ELSEVIER

Computer Physics Communications 117 (1999) 75-85

Computer Physics Communications

Tackling quantitatively large dimensionality problems

F. Campolongo [1], S. Tarantola, A. Saltelli

Institute for Systems, Informatics and Safety, Joint Research Centre of the European Commission, TP 361, 21020 Ispra (VA), Italy

Received 30 April 1998

Abstract

A two-step approach to sensitivity analysis of model output in large computational models is proposed. A preliminary screening exercise is suggested in order to identify the subset of the most potentially explanatory factors. Afterwards, a quantitative method is recommended on the subset of preselected inputs. The advantage of the proposed procedure is that, very often, among a large number of input factors, only a few have a significant effect on the model output. The approach provides quantitative sensitivity measures while controlling the computational cost of the experiment. The procedure has been tested on a recent version of a chemical kinetics model of the tropospheric oxidation pathways of dimethylsulphide, including 68 uncertain factors. © 1999 Elsevier Science B.V.

PACS: 89.60.+x; 02.30.Nw; 02.70.-c
Keywords: Sensitivity analysis; Multi-parameter models; Screening; Quantitative sensitivity measure

1. Introduction

Sensitivity Analysis (SA) is helpful in building models, useful in calibrating them, and essential in the use of models to sustain or disprove hypotheses. This is because our knowledge about the world - what constitutes the input to models - is affected by uncertainties, and the output of a model is likewise uncertain. SA allows the latter uncertainty to be labelled according to source, thus offering one element of reckoning about the consistency between the model internal entailment structure and the world that it tries to emulate. Referring to the elegant formalism of Rosen [1], SA may be considered as useful in the craftsmanship of coding/decoding associated with the process of mimicking reality by models.

Several SA methods are available in the literature and the choice of which SA method to adopt, is a difficult step. Such a choice depends on the problem that the investigator is trying to address, on the characteristics of the model under study, and also on the computational cost that the investigator can afford.

As a rule [2] when the model is nonlinear and various input variables are affected by uncertainties of different orders of magnitude, a global sensitivity method should be used. By global we mean a SA experiment covering the entire space of existence of the input factors, defined in contrast with local, where the input parameters are given a small interval of fractional variation around a nominal value. The use of a local SA implies the assumption (rarely satisfied) that the input–output relationship is linear. If the model is indeed nonlinear, the linear sensitivity approach is unable to assess effectively the impact of possible differences in the scale of variation of the input variables.

[1] E-mail:francesca.campolongo@jrc.it

PII S0010-4655(98)00165-9

In dealing with models that are computationally expensive to evaluate and have a large number of input parameters, the choice of the SA method is restricted to those methods which are computationally cheap, i.e. which require a relatively small number of model evaluations. In general, as a drawback, those "economic" methods can only provide sensitivity measures which are qualitative, i.e. measures capable of ranking the input factors in order of importance, but not of quantifying how much a given factor is more important than another. A quantitative method instead would give, for example, the exact percentage of the total output variance that each factor (or group of factors) is accounting for. Thus, there is a trade-off between computational cost and information.

Following this idea, in Fig. 1 we have represented some well-known classes of methods according to the two properties: "information", i.e. the amount of information produced in terms of the model sensitivity (on the abscissa) and "cost", i.e. the computational cost (on the ordinate). The computational cost is measured in number of model evaluations and is a function of the number of input factors examined (k), and of the complexity of the model.

We have put closest to the origin the class of the "elementary OAT" methods. By "elementary OAT" we mean those methods that change one factor at a time (OAT) and explore what the model does with the new datum. In these analyses the baseline value is kept constant, i.e. the factors are moved away from the baseline only once (or twice) and the baseline is not changed throughout the analysis. While this approach is computationally very cheap ($\sim k$ model evaluations), its limitations are evident because the information that it is produced is only local.

