Estimation of human impact
in the presence of natural fluctuations
EC Fifth Framework Program
SCA Project IST-1999-11313

Deliverable 16

Model reduction techniques for quantification of uncertainty

December 2000

Document Reference:	IMPACT/JRC/WP6/D16
Version:	1
Security:	Pub.
Status:	Final
Authors:	M. Ratto, S. Tarantola, A. Saltelli Joint Research Centre of European Commission, Institute for Systems, Informatics and Safety, Ispra

TABLE OF CONTENTS

1. Introduction

1.1 Uncertainty and sensitivity analysis and model reduction

One of the main objectives of the IMPACT project is to develop new methodologies for the normalisation of observed natural fluctuations based on the merging of mechanistic models and statistical procedures.

Computational models are used to give a simplified mathematical representation of reality. Model input is subjected to many sources of uncertainty including errors of measurement, inadequate sampling resolution, etc. Furthermore, the model itself can include conceptual uncertainty, i.e. uncertainty in model structures, assumptions and specifications. All this imposes a limit on our confidence in the response, or output, of the model. Good modelling practice requires the modeller to provide an evaluation of the confidence in the model predictions; possibly assessing the uncertainties associated with the outcome (response) of the model itself.

Uncertainty Analysis (UA) and Sensitivity Analysis (SA) are therefore prerequisites for model building in any field where models are used. UA allows assessing the uncertainty associated with the model response as a result of uncertainties in the model input. SA is aimed at establishing how the variation in the model output can be apportioned to different sources of variation, in order to establish how the given model depends upon the information fed into it. SA can be useful in model building for identifying, on one side, the relevant parameters, and on the other those who do not drive significant variation on the output. In this way, SA can be used to reduce models: unimportant factors can be fixed to their nominal values and, if parameters are clearly connected to particular processes included in the model, entire parts of the models can also be eliminated or simplified.

The role of UA-SA methods in framework of the IMPACT project is therefore clearly identified as the intermediate step in the merging of mechanistic models and statistical procedures, in which the original model is reduced by allowing unimportant factors to be fixed or eliminated.

1.2 Aims of the report

The present deliverable is focused on the uncertainty aspects of mechanistic modelling.

The methodological framework is identified through the following steps:

1) study the empirical distribution of the model outputs due to propagation of the various input uncertainties through the models themselves. This study is actually an uncertainty assessment, aimed at testing the overall robustness of both the underlying model and the available data. The basic step in the reliability assessment procedure consists of establishing the overall level of uncertainty in the model predictions and, if desired, stating whether they remain within some desired target bounds (e.g. for policy purposes). It is also possible to verify whether the uncertainties come from subjective modelling assumptions, from non-reducible input uncertainties, from poor resolution and so on.
2) investigate model relevance. The model output prediction variation can be apportioned according to source by using sensitivity analysis SA. This study is able to reveal to what extent had uncertain model factor affects the model response. This investigation aims at trying to distinguish "live" components of the model, which drive model response and are hence "relevant", from "dead" ones, which make no contribution to the variation in the model predictions. As a consequence, SA is the basis for the elimination of unneeded complexity from the model and therefore for model reduction.
3) investigate how various resolution levels in the input data can affect the output uncertainty. Global UA and SA can be employed by using different spatial or temporal resolution levels for input data in order to identify - given the task- the optimal level of resolution for spatially referenced data. This could assist in optimising the costs of data collection procedure for decision-making purposes, and help to identify most sensitive data gaps.
4) model reduction criteria. Criteria for model reduction are straightforward from the steps 2) and 3).

The methodological approach identified is applied in test cases of IMPACT. In particular, the analysis of soil nitrogen turnover applying the SOIL/SOILN models is described in the present deliverable and the Elbe River case study applying the WAMPUM model is described in deliverable n. 17 (Ratto et al., 2000b).

1.3 List of publications

Participation to international conferences

1) M. Ratto, N. Giglioli, S. Tarantola, U. Callies, Å. Forsman, Sensitivity analysis and environmental data normalisation: eutrophication case studies, Abstract submitted for the *Third International Symposium on Sensitivity Analysis of Model Output SAMO* 2001, Madrid, June 2001.
2) M. Ratto, N. Giglioli, S. Tarantola, U. Callies, Å. Forsman, Abstract accepted for the *European Safety and Reliability International Conference ESREL 2001*, Torino, September 2001.

Papers for submission to the open literature are also in preparation:

1) Ratto M., Tarantola S., Saltelli A., Sensitivity analysis in model calibration: GSA-GLUE approach, (2000), submitted to *Computer Physics Communications*.

2 Methodologies

Some basic concepts about UA and SA are introduced hereafter. A thorough review on this matter can be found in (Helton, 1993) and (Hamby, 1994).

A computational model, describing a general system, is under investigation. The model has one output variable Y and k input factors, $\mathbf{X} = (X_1, X_2, \ldots, X_k)$, that represent all possible sources of uncertainty that affect the model output. The input factors are variables that correspond either to model inputs (i.e. the data fed into the model) or model parameters. For the purpose of UA and SA the input factors are treated as random variables with a probability density function (pdf), which is assumed known a-priori.

The computational model, i.e. the relationship between the input factors and the output under study, can be represented by a mathematical operator, $f(\cdot)$, which maps the k-dimensional space of the input factors Ω to that of the output variable Y

$$Y = f(X_1, X_2, \ldots, X_k).$$

The output Y has its own pdf, whose estimation is the purpose of uncertainty analysis. The investigator can quantify the impact of input uncertainties on the model response and assess whether or not the response meets the required standards of precision.

2.1 Uncertainty Analysis

Various methods are available for evaluating output uncertainty (Helton, 1993). In the following, the focus is on the Monte Carlo based method, which allows exploring the full range of variation for the input factors and is model-free (no assumptions are required upon the model structure).

The Monte Carlo (MC) method is based on performing multiple evaluations of the model with randomly selected model inputs. The MC-based UA involves four steps:

1. assign a distribution (pdf) to each input factor X$_i$;
2. generate a sample of size N ($\mathbf{X}^j$, $j = 1,..., N$) from the factors' distributions according to an appropriate design;
3. evaluate the model at each sample point $\mathbf{X}^j$;

4. analyse the resulting output values Y_j.

2.2 Sensitivity Analysis

Many techniques for SA have been proposed (e.g. linear regression or correlation analysis, measures of importance, sensitivity indices, screening, etc.). A thorough description of such techniques can be found in (Saltelli et al., 2000). In the following regression/correlation analysis and variance-based techniques are considered. These techniques have been applied for the SOIL/SOILN case study described in Section 3.

2.2.1 Regression analysis

A multivariate sample of the input **x** is generated by some sampling strategy and the corresponding sequence of *N* output values is computed using the model under analysis. If a linear regression model is being sought, it takes the form

$$y_i = b_0 + \sum_j b_j x_{ij} + e_i \quad j = 1,2,...,k \tag{1}$$

where the b_j*'s* are the regression coefficients that must be determined and ε_i is the error (residual) due to the approximation. One common way of determining the coefficients b_j*'s* is to use least squares analysis (Draper and Smith (1981)).

Once the b_j's are computed, they can be used to indicate the importance of individual input variables x_j with respect to the uncertainty in the output *y*. In fact, assuming that **b** has been computed, the regression model can be rewritten as

$$(y - \bar{y}) / \hat{s} = \sum_j (b_j \hat{s}_j / \hat{s})(x_j - \bar{x}_j) / \hat{s}_j \tag{2}$$

where

$$\bar{y} = \sum_i y_i / N, \; \bar{x}_j = \sum_i x_{ij} / N, \tag{3}$$

$$\hat{s} = \left[\sum_i (y_i - \overline{y})^2 /(N - 1) \right]^{1/2} , \ \hat{s}_j = \left[\sum_i (x_{ij} - \overline{x}_j)^2 /(N - 1) \right]^{1/2} \tag{4}$$

The coefficients $b_j \hat{s}_j / \hat{s}$ are called *Standardised Regression Coefficients* (*SRC's*). These can be used for sensitivity analysis (when the x_j's are independent) as they quantify the effect of varying each input variable away from its mean by a fixed fraction of its variance while maintaining all other variables at their expected values.

2.2.2 Correlation measures

Another interesting measure of importance is given *by Partial Correlation Coefficients* (*PCC's*). These coefficients are based on the concepts of correlation and partial correlation. The partial correlation coefficient between the output variable *Y* and the input variable X_j is obtained from the use of a sequence of regression models. First the following two models are constructed:

$$\hat{Y} = b_0 + \sum_{h \neq j} b_h x_h \text{ and } \hat{X}_j = c_0 + \sum_{h \neq j} c_h x_h . \tag{5}$$

Then, the results of these two regressions are used to define the new variables $Y - \hat{Y}$ and $X_j - \hat{X}_j$. The partial correlation coefficient between Y and X_j is defined as the correlation coefficient between $Y - \hat{Y}$ and $X_j - \hat{X}_j$ (Helton, (1993)). Thus, the PCC's provide a measure of the strength of the linear relationship between two variables after a correction has been made for the linear effects of the other variables in the analysis. In other words, the PCC gives the strength of the correlation between Y and a given input X_j after adjusted for any effect due to correlation between X_j and any of the X_i, $i \neq j$. Since SRC's measure the effect on the output variable that results from perturbing an input variable by a fixed fraction of its standard deviation, PCC's and SRC's provide related but not identical measures of variable importance. In particular:

- ♦ SRC's are sensitive to all input distributions; the SRC's can provide a decomposition of the output variance according to the input factors;

- ♦ PCC’s provide a measure of variable importance that tends to exclude the effects of other variables.

However, for the case in which the input variables are uncorrelated, the order of variable importance based either on SRC's or PCC's (in their absolute values) is exactly the same.

2.2.3 Stepwise regression analysis

Stepwise regression analysis (Helton (1993)) provides an alternative to constructing a regression model containing all the input variables. A sequence of regression models is constructed using the following steps:

(i) the first regression model contains the most influential (on the output variable) input variable;

(ii) the second model introduces the next most influential input variables (given the one from the previous step);

(iii) the third model introduces a third variable (given the variables from steps (i) and (ii));

and so on, until the point is reached at which subsequent models are unable to increase, meaningfully, the amount of variation in the output variable that can be accounted for.

The model coefficients of determination R_y^2 computed at successive steps of the analysis provide a measure of variable importance by indicating how much of the variation in the dependent variable can be accounted for by all variables selected at each step. Also the individual SRC's in the individual regression models provide an indication of variable importance. When the input variables are uncorrelated, the size of the coefficient of determination R_y^2 attributable to the individual variables, the absolute values of the SRC's, and the absolute values of the PCC's, are identical. When variables are correlated, care must be used in the interpretation of the results of a regression analysis since the regression coefficients can change in ways that are basically unrelated to the importance of the individual variables as correlated variables are added to and deleted from the regression model, i.e. the results are conditional on what is already in the model.

A delicate part of this technique consists in deciding when to stop the construction process of the consecutive regression models. Calculation of the regression coefficients associated with the input variables, in order to check whether they are significantly different from zero or not, can be a useful method to adopt. F statistic values are conventionally used to control which variables should be included in the model (Draper and Smith (1981), p. 93). More details about the stepwise regression analysis can be found in Draper and Smith (1981).

