marginal distributions and consider the normalized likelihood

$$\bar{L}(\theta^*_i) = \frac{L(\theta^*_i)}{\sum_{i=1}^{n} L(\theta^*_i)}.$$

Now, suppose that the quantity of interest is a function $Q = Q\left[f\left(x,\theta\right)\right]$ with weighted Monte Carlo cumulative distribution given by

$$\hat{F}\left(q\right) = \hat{P}\left(Q \leq q\right) = \sum_{i:Q\left(\theta_i^*\right)\leq q} \bar{L}\left(\theta_i^*\right). \tag{20}$$

For example, if $Q = z$, equation (20) allows the computation of the weighted forecasting quantiles. Moreover, if $Q$ is the $i^{th}$ component of $\theta$, namely $Q = \theta_i$, equation (20) gives the marginal cumulative distribution of $\theta_i$. Hence, $SA$ may be performed on graphical grounds by comparing this marginal with the uniform distribution which may be interpreted as the prior Monte Carlo distribution. In particular, for a hypothetical example, Figure 2 shows the reduction in output uncertainty achievable in $\theta_i$ by multivariate calibration of $\theta$.

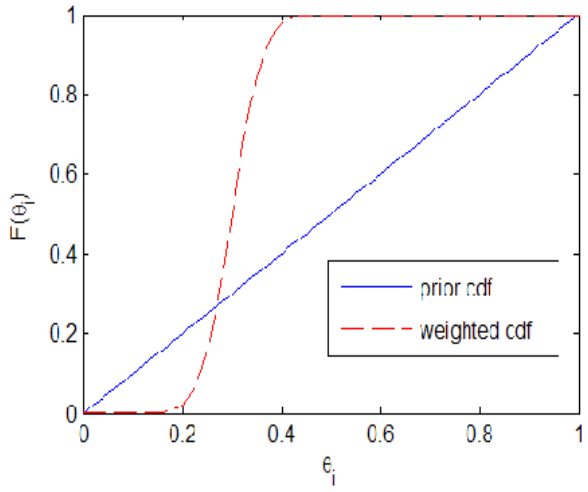



Figure 2: Weighted cumulative distribution of $Q = \theta_i$, $\theta$ Rectangular$(0,1)$, $\hat{F}\left(\theta_i\right) = \sum_{j:\theta_{ij}^*\leq\theta_i} \bar{L}\left(\theta_i\right)$

# 8 Geostatistical $SA$

So far, we have used methods that assume independence of emulator errors between computer runs. In this section, we consider methods which imply more complex modelling and computing time. Hence, they are appropriate for cases where the $CM$ is an "expensive function" and large Monte Carlo computer experiments are not feasible. Moreover, this approach is efficient when we are dealing with a "smooth $CM$" where smoothness here means that $f\left(x\right)$ and $f\left(x'\right)$ are highly correlated for $x$ close to $x'$.

The basic idea of Oakley and O'Hagan (2004) is to consider the model output $f\left(x\right)$ as a stochastic process indexed by the $CM$ input $x$ in the sense that, for a fixed hypothetical sequence of inputs, say $x_1, ..., x_n$, the model

outputs, namely $f(x_1), ..., f(x_n)$, are correlated random variables. This stochastic process representation may be interpreted as Bayesian believes about the $MO$'s, prior to running the code.

Whenever $x$ is assumed to be nonstochastic, it is considered to be unknown with uncertainty distribution $p(x)$. This approach with $x \in D$ can be seen as a geostatistical approach and, in this sense, we will use terms like *space* for $D$. It follows that, the sensitivity quantities introduced in section 3 are stochastic quantities, for example the *spatial* averages (11) and variances (12) are integrals of a stochastic process. Given a set of $MO$'s $(x_1, z_1), ..., (x_n, z_n)$, the above spatial integrals can be estimated by the posterior counterparts of $f(x)$. For example suppose that $\hat{f}(x)$ is an appropriate Bayesian kriging estimate of $f(x)$ given by

$$\hat{f}(x) = E(f(x)|z_1, ...z_n).$$

Then the spatial average (11) is estimated by

$$E^*(z|x_j) = \int_{D_{(j)}} \hat{f}(x) p\left(x_{(j)}\right) dx_{(j)}$$

and similarly, the spatial variance (12) :

$$Var^*(z|x_j) = \int_{D_{(j)}} \hat{f}(x)^2 p\left(x_{(j)}\right) dx_{(j)} - \left(\hat{f}_0\right)^2.$$