A method that, still changing one factor at the time, can be considered as global is the method proposed by Morris [3]. The Morris sensitivity measure is obtained by computing a number r of local measures at different points $\boldsymbol{x}_1, \ldots \boldsymbol{x}_r$ of the input space and then taking their average (so to lose the dependence of the specific point at which the measure was computed). The information on the model sensitivity produced by the Morris method is more general with respect to an elementary OAT in the sense that the method explores the whole input factor space. As a drawback, the computational cost has been increased up to $r \times (k+1)$ model evaluations, where r is usually in the range 5-15 (see Fig. 1).

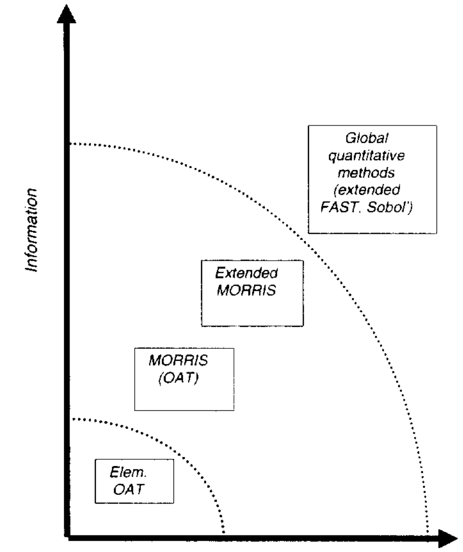

Fig. 1. Representation of various classes of methods for sensitivity analysis according to two properties: "information", i.e. the amount of information produced in terms of the model sensitivity (along the abscissa), and "cost", i.e. the computational cost of the experiment (along the ordinate). The "cost" is measured in terms of the number of model evaluations and is a function of the number of input factors (k) and of the complexity of the model.

The OAT method proposed by Morris – which estimates the main effects of the input factors on the output – has been extended by Campolongo and Braddock [4] to estimate also the two-factor interactions effects. The number of model evaluations required by the extended Morris is $O(k^2)$ (see Fig. 1). Note that both the original Morris and its extended version provide sensitivity measures that are only qualitative.

To obtain a quantitative sensitivity measure, the computational cost has to be further increased. For instance, methods such as the Sobol' indices [5] or the extended FAST [6] that are capable of quantitatively decomposing the total output variance in the percentages that each factor (or combination of factors) is accounting for, require a higher number of model evaluations. The computational cost of these methods is $\sim k \times N$ model evaluations, where N is usually in the range of hundreds to thousands. However, these methods are valuable because they can quantify the importance of each factor. We would recommend their use any time the modeller can afford the computational cost.

Campolongo and Saltelli [7] suggested that, when dealing with models containing a large number of parameters, a possible procedure, which would match sensitivity information achieved with computational cost, would be the one made by two consecutive experiments. A preliminary screening exercise could be conducted with the goal of identifying the parameter subset that controls most of the output variability with low computational effort. Then, the screening could be followed by a quantitative method applied on the subset of preselected inputs. For a successful experiment, both the exercises should be run with global techniques.

This procedure can be effective, especially since often, among a large number of input parameters involved in a model, only a few have a significant effect on the model output. Furthermore, the approach has the big advantage of providing quantitative sensitivity measures while controlling the computational cost of the experiment.

In this paper, the idea is implemented for the first time on a real model. The model under study is KIM (KInetic Model), a chemical kinetics model of the tropospheric oxidation pathways of dimethylsulphide (DMS).

A description of KIM and a brief story of its development are given in Section 2. Section 3 reports the results of a previous qualitative SA conducted on KIM. Section 4 describes the two SA methods adopted within the new analysis, the Morris screening method (Section 4.1), and the quantitative extended FAST (Section 4.2). Results and conclusions are given in Section 5.