2.3 Model reduction

SA allows identifying, on one side, the relevant parameters, and on the other those who do not drive significant variation on the output. As a consequence, SA is the basis for the elimination of unneeded complexity from the model and therefore for model reduction. Removing the unimportant factors / model structures, quantification of uncertainty can be performed with a reduced model.

3 Case study. SOIL/SOILN models.

3.1 Short description of the model

SOIL/SOILN are a chain of models to be implemented in series.

SOIL is a hydrological model, which solves the mass and energy conservation balances at different layers of a soil with given characteristics (clay-sand ratio, soil depth, etc.) (Jansson & Halldin, 1979). The inputs to the SOIL model are time series of daily meteorological data (precipitation, temperature, cloudiness, wind, etc.) for given geographical region and soil physico-chemical properties. Outputs are daily time series of water content, water fluxes and temperature in the different layers of the soil.

The basic structure of the model is a depth profile of the soil. Processes such as snowmelt, interception of precipitation and evapo-transpiration are examples of important interfaces between soil and atmosphere. Two coupled differential equations for water and heat flow represent the central part of the model. The basic assumptions behind these equations are very simple: the law of conservation of mass and energy and flows occur as a result of gradients in water potential (Darcy's Law) or temperature (Fourier's law).

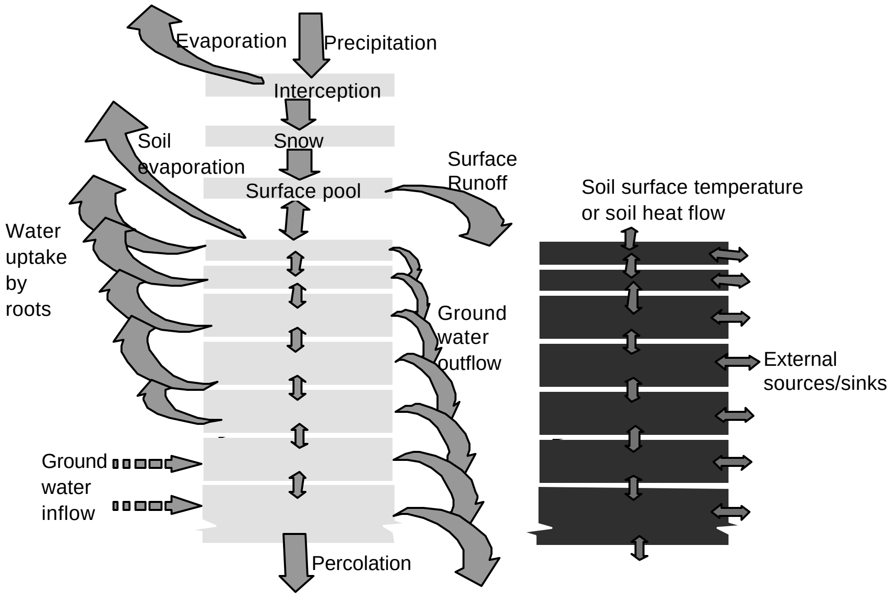

Figure 3.1 Mass balance (left) and heat balance (right) of the SOIL model.

The SOILN model simulates major C and N-flows in agricultural and forest soils and plants (Eckersten et al., 1994) and provides, as a main output, the total nitrogen leached. The model has a daily time step and simulates flow and state variables on a field level. Input variables are daily data on air temperature and solar radiation, management data and variables on soil heat and water conditions simulated by model SOIL. The model depends also on the soil properties (clay content, soil depth, organic content), the rate constants of various chemical-biological processes (denitrification, mineralisation, etc.) and factors depending on the human activity (input of fertilisers and manure, crop type, etc.).

The soil is divided into layers. In each layer mineral N is represented by one pool for ammonium N (immobile) and one for nitrate N (transported with water fluxes). Water flows bringing nitrate between layers is the process finally responsible for N leaching.

3.2 Methodological approach

3.2.1 Definition of the sensitivity analysis role in the SOIL/SOILN model study

The main objective is the study of the SOILN module. The SOIL model has the main aim to provide the input to the SOILN model (see Figure 3.2). The main aspect to consider is that only global quantities are searched, in particular the total nitrogen loss averaged in a period of several years. So, from a very huge quantity of input data only a few global output values are of interest. In this context, the role of sensitivity analysis is defined, as the tool for determining the hydrological data requirements with respect to the long-term average of nitrogen losses. This information will be useful for the lumping of SOIL model (which is the most expensive in terms of computational time) and for evaluating the possibility of applying other models than SOIL. Furthermore, the knowledge of the minimum quantity of data really necessary to implement SOILN model will also allow the use of measured hydrological data, instead of hydrological model simulations, as the input to SOILN.

3.2.2 Definition of a reduced input factor set for SA

Considering SOILN model, implies that a sensitivity analysis has to be performed with a single output variable (the average nitrogen leached) and time series of daily data for many years as the inputs. To overcome the complexity of this task, the hydrological time series in the different layers of the soil have to be characterised by a limited set of parameters. These parameters have been firstly identified: global average over the period under consideration (e.g. 30 years), variance of the annual averages over the 30 years, maximum and minimum values over the 30 years. In this way the sensitivity analysis structure is well defined: input factors are the characteristic parameters of the time series and output is the global nitrogen loss over a long period (e.g. 30 years). As far as the depth variation of the parameters is concerned, we simply considered parameters at layers 1, …, *n*, instead of representing them through a fitting function, whose parameters may not have a clear meaning. In the present analyses 5 layers were considered, for a total of 20 hydrological time series accounted for the SA:

- ✓ 5 series of water flow (notation: WF);
- ✓ ratio between actual and potential transpiration from soil (notation: transp);
- ✓ 5 series of temperature (notation: T);
- ✓ 5 series of water content (notation: WC);
- ✓ 4 series describing water flow in deep percolation, runoff (2 series), water flow in surface pools.

Considering 4 characterisations for each time series, we have 80 input factors. The index of the layers increases from upper layers (*i*=1 for top layer) to bottom layers (*i*=5) for the deepest layer. So, AVtransp stands for the average transpiration, AVWC2 stands for the average water content in the 2nd layer, VARWFsurf stands for the variance of the water flow at the surface, and so on.

After the first analyses were performed (see below), it has been found that among the 80 characterising parameters of the hydrological time series, only 2 global averages (1 for water content and 1 for water flux or transpiration) were relevant. To get a more complete SOILN model input-output mapping, a more detailed representation of water content has been identified. The water content in different soil layers are summarised in frequency tables where the water content (%) is divided in 50 equally large classes and the number of days within each class is counted, there is one table for each layer and month. The frequency tables together with two global averages (e.g. water content in the first layer and water flow from the deepest layer) have been subsequently chosen as the explanatory variables.

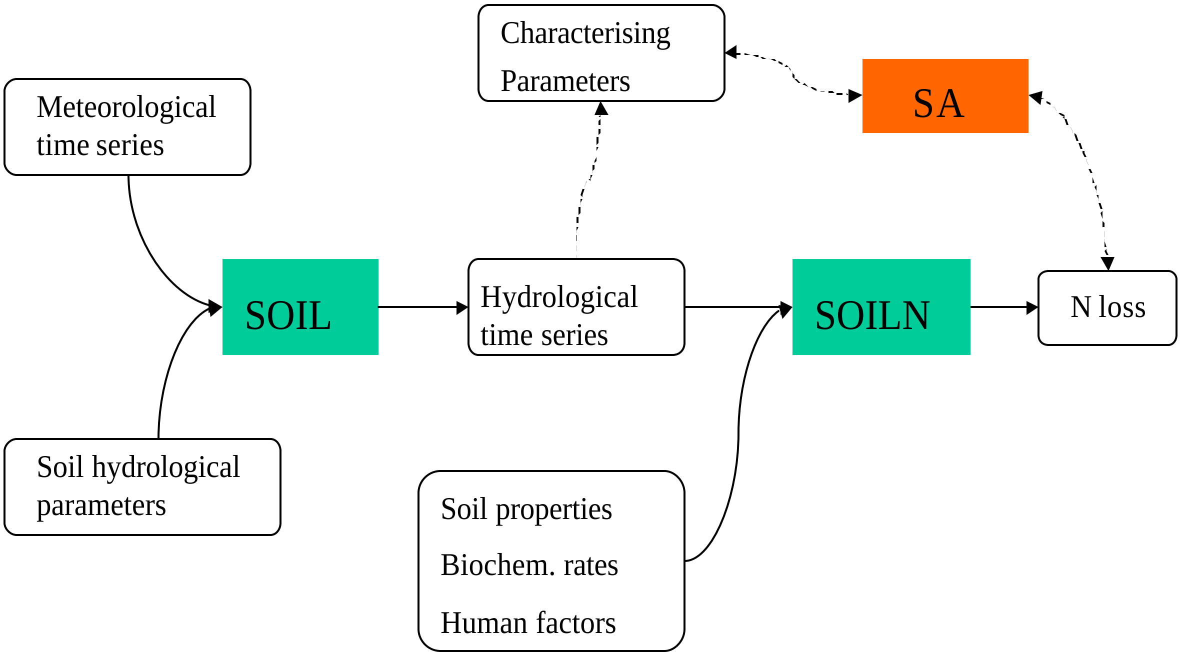

Figure 3.2: SA scheme for the SOILN model.

3.2.3 Sampling methodology

Since the hydrological time series have to fulfil the mass and heat conservation laws, the characteristic parameters of the series cannot be sampled in a purely random way. Instead, the SOIL model has to be applied to sample different replications of the hydrological time series. Such replications can be obtained using artificially generated 30-years meteorological data. Fixed soil conditions (sandy soil) and a fixed class of climate properties (climate of the southern part of Sweden) have been considered. A total of 96 realisations of hydrological time series have been generated with SOIL. Analysing the correlation structure of the sample showed a very strong correlation between the input factors.

3.2.4 SA methodology

Regression/correlation based sensitivity analysis methodology as described in paragraphs 2.2.1-5. It is not possible to perform a variance based analysis, since the input factor have a strong correlation structure given by the mass/heat conservation laws.

In Figure 3.2 the SA scheme is represented by means of a flow sheet, in order to clarify the flow of information between models SOIL and SOILN and object of the present SA. The average nitrogen loss is used as response variable and regression models (stepwise regression analysis) between the characterising parameters of the hydrological time series and the total nitrogen loss are fitted to the simulation results. This kind of analysis allows studying sensitivity of nitrogen loss with respect to a variation in the properties hydrological time series. At the same time, regression models are reduced models mapping the main properties of the hydrological time series to the total average nitrogen loss.

3.3 UA of the average nitrogen leached over 30 years

The frequency histogram of the N-loss from the soil layer, averaged over a period of 30 years is shown in Figure 3.3. The distribution has a smooth behaviour, near to normality. The main statistical properties of the N-loss are:

	Min	**Max**	**Mean**	**Std.dev**
N-loss [g (m^2 day)$^{-1}$]	0.0102	0.0135	1.144E-02	5.230E-04

The standard deviation is about 5% of the mean of the distribution.