To do this, the prior uncertainty on the model output $f(x)$ before actually running the $CM$ is modelled by a Gaussian stochastic process with mean value given by

$$E(f(x)|\beta) = h(x)'\beta$$

where $h(x)$ is a known input transformation as in equation (18) and $\beta$ is a hyperparameter. The covariance function of $f(x)$ is given by

$$Cov\left(f(x), f(x')|\sigma^2\right) = \sigma^2 c(x, x')$$

where $c(x, x')$ is a geostatistical correlation function, for example, in the stationary isotropic case, we have

$$c(x, x') = c(|x - x'|);$$

moreover, $c(0) = 1$ and $c(t)$ decreases with increasing $t$ and, in general, may depend on some further hyperparameters, say $\gamma$.

If $\gamma$ is known and the hyperparameters $(\beta, \sigma^2)$ have prior

$$p\left(\beta, \sigma^2\right) \propto \sigma^2 \tag{21}$$

then $\hat{f}$ and $\hat{c}$ have closed form representation and, marginally to $(\beta, \sigma^2)$, the $MO$'s have a multivariate $t$ distribution. In particular

$$t(x) = \frac{f(x) - \hat{f}(x)}{\sqrt{\hat{c}(x, x)}} \tag{22}$$

has a $t$ distribution with $k+n$ degrees of freedom.

If the prior distribution is not as in (21) or $\gamma$ is unknown, the closed form posterior distribution (22) does not hold, and Markov chain Monte Carlo integration is required giving a considerably increased computational burden. To avoid this, it is common practice in Bayesian Kriging to use a plug-in approach based on substituting the posterior estimate for $\gamma$, say $\hat{\gamma}$, into $c(x, x')$ and, conditionally on this use the above methods.

### 8.1 *DOE*

In this frame, the input design is different from the Monte Carlo approach of section 5 because here, $x$ is nonstochastic but the integrals to be estimated are stochastic ones. As a matter of fact, Sacks et al. (1989) discuss the extension of the classical *DOE* of section 4 to *DOE* for stochastic processes. In general terms, it is based on the optimization of the integrated mean squared error

$$IMSE(x_1, ..., x_n) = \int_D E\left(\left(f(x) - \hat{f}(x)\right)^2 | x_1, ..., x_n\right) p(x)\, dx.$$

giving both sequential and nonsequential design algorithms are reviewed. Since the *MO*'s are not independent, algorithms are nonstandard and may be time consuming. Of course this is worthwhile if the computer runs are more expensive.

## References


[1] Beven K.J. (2001) *Rainfall-runoff modelling.* Wiley.

[2] Becciu G., Bianchi A., Fassò A., Fassò C. A., Larcan E. (2000) Quick calculation of minimum in-stream flow in drainage basins of Central Alps. In Proceeding of New Trends in Water and Environmental Engineering for Safety and Life, Maione, Majone Lehto & Monti (eds), 2000 Balkema, Rotterdam.

[3] Box G.E.P., Hunter W., Hunter J. (1978) *Statistics for experimenters. An introduction to design, data analysis and model building.* Wiley

[4] Chan K., Tarantola S., Saltelli A., Sobol' I. (2000) Variance-based methods. In *Sensitivity Analysis,* Saltelli, A., Chan, K., Scott, M. eds. Wiley.

[5] Fassò A., Esposito E., Porcu E., Reverberi A.P., Vegliò F. (2002) Sensitivity Analysis of heavy metals biofilter models, GRASPA Working paper n.14, www.graspa.org

[6] Fassò A., Esposito E., Porcu E., Reverberi A.P., Vegliò F. (2003) Statistical Sensitivity Analysis of Packed Column Reactors for Contaminated Wastewater. *Environmetrics.* Vol. 14, n.8, 743 - 759.