2. The KIM model

KIM stands for KInetic Model for OH-initiated oxidation of DMS (CH_3SCH_3), and incorporates a description of the tropospheric reaction pathways for the formation of sulphur-containing molecules, such as sulphur dioxide (SO_2) and methane sulphonic acid (MSA, CH_3SO_3H), from DMS. KIM is 0-dimensional and includes multiphase (droplets-air) transport and chemistry. The KIM model is relevant to climate change studies, because of the important contribution of DMS emissions to the formation of climatically active atmospheric aerosols and in particular the hypothesised feedback mechanism linking the biogenic sulphur cycle to the greenhouse effect [8,9].

The oxidation can be enhanced by the presence of water droplets in the troposphere, and provide an aqueous pathway for the formation of sulphur containing molecules.

The process of building the KIM model involved tackling uncertain mechanisms and reaction rates. Certain mechanistic aspects of the chemistry of DMS were so poorly understood that the knowledge of the various reactions and of their relative weights was very imprecise. Thus, the reaction scheme adopted in the model was affected by large uncertainty (structural uncertainty). Furthermore, given the large uncertainties in the parameter values governing DMS oxidation kinetics, the error bars associated with the rate constants involved were large, and in some instances almost arbitrary (parametric uncertainty).

All the above reasons, together with the scarcity of observed data for a proper model calibration, led to the implementation of a model building process where uncertainty and sensitivity analysis played a central role. The KIM model was run within a Monte Carlo driver, capable of propagating the uncertainty in the input parameters onto the output variables. The MC analysis of Saltelli and Hjorth [10] allowed the quantification of (1) the uncertainty in model prediction, and (2) the relative importance of each input parameter in determining such an uncertainty. The study was based on KIM-I, a purely gas-phase chemistry version without droplets.

One of the main limitations of studies done with KIM-I was that it neglected the temperature effects on the DMS-oxidation process. In 1994, Remedio et al. [11] extended the KIM-I model to include the latitude dependency, producing a second version of the model, KIM-II. The Monte Carlo analysis was then performed again on KIM-II, and a latitudinal analysis, emphasising the possible regional differences on the main oxidation pathways of DMS and on the relative amounts of end-products formed, was carried out. Results of the analysis agreed generally with those found in a recent bibliography by Koga and Tanaka [12], and several conclusions could be drawn. However, those conclusions were still conditional upon the model and data assumptions underlying the experiment. Among these assumptions, the non-inclusion of the heterogeneous chemistry (aqueous phase) and dry deposition, by far the largest sink for SO_2 molecules, was the most severe.

In 1997, a third version of the KIM model (KIM-III) was produced [13]. In KIM-III, the heterogeneous chemistry is dealt with, and a first attempt is made to include some elements of cloud processing in the model. Liquid phase chemistry occurs inside the water droplets in the troposphere, and involves transfer to the droplet, chemical reactions inside the droplet, and the sink terms for the droplet (e.g. washout).

Uncertainties in the heterogeneous oxidation of DMS and of its intermediates in the liquid phase are even more severe than for the homogeneous chemistry mechanism [13].

As discussed by Ayers et al. [14], the concentration ratio in marine aerosol between MSA and non-sea-salt sulphates (nss-$SO_4^=$, including SO_2 and H_2SO_4), i.e.

$$\alpha = \mathrm{MSA}/(\mathrm{SO_2} + \mathrm{H_2SO_4}),$$

seems to offer the best opportunity for comparing observed data to those predicted by models, in particular, for considering the temperature dependencies involved in the various branches of the oxidation processes. Further, the MSA/nss-$SO_4^=$ ratio may be used for estimating the actual contribution of DMS to observed nss-$SO_4^=$ from measurements of MSA, if the dependence of the ratio on temperature and other ambient conditions are sufficiently well known [15,16].

The temperature dependency of α was the focus of the study by Campolongo et al. [13]. The values of α predicted by KIM-III were compared with field observations of MSA to non-sea-salt-sulphate ratios [17].