When the effect of reducing time resolution in the input is analysed, the prediction of the N loss is affected by a systematic error in excess, significant already for the 2 days averaging (see Table 3.15 in section 3.5). This means that accurate predictions require high time resolution in the input data. On the other hand, the relative error of the N loss prediction with averaged input data is bounded at the 5% even for the 32 days averaging. This may be acceptable remembering that the prediction with coarse time resolution in the input time series is conservative and that the standard deviation of the prediction itself is of the order 5%, too. Further discussion about time resolution is developed in the SA section 3.5.

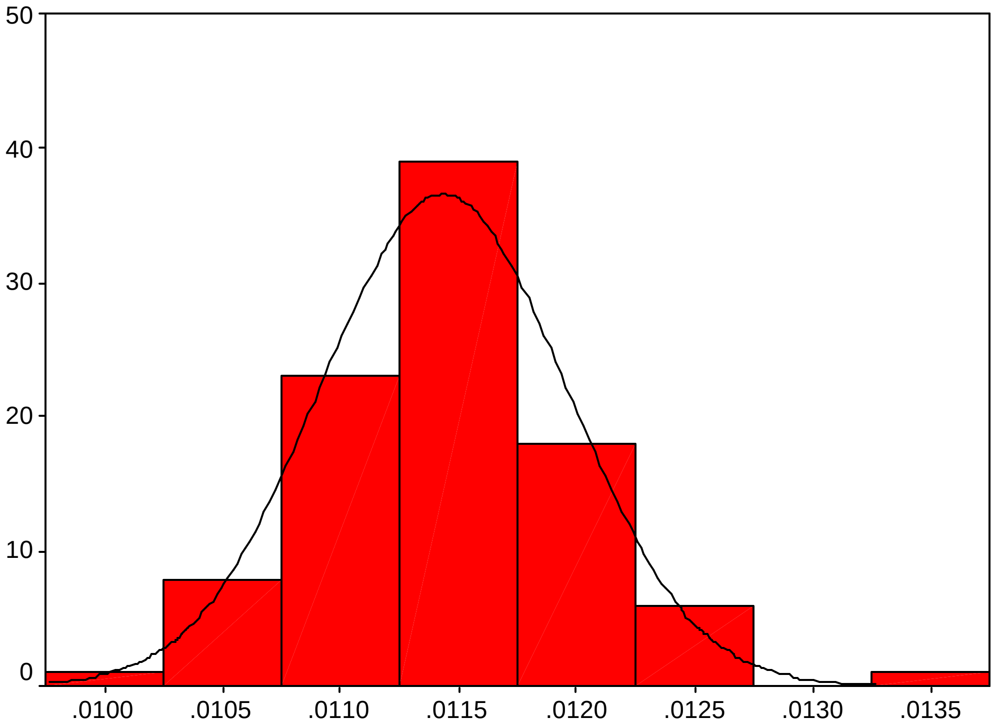

Figure 3.3: Histogram of the average N-loss from the soil layer [(g of N) / (m^2 day)].

3.4 SA of SOILN to the hydrological time series

3.4.1 Hydrological time series characterisations

With the average water flow AVWF5 as the only explanatory variable, the regression model explains 54 % of the variation in the response variable, Table 3.1. When the average water content in the first layer AVWC1 is added to the model, the R-square increases to 68 %, Table 3.2. Standardised regression coefficients (SRC) are also shown.

By considering a stepwise regression by considering all 64 time series characterisations, parameters AVWFsurf and AVtransp give the best R-square, Table 3.3. Results are qualitatively almost identical, so the choice between the 2 possibilities is arbitrary. In the following we will apply AVWFsurf and AVtransp.

#	Entered	Removed	R**2	Partial R**2	T	Prob > |T|
1	AVWFL5	.	0.545	0.545	10.661	0.0001

Table 3.1: Regression model with average water flow from the deepest layer (AVWFL5) as explanatory variable.

#	Entered	Removed	R**2	Partial R**2	T	Prob > |T|	SRC
1	AVWFL5	.	0.545	0.545	4.638	0.0001	0.373
2	AVWC1		0.69	0.146	6.643	0.0001	0.522

Table 3.2: Regression model with average water flow from the deepest layer and average water content (AVWC1) as explanatory variables.

#	Entered	Removed	R**2	Partial R**2	F	Sig. Var. F	SRC
1	AVWFSURF	.	0.685	0.685	204.672	0	0.657
2	AVTRANSP	.	0.759	0.073	28.269	0	0.32

Table 3.3: Stepwise regression analysis using the 64 time series characterisations.

3.4.2 Water content frequency tables

Next, the water content frequency tables from the *first* layer only were added to the list of explanatory variables and a model was fitted with the stepwise regression method. In Figure 3.4-6, the shapes of the frequency tables in the upper layer at different months are shown in four representative cases. Such shapes are represented through the mean (dashed line) and standard deviation (error bar) over the 96 replications. In Figure 3.4-6 the most important frequencies detected with the stepwise regression analysis are highlighted.

The results of the stepwise regression performed by considering global averages of water flow at the surface and of transpiration and the water content frequency tables in the upper layer are shown in Table 3.4: most of the variation in the response variable is now explained.

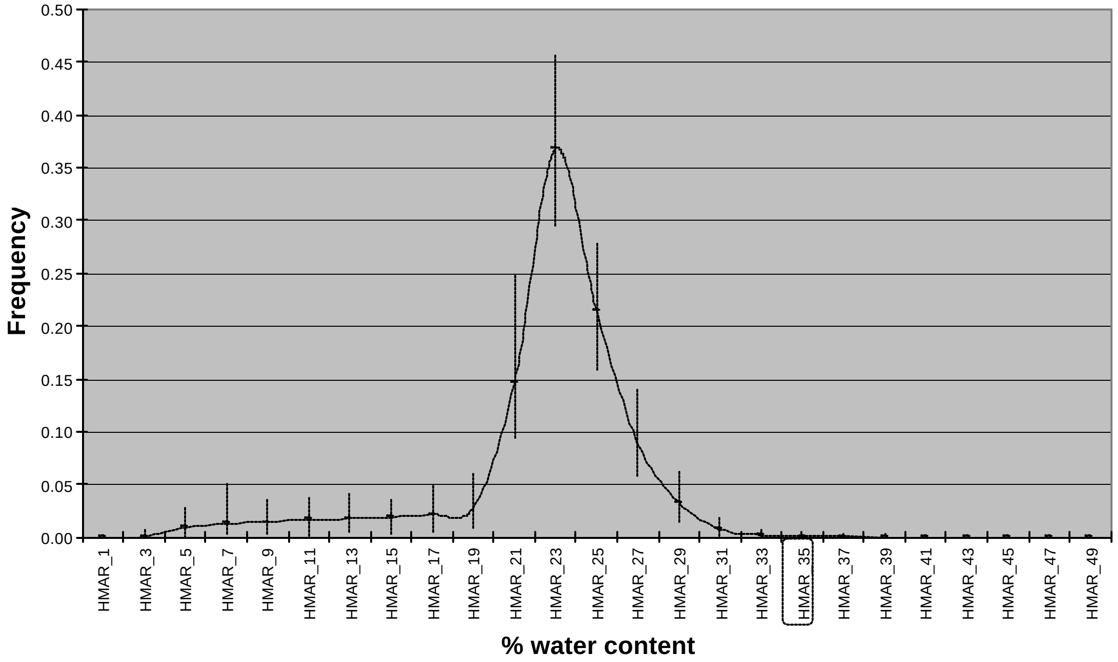

Figure 3.4: mean and standard deviation of the water content frequency tables in March ('dry' month).

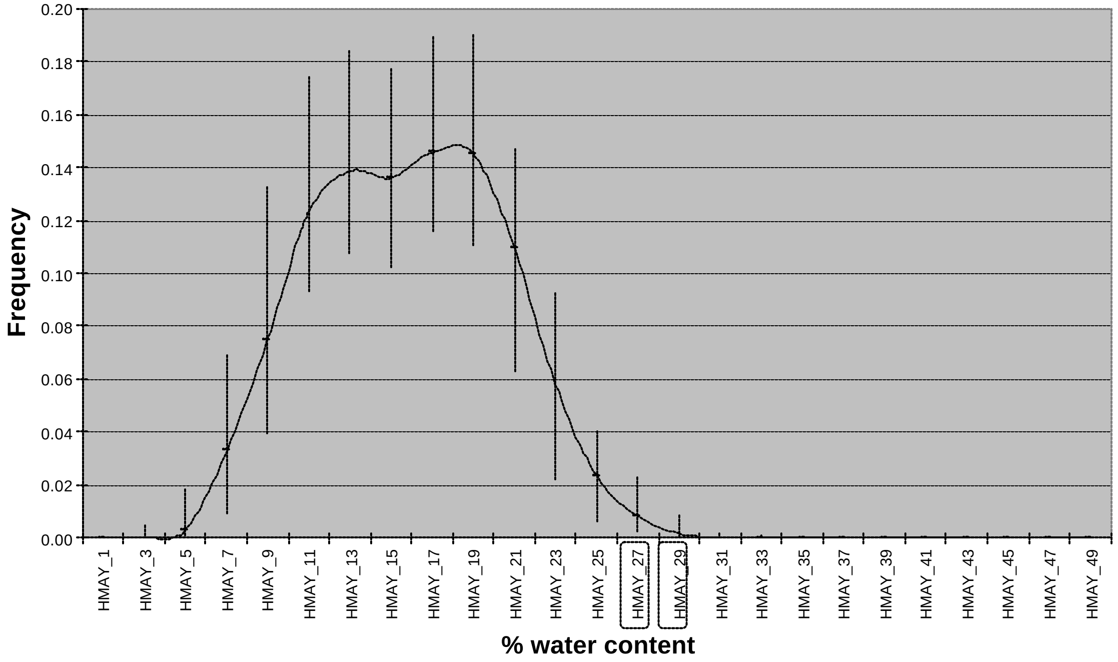

Figure 3.5: mean and standard deviation the water content frequency tables in May ('dry' month).

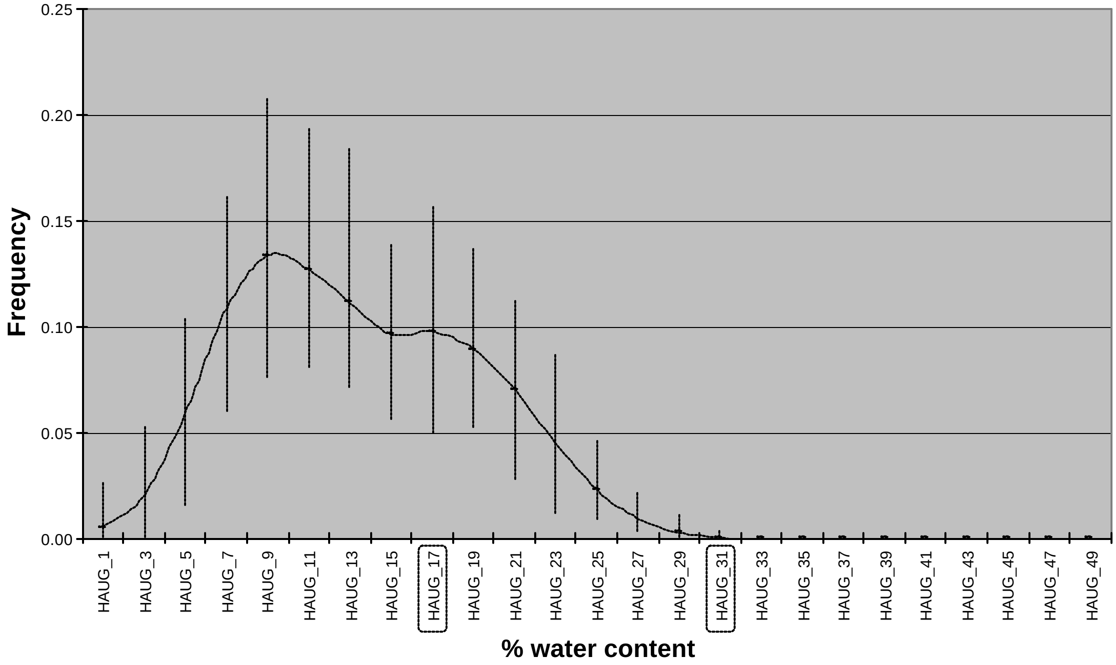

Figure 3.6: mean and standard deviation of the water content frequency tables in August ('dry' month).