[7] Fassò A. and Nicolis O. (2005) Space-time integration of heterogeneous networks in air quality monitoring. Proceeding of the Italian Statistical Society Conference on "STATISTICA E AMBIENTE", Messina, 21-23 September 2005, Vol.1.

[8] Fassò A. Perri P.F. (2001) Sensitivity Analysis in A. El-Sharaawi & W. Piegorsch (eds) *Encyclopedia of Environmetrics*. Wiley. Vol.4, 1968-1982.

[9] Jamieson R., Gordon R., Joy D., Lee H. (2004) Assessing microbial pollution of rural surface waters: A review of current watershed scale modeling approaches, *Agricultural Water Management* Volume: 70, Issue: 1, October 15, 2004, pp. 1-17.

[10] Helton J.C., Davis F.J., Johnson J.D. (2005) A comparison of uncertainty and sensitivity analysis results obtained with random and Latin hypercube sampling. *Reliability Engineering and System Safety.* 89, 3, 305-330

[11] Homma T. and Saltelli A. (1996) Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering and System Safety*, **52**, 1-17.

[12] Huntington D. E., Lyrintzis C. S. (1998) Improvements to and limitations of Latin hypercube sampling (STMA V40 3950), *Probabilistic Engineering Mechanics*, 13 , 245-253.

[13] Kennedy M.C. O'Hagan A. (2001) Bayesian calibration of computer models, *JRSS-B*, **63**, 425-464.

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

[15] McIntyre N.R. and Wheater H. S. (2004) Calibration of an in-river phosphorus model: prior evaluation of data needs and model uncertainty. *J. Hydrology*, 290, 100-116.

[16] Miller B.M., Peirson W.L., Wang Y.C., Cox R.J. (1996) An Overview of Numerical Modelling of the Sydney Deepwater Outfall Plumes. *Marine Pollution Bulletin* Volume: 33, Issue: 7-12, pp. 147-159.

[17] Norton J.P., Newham L.T., Andrews F.T. (2004) Sensitivity Analysis of a Network-Based, Catchment-Scale Water-Quality Model. *Transactions of the 2nd Biennial Meeting of the International Environmental Modelling and Software Society*, iEMSs: Manno, Switzerland, 2004. ISBN 88-900787-1-5.

[18] Oakley J. and O'Hagan A. (2004) Probabilistic sensitivity analysis of coplex models: a Bayesian approach. *J.R.Statist. Soc. B*, **66**, 3, 751-769.

[19] Olyphant G.A. and Whitman R. (2004) Elements of a Predictive Model for Determining Beach Closures on a Real Time Basis: The Case of 63rd Street Beach Chicago. *Environmental Monitoring and Assessment* Volume: 98, Issue: 1, pp. 175-190.

[20] Reynolds C. S. (1999) Modelling phytoplankton dynamics and its application to lake management, *Hydrobiologia* Volume: 395, February, pp. 123-131.

[21] Sacks J., Welch W.J., Mitchell T.J., Wynn H.P. (1989) Design and Analysis of Computer Experiments, *Statistical Science*, Vol. 4, No. 4. 409-423.

[22] Saltelli A., Chan K., Scott M. (2000) *Sensitivity Analysis*, Wiley, New York.

[23] Sincock A.M., Wheater S.S., Whitehead P.G. (2003) Calibration and sensitivity analysis of a river water quality model under unsteady flow conditions. *J. Hydrology*, 277, 214-229.

[24] Stein M.L. (1987). Large sample properties of simulations using latin hypercube sampling. *Technometrics*, 29(2):143-151, .

[25] Vinten A.J.A., Lewis D.R., McGechan M., Duncan A., Aitken M., Hill C., Crawford C. (2004) Predicting the effect of livestock inputs of E. coli on microbiological compliance of bathing waters. *Water Research* Volume: 38, Issue: 14-15, pp. 3215-3224.

[26] Whitehead P.G., Williams R.J., Lewis D.R. (1997) Quality simulation along river systems (QUASAR): model theory and development. *The Science of the Total Environment* 194/195, 447–456.

[27] Wu J., Hamada M. (2000) *Experiments: planning, analysis and parameter design optimization.* Wiley.