3. Previous SA on KIM

Campolongo et al. [13] carried out a qualitative SA study to identify the parameters most influential on the ratio α. The input parameters included in the analysis were not only temperature dependencies involved in the gas phase chemistry but also the anticipated temperature dependencies of the interaction between gas phase (homogeneous) and liquid phase (heterogeneous) chemistry. Results indicated that these latter temperature dependencies might, to a large extent, explain the actual observed values of the α ratio. Thus the analysis highlighted the potential role of multiphase atmospheric chemistry not only in the case of SO_2, but also of other oxidation products of DMS and, particularly, of DMS itself.

The KIM-III version of KIM involves 68 uncertain input variables. It follows that with such a large number of potential explanatory variables, any SA measure based on a regression (such as the Standardised Regression Coefficient SRC, the Partial Correlation Coefficient PCC, etc.) cannot be trusted with any high level of confidence [18]. In order to perform the SA on KIM-III by focusing only on a limited number of variables, Campolongo et al. [13] conducted a preliminary analysis and identified the 20 more influential input variables out of the total 68. The preliminary analysis was carried out by computing the SRC's 10 different times (on the base of ten different Monte Carlo simulations) for all the 68 input variables, and then selecting those variables which had been identified at least three times (out of the ten MC analysis executed).

A more rigorous Monte Carlo type sensitivity analysis was then carried out on the 20 preselected variables. The other variables were fixed to a "nominal" value and kept constant in the succeeding simulations. The SRC regression coefficients, from the least-square regression analysis applied to the Monte Carlo simulation, were computed for the 20 variables of interest and used to rank them in order of importance [13].

4. The present SA on KIM

4.1. The screening exercise

One of the factors identified as very important by SRC sensitivity measures was the kinetic parameter k_{21} [13]. Unfortunately, estimates of the k_{21} value presented in the literature show strong discrepancies [10], and this value is still very uncertain. For these reasons, Campolongo et al. [13] felt appropriate to repeat their analysis of the latitude dependency of the ratio α after replacing the value distribution of k_{21}, which originally was taken from Ray et al. [19], by an alternative value, which was reported in Mellouki et al. [20]. The disagreement found comparing the two sets of model outcomes (obtained respectively with the first and second choices of the k_{21} value) confirmed the key role played by this factor and the need to obtain a accurate estimate for its value.

In this work, we investigate the sensitivity of the final version of the KIM model, which is the KIM-III but with the k_{21} value as given in Mellouki et al. [20]. The preliminary screening analysis is conducted here by using the screening method of Morris [3].

This method varies one-factor-at-a-time across a certain number of levels selected in the space of the input factors, Ω. The method requires a total number of model evaluations that is of the order of k, $O(k)$, where k is the number of model input factors. Two sensitivity measures are provided for each input factor: a measure μ of the "main" effect, and a measure σ that is the sum of all the second and higher order effects in which the factor is involved (including curvatures and effects due to interaction with other factors). Note that the measure μ computed by Morris is global. In fact, μ is obtained by computing a number r of local measures – called Elementary Effects – at different points $\boldsymbol{x}_1, \dots \boldsymbol{x}_r$ of the input space, and then taking their average (so to lose the dependence of the specific point at which the measure was computed). The number r of selected points is called the sample size of the experiment [3]. In this work, we adopted a sample size $r = 10$ and a number of levels $l = 4$.

At present, all the examples of application of the Morris method available in the literature are based on the assumption that the distribution of each input factor is uniform [3,4,7]. Such an assumption, although generally weak, is the only acceptable when the knowledge of the input parameters is quite poor. When the assumption of uniform distributions holds, the levels of the experiment are then simply obtained by dividing in equal parts the interval in which each factor varies. For example, if a factor varies in [0, 1] and we want to select a number of 4 levels, these levels are $l_1 = 0$, $l_2 = 1/3$, $l_3 = 2/3$, $l_4 = 1$.