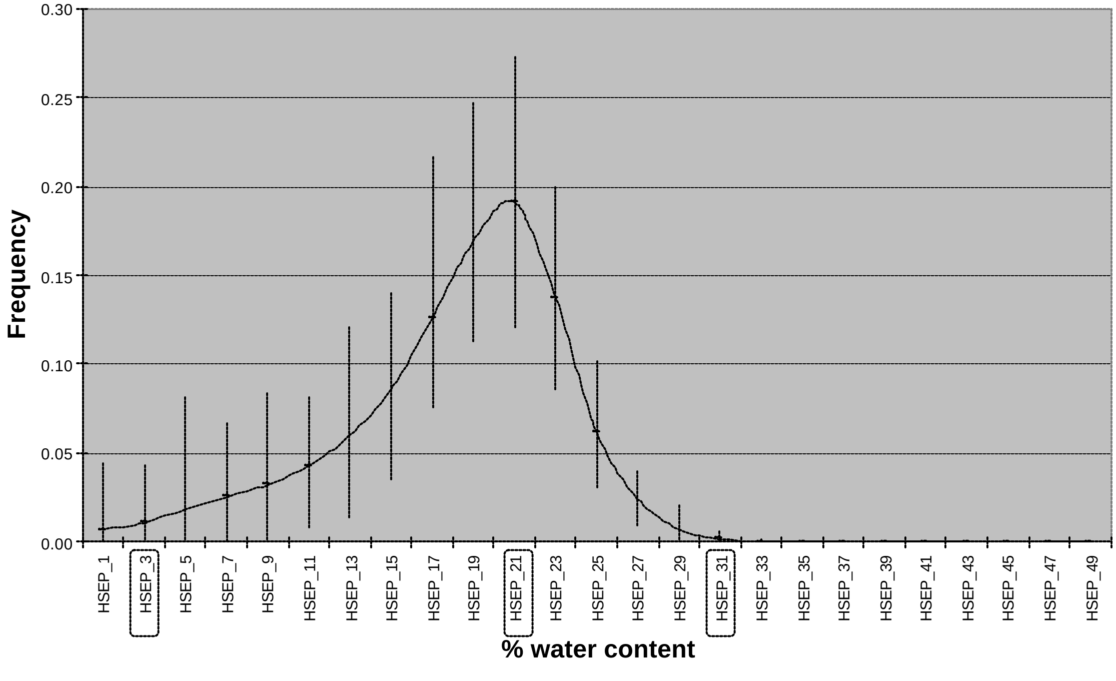

Figure 3.7: mean and standard deviation of the water content frequency tables in September ('wet' month).

#	Entered	Removed	R**2	Partial R**2	F	Sig. Var. F	SRC
1	AVWFSURF		0.685	0.685	204.672	0	0.695215
2	HMAY_27		0.796	0.111	50.393	0	0.210063
3	AVTRANSP		0.834	0.038	21.156	0	0.090541
4	HAUG_17		0.856	0.022	13.715	0	0.127839
5	HSEP_31		0.876	0.02	14.728	0	-0.13342
6	HNOV_21		0.887	0.011	8.413	0.005	0.117627
7	HAUG_31		0.896	0.009	7.459	0.008	-0.06957
8	HSEP_21		0.905	0.009	8.26	0.005	0.107525
9	HMAR_35		0.911	0.006	6.129	0.015	0.096387
10	HMAY_29		0.917	0.006	6.357	0.014	0.105322
11	HSEP_3		0.923	0.006	6.46	0.013	-0.0939
12	HJUN_27		0.927	0.004	5.017	0.028	0.081351
13	HOCT_19		0.932	0.004	5.047	0.027	0.072208

Table 3.4: Results from stepwise regression, considering only global averages and water content frequencies from the upper layer as explanatory variables. The numbers at the end of the variable names refers to the midpoints of the water content classes.

Replacing the water content tables with frequency tables from deeper layers resulted in lower R-square values with the lowest value for the deepest soil layer. The water content in the soil is important to the nitrogen loss mainly because it affects the mineralisation process. Most of the mineralisation takes place in the upper soil layer because of the high content of humus; this can explain the decreasing explanation rates.

3.4.2.1 Grouping of frequency tables

Further stepwise regression models have been considered, where the 50 water content classes are grouped to five and four classes respectively. The classes were combined by taking for reference the soil moisture response function (Figure 3.8). The curve in Figure 3.8 represents the 'efficiency' of mineralisation, scaled from 0 to 1, as a function of the moisture content in the layer. The grouping is performed under the hypothesis that mineralisation is the process mainly controlling nitrogen loss and the mineralisation response curve is assumed as the main interpretation key for synthetic description of the mapping of water content frequency tables through the SOILN model.

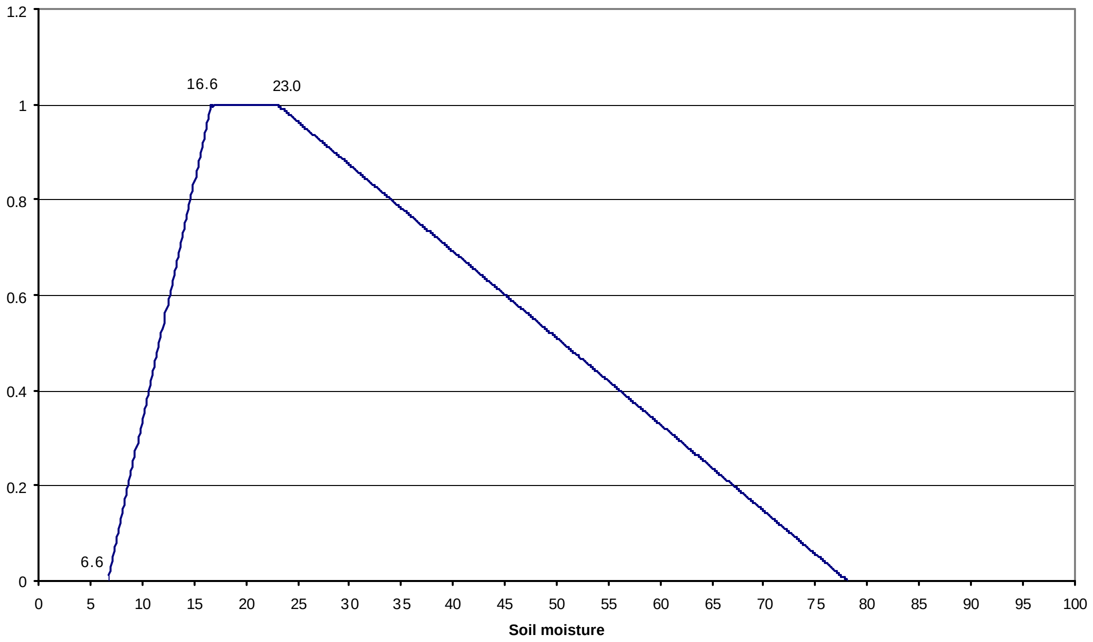

Figure 3.8: Soil moisture response curve for mineralisation in a sandy soil.

In addition, the monthly water content frequency tables are replaced with quarterly tables (Q1=GEN-MAR and so on). The results for the two groupings (5 and 4 classes respectively) are shown in Table 3.5-6. The explanation rates decreases to 82% of the total model output variation. Considering the very strong reduction of the input factors, this seems to suggest that for the prediction of the total average nitrogen loss over 30 years only grouped input factors can be considered. It seems also that taking for reference the mineralisation response curve is effective. In fact, it is interesting to note that the signs of the SRC's can be (partially) interpreted in the light of the mineralisation response curve, e.g. the sign of the SRC for HQ3_16_24 (i.e. the frequency of water content corresponding to the optimal mineralisation) is positive. The same soundness holds also true for the global averages of water flow and transpiration: their increase let N-loss increase.

#	Entered	Removed	R**2	Partial R**2	F	Sig. Var. F	SRC
1	AVWFSURF	.	0.685	0.685	204.672	0	0.615
2	AVTRANSP	.	0.759	0.073	28.269	0	0.22
3	HQ2_24_30	.	0.788	0.03	12.859	0.001	0.164
4	HQ3_16_24	.	0.805	0.017	7.923	0.006	0.153
5	HQ3_30_100	.	0.82	0.015	7.259	0.008	-0.132

Table 3.5: Stepwise regression considering global averages of the time series & water content frequency tables for the upper layer grouped into 5 classes and 4 quarters. The frequencies are grouped as follows: 0 - 6 %; 6 - 16%; 16 -24 %; 24 - 30 %; 30 - 100%.

#	Entered	Removed	R**2	Partial R**2	F	Sig. Var. F	SRC
1	AVWFSURF	.	0.685	0.685	204.672	0	0.625
2	AVTRANSP	.	0.759	0.073	28.269	0	0.237
3	HQ2_24_100	.	0.788	0.03	12.839	0.001	0.168
4	HQ3_16_24	.	0.805	0.017	7.89	0.006	0.184
5	HQ3_24_100	.	0.817	0.012	5.965	0.017	-0.136

Table 3.6: Stepwise regression considering global averages of the time series & water content frequency tables for the upper layer grouped into 4 classes and 4 quarters. The frequencies are grouped as follows: 0 - 6 %; 6 - 16%; 16 -24 %; 24 - 100 %.

3.4.2.2 Is mineralisation the key process for interpretation of data?

To verify the hypothesis that the propagation of the 96 hydrological time series through the SOILN models is mainly driven by the soil moisture response curve for mineralisation, WC frequency tables have been re-grouped into five classes, corresponding to 5 equally large ranges for the mineralisation response, Table 3.7.

Name	Classes of mineralisation response	Water content ranges for each class
Res20	0-0.2	0-8%; 66-100%
Res40	0.2-0.4	8-10%; 56-66%
Res60	0.4-0.6	10-12%; 44-56%
Res80	0.6-0.8	12-14%; 34-44%
Res100	0.8-1	14-34%

Table 3.7: grouping of water content frequency tables used for regression analyses shown Table 3.8 and Table 3.9.