In the present work, more accurate information is available about the input factors of the KIM model. Statistical distribution functions have been selected mostly based on the literature (uniform, log-uniform, normal, log-normal, ...). In this case, a simple choice of levels as the one mentioned above, would result in a loss of information, since it would neglect the statistical information contained in the distribution functions. The procedure we adopted for the KIM model exercise is the following: instead of sampling the input values directly in Ω, we first sampled in the space of the quantiles of the distributions, which is a k-dimensional hyper-cube (each quantile varies in [0, 1]). Then, given a quantile value for a given input factor, the actual value taken by the factor was derived from its known statistical distribution.

Results of the Morris screening exercise are given in Fig. 2. The two Morris sensitivity measures μ and σ are plotted for the 68 input factors (only the 10 most important factors are named). Note that μ and σ provide two complementary measures of sensitivity; however, for this exercise, the two sets of the 10 most important factors as identified by μ and by σ are identical. The ranking of the factors obtained according to μ is given in Table 1.

4.2. FAST on KIM

4.2.1. The FAST method

The Fourier Amplitude Sensitivity Test (FAST) was proposed in the 70's [2,21,22] and was successfully employed in investigating the sensitivity of large sets of coupled reaction systems to uncertainties in rate coefficients.

In a further article [23] the method was reviewed and reinterpreted as to fit into an ANOVA setting. In an ANOVA setting the total output variance D is decomposed into orthogonal terms of increasing dimensionality, e.g. for a model with three factors,

$$D = D_1 + D_2 + D_{12} + D_{13} + D_{23} + D_{123} . \qquad (1)$$

The same decomposition of response into effects is commonly used in experimental design [24].

Here the first order term D_i captures the effect on the output uncertainty due to variations in factor i, while all the other factors are averaged over their range of uncertainty. The second order term D_{ij} is a two-way interaction between factors i and j not including the individual effects due to i and j, which are already taken into account by D_i and D_j. Higher order partial variances express the influence on the output uncertainty due to higher order interactions among factors, and are defined in a similar way. Dividing Eq. (1) by D one obtains

$$1 = S_1 + S_2 + S_3 + S_{12} + S_{13} + S_{23} + S_{123} ,$$

where $S_{i_1,i_2,\dots}$ are the so-called sensitivity indices.

In FAST, the input factors of a model are assumed to be noncorrelated and all of them are varied simultane-

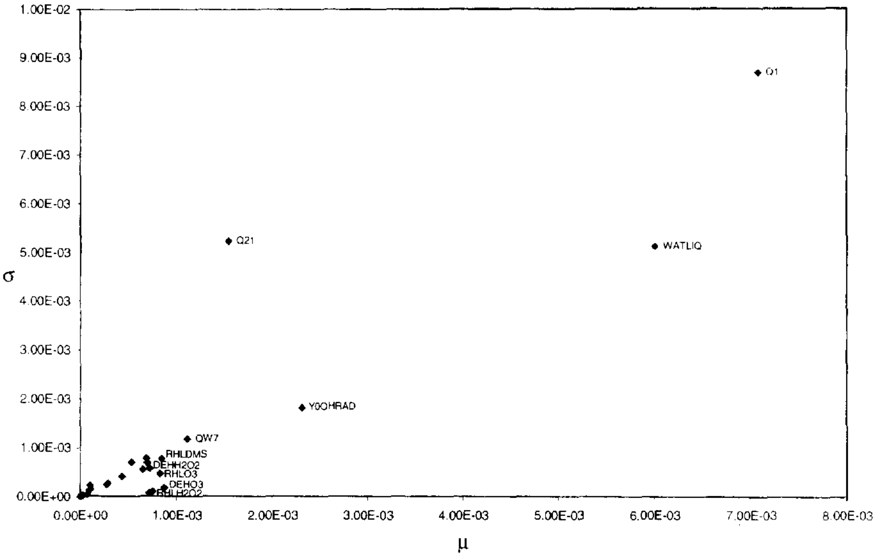

Fig. 2. Results of the Morris screening exercise. The estimated quantities for μ (along the abscissa) and σ (along the ordinate) are illustrated for all the factors. The first 10 most important factors have been labelled.

ously over their ranges of uncertainty, so that a global appreciation of the sensitivities can be achieved.