Results of the new stepwise regression analysis are shown in Table 3.8-9, for monthly data and quarterly data (Q1=GEN-MAR and so on) respectively.

#	Entered	Removed	R^2	Partial R^2	F	Sig. Var. F	SRC
1	AVWFSURF	.	0.686761	0.686761	206.0901	1.34E-18	0.555831
2	AVTRANSP	.	0.760645	0.073885	28.70751	6.08E-07	0.349677
3	JUN_res40	.	0.776002	0.015357	6.307274	0.013768	-0.16556
4	NOV_res40	.	0.790076	0.014074	6.100991	0.015375	-0.12624
5	JUN_res60	.	0.801614	0.011538	5.234111	0.024491	0.110048
6	SEP_res100	.	0.811228	0.009615	4.533042	0.036008	0.12342
7	MAY_res60	.	0.822114	0.010886	5.385212	0.022621	0.111793

Table 3.8: stepwise regression analysis by considering the grouping of Table 3.7 for monthly data of water content frequency tables.

#	Entered	Removed	R^2	Partial R^2	F	Sig. Var. F	SRC
1	AVWFSURF	.	0.686761	0.686761	206.0901	1.34E-18	0.604696
2	AVTRANSP	.	0.760645	0.073885	28.70751	6.08E-07	0.273992
3	HQ3_res100	.	0.773921	0.013275	5.402263	0.022312	0.144782

Table 3.9 stepwise regression analysis by considering the grouping of Table 3.7 for quarterly data of water content frequency tables.

No improvement of R-square is detected with this new grouping. An additional limitation of the present analysis is that only one frequency value is entered for the quarterly data, while in Table 3.5 three values were entered. On the other hand it is interesting to note that the signs of the SRC's have a clear interpretation in the light of the mineralisation response curve: frequencies corresponding to a response smaller than 40% are negatively correlated to the nitrogen leaching, and *vice-versa*. So, the results in terms of SRC are completely sound: this means that the response curve for mineralisation is a correct key for the interpretation of results, but it is not comprehensive (R-square decreased).

In this context, it is interesting to compare the placement in the frequency table and in the mineralisation response curve of the most important ungrouped factors detected with the stepwise regression showed in Table 3.4. The placements are shown in Table 3.10.

Almost all the most influent parameters fall inside the 0.8-1 class (WC=14%-34%): it is clear that by grouping all the parameters falling in that class, a lot of information is lost. Again, we can see that taking the mineralisation response curve for the interpretation of the SRC's values, soundness of results is often verified. However the complex correlation structure of the sample hints a clear comprehension of results.

Factor	Response curve class	SRC	Soundness
HMAY_27	0.8-1	0.210063	Y
HAUG_17	1	0.127839	Y
HSEP_31	0.8-1	-0.13342	N
HNOV_21	1	0.117627	Y
HAUG_31	0.8-1	-0.06957	N
HSEP_21	1	0.107525	Y
HMAR_35	0.6-0.8	0.096387	??
HMAY_29	0.8-1	0.105322	Y
HSEP_3	0	-0.0939	Y
HJUN_27	0.8-1	0.081351	Y
HOCT_19	1	0.072208	Y

Table 3.10: Interpretation of SRCs in the light of mineralisation response curve.

To further understand the input/output behaviour of SOIL/SOILN models, the distributions of the WC frequencies through the different replications have been more deeply analysed. In particular, the analysis of the different WC frequencies as they are (and not through some synthetic representation) has been considered. The variability of the frequency tables is not homogeneous either across months or across the different frequency values within each month (see Figure 3.4-6). This is obvious, considering that the sampling procedure has to follow the characteristics of the chosen climate and fulfil the conservation laws. More precisely, variability across months could be grouped into wet months (JAN-APR, SEP-DEC; with September having a higher variability with respect to the others), dry months (JUN-AUG); May is left out because it seems having properties different from all other months. In wet months, most of the values are placed in the class 0.8-1 for the mineralisation response. In dry months, the pdf has the maximum corresponding to about 0.5 of the mineralisation response. In May, the distribution is more uniform across the different classes of the mineralisation response curve.

In the light of these last observations, it seems evident that the SA reflects also some peculiar properties of the sample, other than the occurrence of a particular process. For example:

- HMAY_27, HMAY_29 are probably important because they are the extreme-high values of the distribution, whose variation changes the queue of the distribution;

- AUG_17 is a value mid of the distribution, which varies in wide range;
- SEP_31 is the maximum of the distribution;
- SEP_3 is a extreme-low value of the distribution;
- other frequencies are not important may be just because they do no change across the replications or because they are not sampled at all.

These considerations may suggest that a good strategy for grouping the data should also reflect the actual pdf shape of the sampled WC frequencies. For example, if the monthly data of the present analysis are grouped into 3 four-monthly groups, based on the similarity of the WC distributions: GEN-APR (Wet1), MAY-AUG (Dry), SEP-DEC (Wet2), result of Table 3.11 are obtained. The R-square is now larger with respect to Table 3.9, even if time resolution is worse, implying that the new grouping is more effective.

#	Entered	Removed	R^2	Partial R^2	F	Sig. Var. F	SRC
1	AVWFSURF	.	0.686761	0.686761	206.0901	1.34E-18	0.594621
2	AVTRANSP	.	0.760645	0.073885	28.70751	6.08E-07	0.287438
3	Dry_res40	.	0.770937	0.010291	4.133377	0.044927	-0.15973
4	Dry_res60	.	0.784494	0.013557	5.724617	0.018784	0.141098
5	Wet2_res100	.	0.798593	0.014099	6.300373	0.013858	0.127548

Table 3.11: Stepwise regression analysis grouping water content frequency tables into 3 classes of four-monthly data.

3.4.2.3 Principal component transformation of the water content frequency tables

A final study of the water content frequency tables has been performed applying a principal component analysis to the water content frequency tables. This analysis allowed a better representation of the input-output relationship of SOIL/SOILN models and, due to the elimination of the main correlation structure of the input factors, also a clearer interpretation of the results.

Principal component transformation has been done separately for each monthly data frequency tables. Furthermore, monthly principal components of water content are rotated to minimise the number of variables mainly affecting each component. This procedure has the advantage of allowing an easier interpretation of results, by always zeroing the correlation within each month. It has the drawback that non-null correlation remains between the principal components of different months. However, the residual correlation structure is much weaker: correlation coefficients rarely exceed 0.25 and never get over 0.4. On the contrary, in the unmodified input data set correlation coefficients larger than 0.9 were present.

Only components having eigenvalues larger than 1 have been considered. From the initial tables of 50 equally large frequency classes, the following reduced set of principal variables is obtained (72 elements), which is slightly larger than the set with 5-classes grouping (60 elements):

Month	**# of components**
Jan	7
Feb	7
Mar	7
Apr	6
May	5
Jun	6
Jul	5
Aug	6
Sep	6
Oct	5
Nov	7
Dec	5
Tot	72

Table 3.12: Reduction of the # of input factors by applying PCA in the water content frequency tables (original # of frequencies was 600).

Stepwise regression analysis has subsequently been performed for the principal components. Results are shown in Table 3.13.

#	Entered	Rem.	R^2	Part. R^2	F	Sig. Var. F	SRC	Sound
1	AVWFSURF	.	0.685274	0.685274	204.6722	1.34E-18	0.595938	Y
2	MAY_PC2	.	0.795512	0.110238	50.13552	2.29E-10	0.269235	Y
3	AVTRANSP	.	0.83673	0.041218	23.22568	5.66E-06	0.273392	Y
4	FEB_PC1	.	0.848516	0.011786	7.08011	0.009213	0.110774	N
5	SEP_PC2	.	0.860275	0.01176	7.57461	0.007159	-0.12706	?
6	MAY_PC1	.	0.871773	0.011498	7.980318	0.005837	-0.10602	N
7	JUN_PC6	.	0.880166	0.008393	6.163673	0.014937	0.09079	Y
8	JUN_PC4	.	0.887374	0.007208	5.567977	0.02053	0.085782	N
9	OCT_PC3	.	0.892709	0.005335	4.27594	0.04166	-0.07708	Y

Table 3.13: Stepwise regression analysis using principal components of water content frequency tables (the monthly data).

The R-square is much better than in any previous analysis with grouped variables, and the % of output variance explained is only slightly smaller than with the ungrouped frequency tables. The loadings of the principal components selected with the stepwise regression are visualised in Figure 3.9-12:

- MAY_PC2 and SEP_PC2 correspond to the high values of the distribution, placed in 0.8-1 class of the response curve (at the right-hand side of the optimal response); in SEP_PC2 there is also a negative correlation to the maxima of the distribution (optimal mineralisation response);
- FEB_PC1 represents the low queues of the distribution (mineralisation response <0.5), which are negatively correlated to the maxima (optimal mineralisation response);
- MAY_PC1 represents maxima of the distribution (optimal mineralisation response), negatively correlated to the low-ascending part of the distribution (mineralisation response < 0.5);
- JUN_PC4 and JUN_PC6 represent the low (response <0.2) and high (response 0.8-1) queues of the distributions respectively;
- OCT_PC3 represents the low values of the distribution (response <0.2).

As far as the residual correlation structure of the transformed/reduced hydrological data set, the only significant correlation between the most important input factor is detected for the following variables:

	AVWFSURF	AVTRANSP	MAY_PC2
AVWFSURF	1	0.533296	0.358428
AVTRANSP	0.533296	1	0.367289
MAY_PC2	0.358428	0.367289	1

The other correlation coefficients are smaller than 0.2. Hence, the transformation of input data to principal components allows a more efficient representation of the input-output relationship of the SOIL/SOILN models when meteorological time series are replicated. Such a representation is more easily interpretable, since the variables can be treated as almost independent.

If we try to explain the results in terms of the mineralisation response curve, not all the SRC's can be explained. For FEB_PC1, MAY_PC1 and JUN_PC4 results are the opposite than expected. On the other hand, for SEP_PC2, the criterion remains ambiguous: in SEP_PC2 frequencies with response in the range 0.8-1 are negatively correlated to frequencies with optimal response. The negative sign of the SRC can be interpreted by the prevailing effect of the optimal response frequencies, but nothing could be said *a priori*.

These results confirm again that the mineralisation response curve provides a key, which is quite good for the qualitative interpretation of the effect of fluctuations in the water content on the N-loss. On the other hand, PCA allows to better highlight the limitation of such a criterion. These limitations can be important, if the detailed effect of meteorological fluctuations on the N-loss is searched.

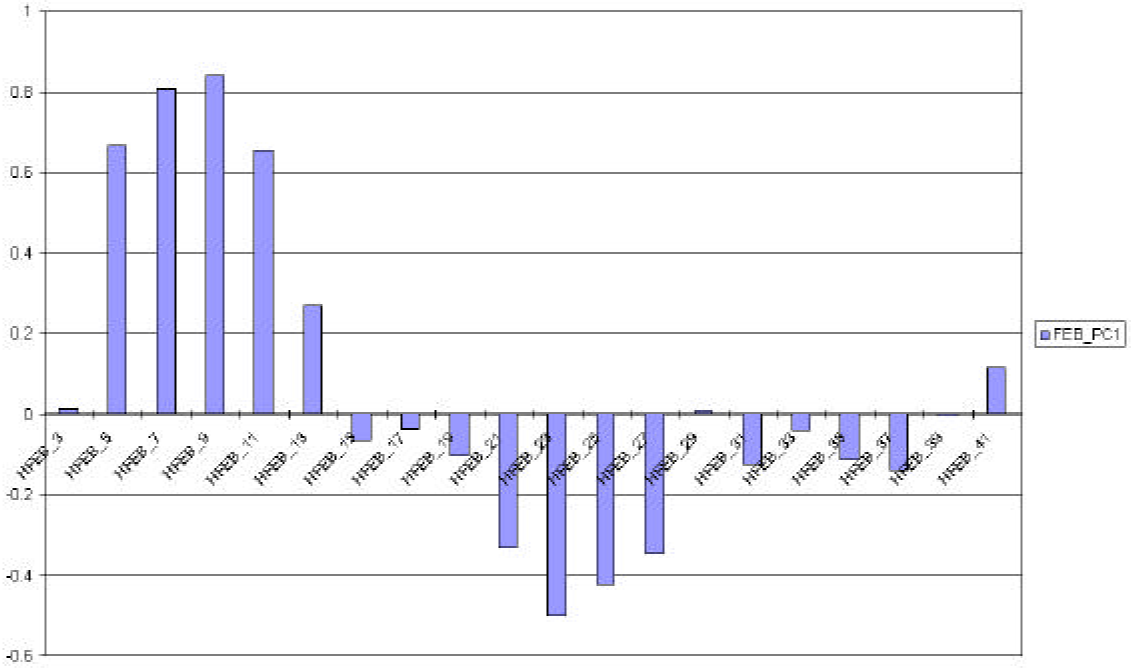

Figure 3.9: loadings for principal component FEB_PC1.

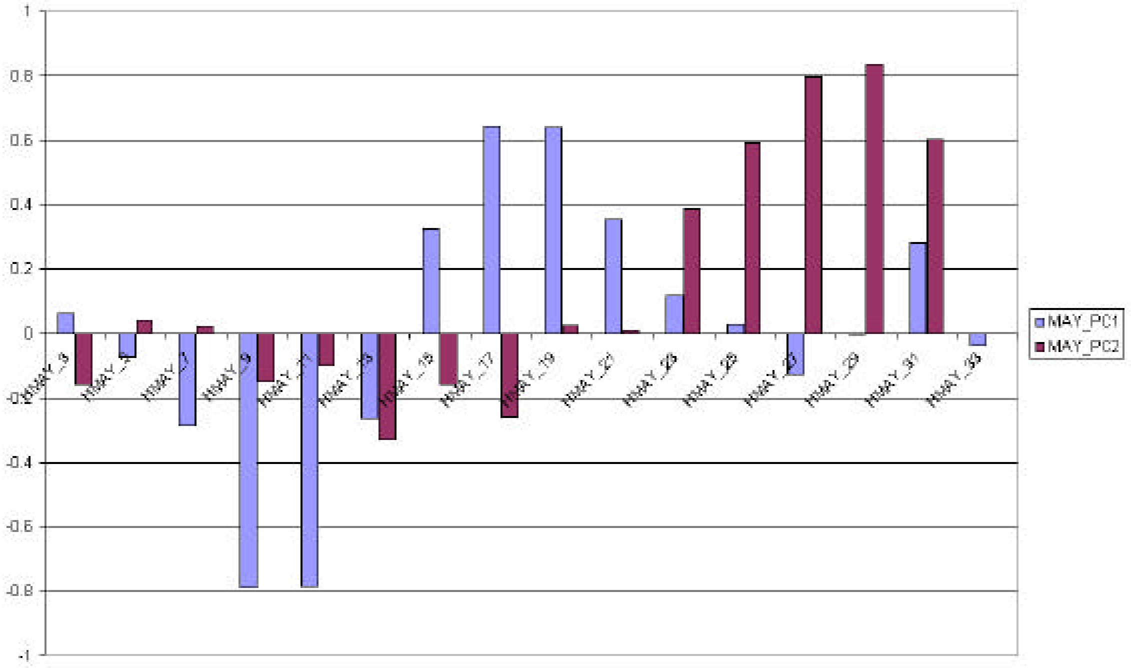

Figure 3.10: loadings for principal components MAY_PC1, MAY_PC2.

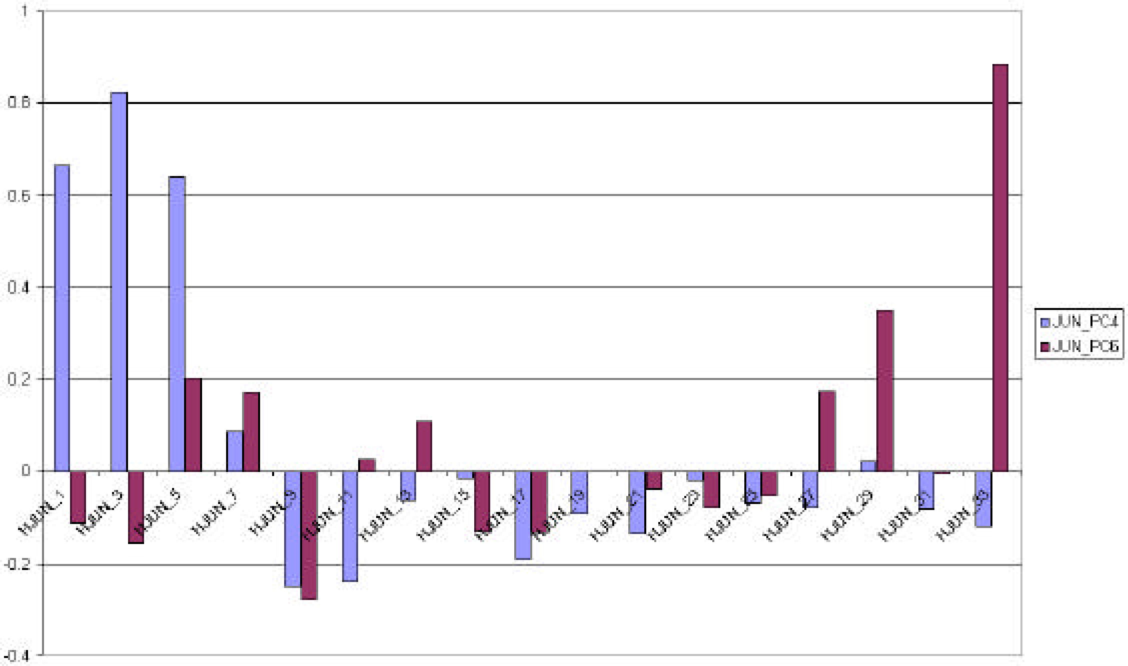

Figure 3.11: loadings for principal components JUN_PC4, JUN_PC6.

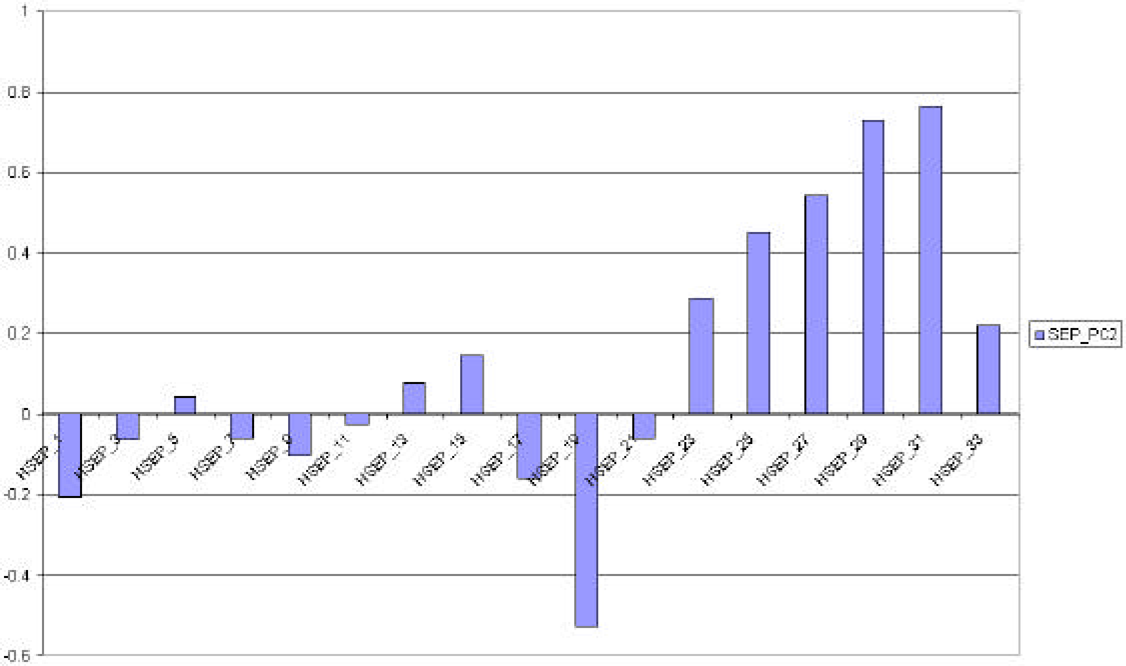

Figure 3.12: loadings for principal component SEP_PC2.

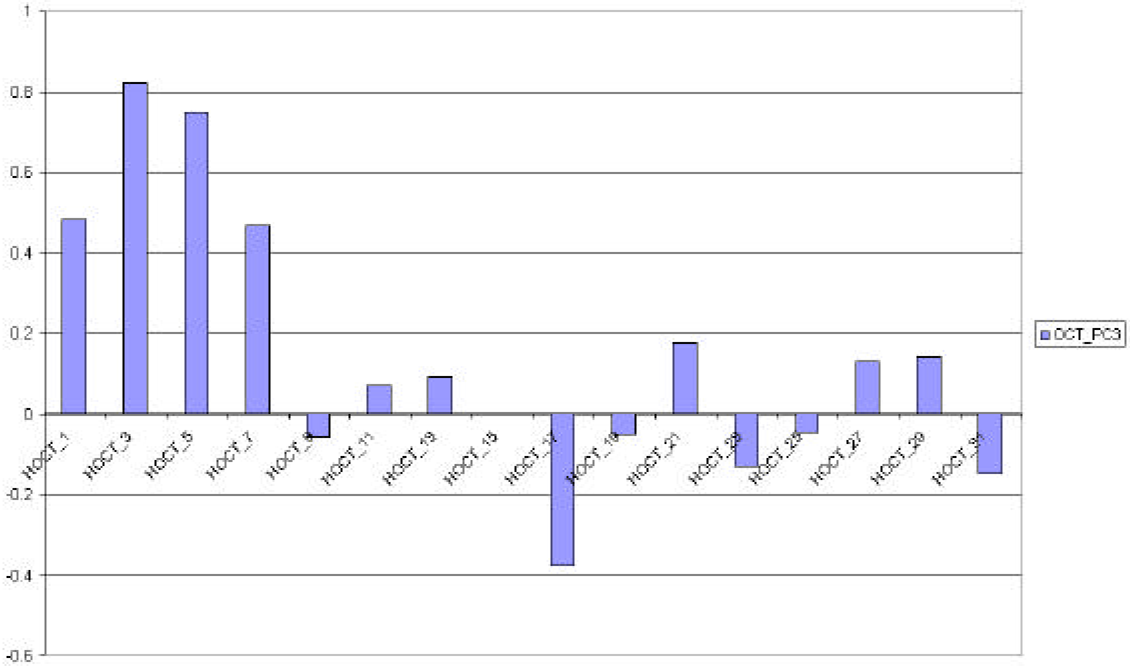

Figure 3.13: loadings for principal component OCT_PC3.