FAST enables the estimation of the first order partial variances D_i as well as the total output variance D, and hence first order sensitivity indices S_i. In FAST each uncertain input factor x_i is related to a frequency ω_i, and a set of suitably defined parametric equations

$$x_i(s) = G_i(\sin(\omega_i s)) \tag{2}$$

allows each factor to be varied in its range, as the new parameter s is varied. The parametric equations define a curve that systematically explores the input parameter space Ω. The curve is supposed to be space-filling, so that summary statistics on the output can be computed, according to the theorem of Weyl [25], by integrating either over Ω or along the curve itself. The quadratures are employed by using a set of N points which are selected along the curve and are usually equally spaced.

N represents the sample size required for evaluating the whole set S_i. N coincides with the number of model evaluations and, hence, with the total cost of the analysis. The larger the maximum value $\Omega_{\max}$ of the input frequencies Ω_i, the larger the sample size N.

A Fourier analysis is performed on the output $y = f(x_1(s), x_2(s), \ldots) \equiv f(s)$ considered as a function of s: the spectrum $\Lambda^2(\omega)$ of $f(s)$ at each frequency ω is computed by

$$\Lambda^2 = A^2 + B^2 ,$$

where

$$A(\omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(s) \cos \omega s \, \mathrm{d}s$$

and

$$B(\omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(s) \sin \omega s \, \mathrm{d}s$$

are integrals numerically evaluated over s. Finally, the S_i's are obtained by computing the ratio between D_i and D, which are estimated according to

$$D_i = 2 \sum_{p=1}^{+\infty} \Lambda^2(p\omega_i) , \tag{3}$$

$$D = 2 \sum_{j=1}^{+\infty} \Lambda^2(\omega_j) . \tag{4}$$

A set of incommensurate frequencies should be used in Eq. (2) for obtaining a space-filling curve. Actually,

Table 1
Results of the Morris experiment on the KIM model. Factors are ranked in order of importance according to the SA measures μ. The 10 most important factors are displayed in bold font.

Factor	Rank	Factor	Rank
WATLIQ	**2**	W15	13
RAIN	66	**Y0OHRAD**	**3**
B32	24	**Q1**	**1**
B5	36	Q5	33
B6	21	Q6	19
B7	16	Q7	15
R8	18	QM14	27
R9	56	QM15	62
R10	11	QM19	29
R12	28	QM20	49
R13	31	**QW7**	**5**
R14	26	**Q21**	**4**
R15	44	Q22	55
R16	30	QM28	39
R17	34	RHLCO2	50
R18	12	RHLSO2	40
R19	25	RHLNH3	67
R20	38	**RHLO3**	**8**
R23	45	**RHLH2O2**	**9**
R25	51	RHLHNO3	61
R26	35	**RHLDMS**	**7**
R28	60	RHLDMSO	53
R29	23	RHLDMSO2	32
R30	20	RHLMSA	42
R31	14	DEHCO2	41
R34	59	DEHSO2	52
PRATE34	48	**DEHH2O2**	**10**
R35	65	**DEHO3**	**6**
R36	63	DEHNH3	68
R37	46	DEHHNO3	47
W8	58	DEHDMS	17
W10	37	DEHDMSO	54
W11	22	DEHDMSO2	43
W14	57	DEHMSA	64

a space-filling curve is only an idealisation since the frequencies cannot, in practice, truly be incommensurate, due to the finite precision of computers. Rational, or equivalently, integer frequencies are therefore employed. This fact poses a problem of interferences between frequencies: the Fourier coefficients evaluated at the input frequency ω_i and its multiples reflect the sensitivity of the output to the ith factor. If interference occurs at a given frequency value, then the corresponding Fourier coefficient will reflect simultaneously sensitivities to more than one factor, thus rendering the analysis impracticable.