3.4.2.4 Principal component transformation of the hydrological time series characterisations

The last step has been the performance of the PCA for the global statistical characterisations of the hydrological time series (means, variances, max, min). The total number of principal components extracted was 13 out of 80 original factors. This analysis was aimed at verifying if there exists an optimal combination of global statistics, which better explains the output variance. The results of the stepwise regression are shown in Table 3.14.

#	Entered	Removed	R^2	Partial R^2	F	Sig. Var. F	SRC
1	STAT_PC2	.	0.5573	0.5573	118.3333	1.34E-18	0.746525
2	STAT_PC10	.	0.636392	0.079093	20.22954	1.98E-05	0.281234
3	STAT_PC1	.	0.685696	0.049304	14.43181	0.000261	0.222045
4	STAT_PC3	.	0.702725	0.017028	5.212602	0.024752	0.130493
5	STAT_PC11	.	0.715448	0.012723	4.024171	0.047855	0.112797

Table 3.14: Stepwise regression analysis for the principal components of the hydrological time series characterisation.

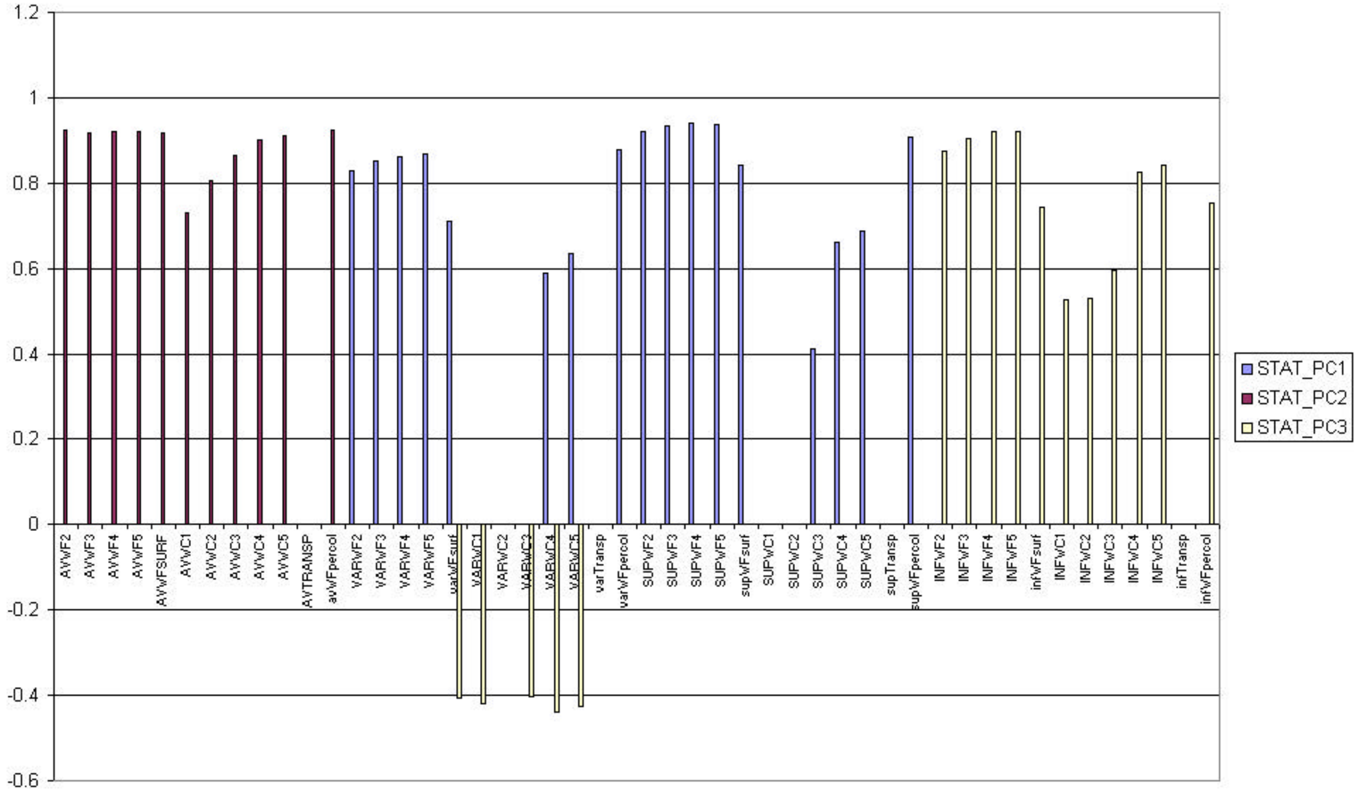

Figure 3.14: loadings for components STAT_PC1/2/3.

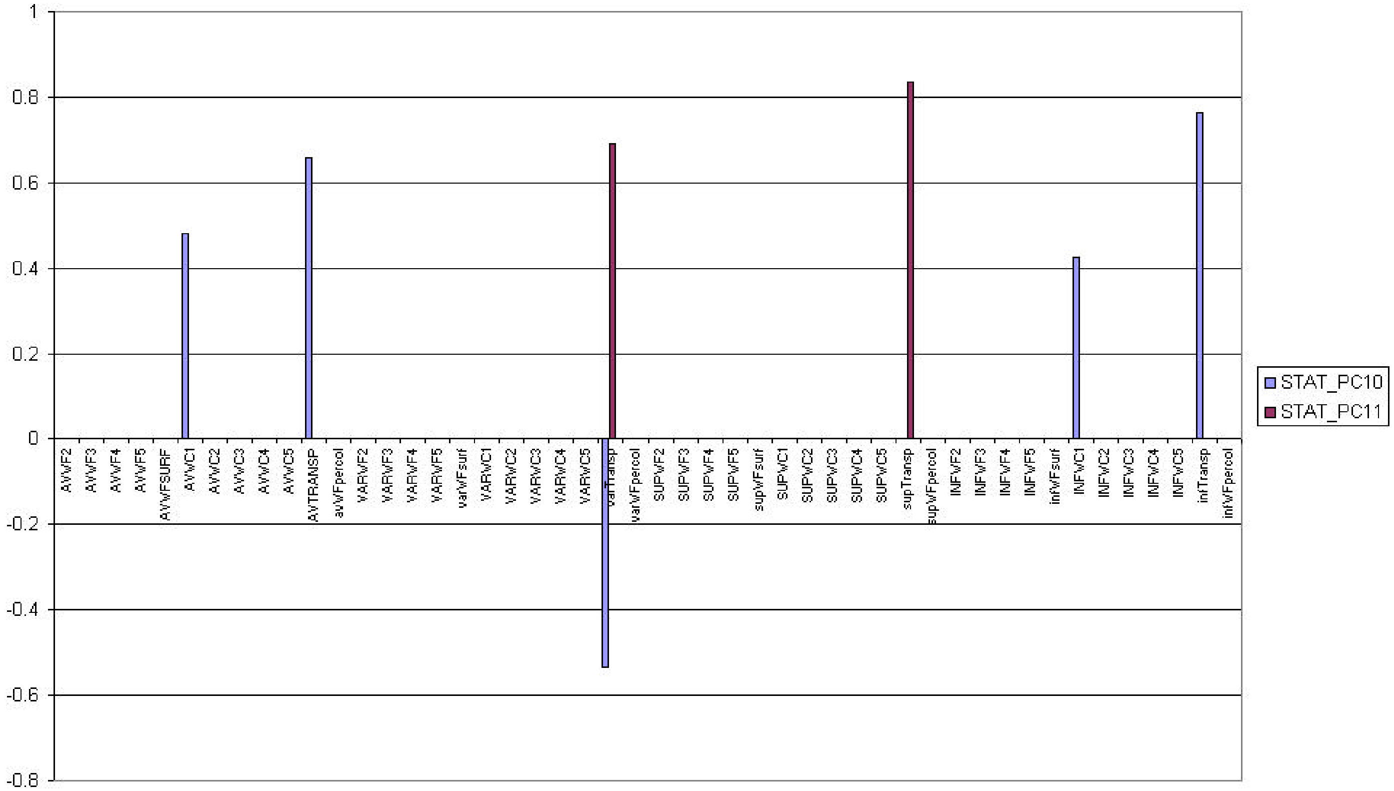

Figure 3.15: loadings for the principal components STAT_PC10 and STAT_PC11.

The loadings for the 5 most important components are shown in Figure 3.14-14, where only loadings larger than 0.4 are shown. The following physical interpretation can be drawn:

- STAT_PC2 represents the average water flows and water contents across the 5 soil layers;
- STAT_PC1 mainly represents the variance of water flows and water contents and the maxima of WF and WC across the 5 layers;
- STAT_PC3 mainly represents the minima of WF and WC across the 5 layers;
- STAT_PC10 and STAT_PC11 are a combination of all the statistics of the ratio between actual and potential transpiration, correlated also with AVWC1 and INFWC1.

All the SRC's are positive: by looking at Figure 3.14-14 it is worth noting that an increase of the inter-annual variance provides a smaller total N-loss.

This last PCA confirms that the best strategy for the reduction of the hydrological time series is to choose a global average of water flow or of water content plus the average ratio between actual and potential transpiration.

3.4.3 Conclusions for the reduction of SOIL/SOILN models

From the regression analysis of N-loss to the characterisations of the hydrological time series, the following conclusions can be drawn.

- Most of the variation in the average nitrogen loss can be explained with 2 global averages: these can be average water flow from the deepest layer and water content from the upper layer or average water flow at the surface and average ratio actual/potential transpiration.
- A high resolution with respect to the actual water content values and the time steps is necessary to reach values of R-square near to 1. However reducing the number of frequencies by grouping histograms maintains a satisfactory R-square value (> 0.8). So, also considering the relatively small sample dimension of the SA, the hydrologic time series can be satisfactorily represented by a very small set of parameters.

- Information about soil water content from the upper layer is sufficient; water content values from deeper layers do not add any important information.
- The reduced models based on the regression analyses can be physically interpreted in the light of the mineralisation process. This corroborates the possibility of taking the regression models as 'sound' models. In particular, it seems that at least the information for model reduction obtained in terms of global averages can have general validity.
- Taking the mineralisation response curve for reference for the grouping of water content frequency tables provides a key for an easy interpretation of results. However this key is only qualitative, since results also show that the response curve is not the unique key-process for the interpretation of results.
- Principal component analysis is the most effective strategy to group input factors in such a way that the largest part of the output variance is explained.
- The more detailed input-output representation obtained by considering the WC frequency tables (PCA or unmodified tables) is strongly connected to the particular/local characteristics of the distributions of the meteorological time series (in connection to the particular climate considered) and are not comprehensively explained through the response of the mineralisation process. So, the grouping of WC frequency tables following a physical criterion, even if allows a sound interpretation of results, provides a loss of information of such local properties.
- The comprehensive interpretation of PCA results would require the inclusion of other processes in addition to mineralisation.