Actually, the set of frequencies is chosen so that they are free of interferences up to order M, where M is a parameter at disposition of the investigator. Eqs. (3) and (4) are therefore cut to the Mth term, usually $M = 4$ or 6. Note that incommensurability would correspond to $M = \infty$. Of course, the higher M the more accurate the estimates. The drawback is that N, the sample size, is strongly constrained by M. In a case with 8 factors, for example, if $M = 4$, the minimum sample size required is $N = 486$; if M is set to 6, the minimum N is much higher ($N = 3492$). N is also constrained by k given that, as the number of factors increases, it is necessary to choose higher $\omega_{\max}$ in order to obtain a set $\{\omega_i\}$ free of interferences up to a given order.

Cukier et al. [21] proposed an empirical formula yielding the frequency sets $\{\omega_i\}$ and the (minimum) sample size N for models up to $k = 50$ factors, assuming $M = 4$.

Another formula, based on the Nyquist criterion, gives N in terms of $\omega_{\max}$ and M,

$$N = 2M\omega_{\max} + 1\,, \qquad (5)$$

As discussed in [26], N grows approximately as the square of the number of factors, for $M = 4$. In a case with 50 factors, for instance, no less than $N = 43606$ simulations are required, thus rendering FAST in some cases computationally infeasible.

Saltelli et al. [6] extended the FAST method to estimate the total effects, S_{T_i}, $\forall i = 1, 2, \ldots, k$. A total effect index S_{T_i} (see [27]) is defined as the sum of the indices $S_{i_1, i_2, \ldots}$, which include the index i. For instance, in a model with three factors the total indices look like

$$S_{T_1} = S_1 + S_{12} + S_{13} + S_{123}\,,$$

$$S_{T_2} = S_2 + S_{21} + S_{23} + S_{213}\,,$$

$$S_{T_3} = S_3 + S_{31} + S_{32} + S_{312}\,.$$

It is worthwhile noticing that, as in experimental design, $S_{123} \equiv S_{213} \equiv S_{312}$, which expresses the sensitivity to the output of the third-order interaction among the factors of the model. Similar symmetries occur for the second-order interactions, i.e. $S_{12} \equiv S_{21}$, $S_{13} \equiv S_{31}$, $S_{23} \equiv S_{32}$. The total index S_{T_i} quantifies the overall effect on the output uncertainty due to the factor i: it includes the single effect S_i as well as the effects due to the interactions with the other factors, at any order.

The pair of indices S_i and S_{T_i} for the factor i can be obtained by choosing a "high" value for the frequency ω_i and a set of "low" values for the other frequencies, $\omega_{(-i)}$, corresponding to the remaining factors. In the extended FAST, therefore, $\omega_i \equiv \omega_{\max}$.

By evaluating the Fourier spectrum in the "low" range, the total index S_{T_i} can be estimated, whereas the first order index S_i is obtained as in the classical FAST (see Eqs. (3) and (4)). To estimate the sensitivity indices for the factor j, a permutation of the frequencies is necessary, because a "high" frequency value must be assigned to the factor of interest. This computation requires a new set of N sample points within Ω. Hence, the total cost of the analysis for computing all the pairs of indices is $k \times N$.

When performing the extended FAST, the problem of interference is easier to manage than in the classical FAST: it consists of the overlap occurring between the "low" and the "high" frequency bounds. The problem of interference can be escaped by choosing ω_i such that

$$\omega_{\max} \equiv \omega_i \geq 2M \max\{\omega_{(-i)}\}\,. \qquad (6)$$

This formula imposes a constraint on the sample size N, i.e. $N = 4M^2 + 1$. This relation is deduced by Eqs. (5) and (6) where $\{\omega_{(-i)}\} = 1$ and $\omega_{\max} = 2M$. The use of such a sample size N guarantees an analysis free of interferences up to order M.