3.5 SA: Effect of time resolution of input data

3.5.1 Methodological approach

To verify the necessity of considering daily data as the input to predict to total average N-loss, SOILN model runs have been made by averaging the daily data obtained with the SOIL model at increasing time periods. Specifically, 2, 4, 8, 16, 32 days averaging was considered.

The SA has been performed considering 4 input factors:

- the first three parameters identified in Table 3.6 to characterise the climatic-hydrological scenario;
- an additional parameter (Resol) representing time resolution, having the values:

  Resol=1, 1/2, 1/4, 1/8, 1/16, 1/32.

3.5.2 Results

The runs with averaged input data introduce a systematic error in excess for the prediction of the nitrogen loss (both for the average and for the standard deviation). The statistical properties of various N loss distributions are shown in Table 3.15.

	Min	Max	Average	St. dev.	Kurtosis
Daily data	1.02E-02	1.35E-02	1.14E-02	5.23E-04	1.652
AVE_2	1.02E-02	1.36E-02	1.15E-02	5.26E-04	2.008
AVE_4	1.02E-02	1.36E-02	1.15E-02	5.31E-04	1.637
AVE_8	1.04E-02	1.37E-02	1.16E-02	5.28E-04	1.66
AVE_16	1.05E-02	1.38E-02	1.17E-02	5.42E-04	1.301
AVE_32	1.08E-02	1.41E-02	1.20E-02	5.49E-04	1.297

Table 3.15: statistical properties of the predicted nitrogen loss.

In Figure 3.16 the standardised regression coefficients for the stepwise procedure applied to the 4 input factors considered are shown for the mixed input/output data obtained by considering the combinations of daily data runs with runs of increasing averaging. The SA has been performed by considering separately the runs with different averaging, in order to appreciate the increasing importance of averaging in modifying predictions.

The SRC for Resol is negative: in fact, as the time resolution increases, the systematic error in excess tends to vanish, implying smaller values for the N-loss. Time resolution begins to be significant at the 4 days averaging and is prevailing starting from 8 days averaging. On the other hand, if the percentage change in the mean and standard deviation is analysed (Figure 3.17), we can see that it does not exceed 5% even for the 32 days averaging. Moreover, the deviation of data is in excess (i.e. conservative).

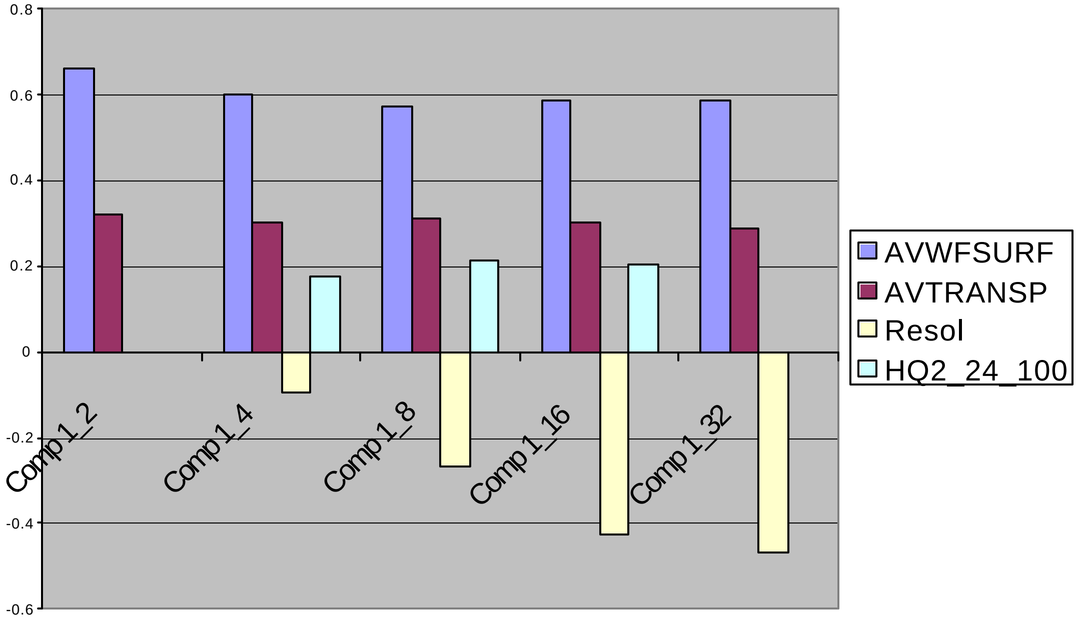

Figure 3.16. Standardised correlation coefficients for the runs of daily data combined with: 2 days averaging (Comp 1_2);4 days averaging (Comp 1_4); 8 days averaging (Comp 1_8); 16 days averaging (Comp 1_16); 32 days averaging (Comp 1_32).

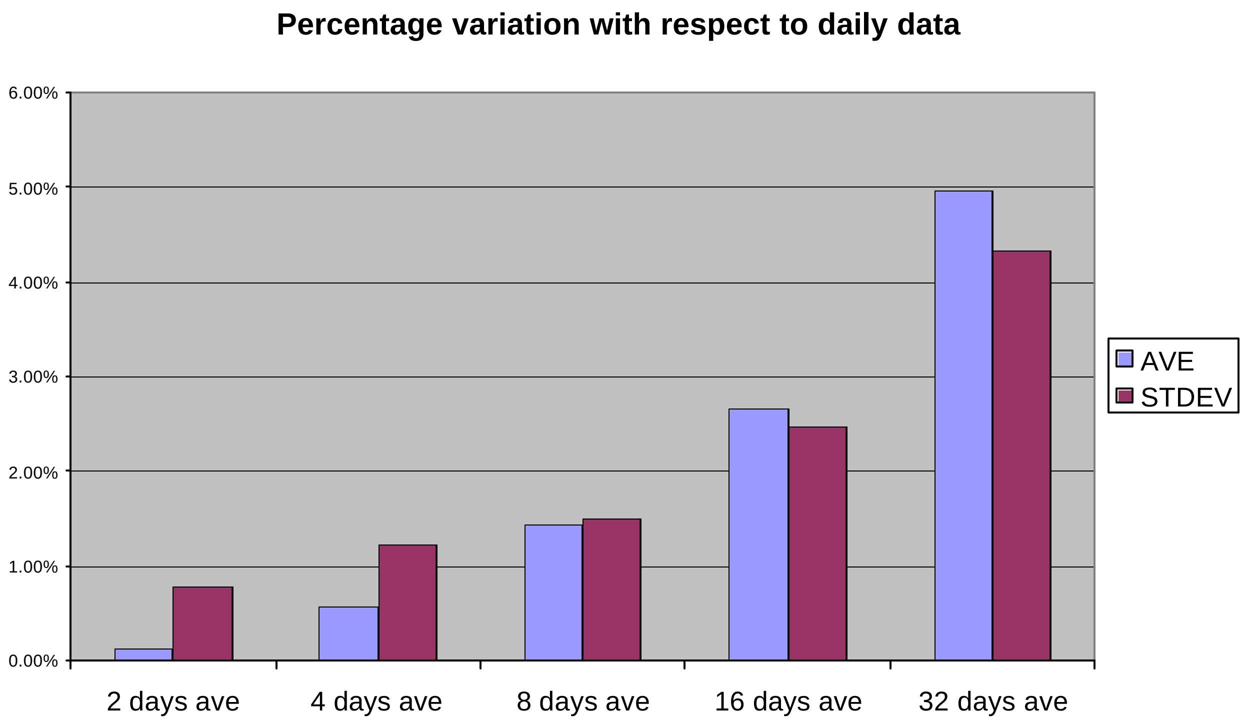

Figure 3.17. Variation of the average and the standard deviation of the N loss by running SOILN model with averaged input data.

3.5.3 Conclusions

Considering the SOILN outputs obtained by averaging hydrological data, the prediction of the N loss is affected by a systematic error in excess which is significant already for the 2 days averaging: this means that accurate predictions require high time resolution in the input data.

On the other hand, the relative error of the N loss prediction with averaged input data is bounded at the 5% even for the 32 days averaging. This may be acceptable remembering that the prediction is conservative and that the standard deviation of the prediction itself is of the order 5%, too.

4 Concluding remarks

In the present report, a methodological approach is presented for model reduction, based on the performance of a sensitivity analysis of model output to input factors. The aim of the methodological study is to identify criteria to reduce models, avoiding unneeded complexity, used for the uncertainty prediction of environmental systems. The model reduction technique is based on the results of the SA. SA allows apportioning the model output variation (i.e. uncertainty) to the different sources of uncertainty in the inputs. So, factors providing negligible contributions to the model output are clearly identified and model reduction criteria are easily defined accordingly.

SA applied to a case study of IMPACT, SOIL/SOILN model, was able to assess the hydrological data requirements for the prediction of the N-loss from the soil averaged over a period of 30 years. The present case study allowed also studying the effect of different levels of resolution in the input data.

The tools of UA and SA appear adequate to help the analyst for the reduction of mechanistic models, to facilitate the interpretation of temporal changes in the state of the environment. The methodological approaches identified will be applied in the further activities of the IMPACT project, where mechanistic models are involved.

5 References

Draper, N. R., and Smith H., (1981). Applied Regression Analysis. John Wiley & Sons, New York.

Eckersten, H., Jansson, P-E. & Johnsson, H. (1994) SOILN model, ver. 8, User's manual, 2:nd edition. Division of Hydrotechnics, Communications 94:4, Department of Soil Sciences, Swedish Agricultural University, Uppsala, ISRN SLU-HY-AVDM--94/4--SE. 58 pp.

Freer, J., Beven, K., and Ambroise, B., 1996, Bayesian estimation of uncertainty in run-off predictions and the value of data: an application of GLUE approach, Water Resources Research, 32(7), 2161-2173.

Hamby D. M. (1994) A Review of Techniques for Parameter Sensitivity Analysis of Environmental Models, *Environmental Monitoring and Assessment*, **32**, 125-154

Helton, J. C. (1993) Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal, *Reliability Engineering and System Safety*, **42**, 327-367.

Helton J. C., and D. E. Burmaster, (1996), On the treatment of aleatory and epistemic uncertainty in performance assessment for complex systems, special issue: Reliability Engineering and System Safety, 54, 91-94.

Jansson, P-E. & Halldin S., 1979. Model for annual water and energy flow in layered soil. In: Halldin (ed.) Comparison of forest water and energy exchange models. Int. Soc. Ecol. Modelling (Copenhagen) pp.145-163.

Ratto M., S. Tarantola, A. Saltelli, Sensitivity analysis in model calibration: GSA-GLUE approach, submitted to *Computer Physics Communications,* 2000a.

Ratto M., Tarantola S., Saltelli A., U. Callies, Model reduction techniques for time series normalisation, IMPACT Project, Deliverable 17, 2000b.

Saltelli, A., K. Chan, M. Scott, Editors, (2000), Sensitivity analysis, John Wiley & Sons publishers, Probability and Statistics series.