The above constraint is much weaker compared to the classical FAST, as the minimum value required for the sample size is $N = 65$ when M is set to 4. An important remark is that N is not constrained by k in any way.

The parametric representation of the curve exploring Ω, introduced in Eq. (2), has been standardised. With extended FAST we generate standard samples of noncorrelated input factors that are uniformly distributed in the range [0, 1). The proposed parametric equations are

$$x_i = \tfrac{1}{2} + \tfrac{1}{\pi}\arcsin(\sin(\omega_i s))\,.$$

In a model where the probabilty distribution functions (p.d.f.'s) and the ranges of the input factors are generic, a differential form can be solved to transform the standard sample into the required one. The ODE for the factor i is

$$\pi(1 - x_i^2)^{1/2} P_i(G_i)\frac{dG_i(x_i)}{dx_i} = 1\,,$$

where P_i is the p.d.f. of x_i.

The extended FAST is flexible as it allows using different curves systematically exploring the input factor space Ω. Let us call the number of curves employed N_r. The sample size N is given by

$$N = (2M\omega_{\max} + 1)N_r\,.$$

4.2.2. *The extended FAST applied to KIM*

The 10 most important factors resulting from the Morris screening exercise have been selected for further investigation. The selection of these factors has been made on the basis of the results displayed in Fig. 2. The distance that a given point in the figure has to the vertical axis represents a "qualitative" measure of the importance of the corresponding factor. In other words, a "qualitative" ranking is obtainable by means of the set of μ values for the various factors.

In the analysis we decided to consider the set of the 10 most important factors according to the ranking offered by Morris. A quantitative appreciation of their influence on the output variable has been obtained by performing the extended FAST.

The ranges and distributions adopted for the 10 factors are the same as in the screening exercise. The remaining factors have been assumed irrelevant and, therefore, fixed to their nominal values.

The sample size $N = 1026$ has been used for the calculation of each pair of indices (S_i, S_{T_i}), with $M = 4$, $\omega_i = 64$ and $N_r = 2$. The total cost of the analysis is $C = N \times k = 10260$, being $k = 10$. The results are illustrated in Table 2 and in Fig. 3, where two pie charts, for the S_i's and for the S_{T_i}'s, represent the percentage contributions of the 10 factors to the total output variance.

The execution of the classic FAST would have yielded the pie chart (a) only. According to Cukier''s empirical formula, the computational cost of the classic FAST would have been $C = 806$. In this particular exercise, standard FAST would have likely sufficed, given that within the error the model behaves additively (the sum of the S_i's is one on the (a) chart). The (b) pie indicates nonnegligible interactions, though these are likely to be an overestimate. Indeed S_{Q_1} is 0.93, while S_{TQ_1} is 0.97; once all the S_{T_i}'s are renormalised with the sum of the total indices, the

Table 2
The FAST indices computed for the 10-factor model. The remaining 58 uncertain factors were kept fixed to their nominal values.

Factor	First order index (S_i)	Total index (S_{T_i})
WATLIQ	.07595	.13668
Y0OHRAD	.00663	.01726
Q1	.93284	.97196
QW7	.01207	.05923
Q21	.00003	.01391
RHLO3	.00150	.02345
RHLH2O2	.00008	.01372
RHLDMS	.00201	.01898
DEHH2O2	.00007	.01186
DEHO3	.00013	.01752

influence of Q_1 ends up being reduced. The true relevance of interactions is likely in between the values given by the two charts.

The use of extended FAST even for additive cases is motivated by the fact that a priori one cannot anticipate additivity (most versions and cases run with KIM indicated substantial non-additivity [10,11]).

5. Discussion and conclusions

The (b) pie chart in Fig. 3 yields useful quantitative information about the overall relative importance of the 10 factors. It shows that the factor Q_1, which is the quantile of k_1, is by far the most important among the 10 factors, accounting approximately for (76–90)% of the output variance D.

This is not surprising since the kinetic constant k_1 is involved in the reaction between OH and DMS and the reactions competing with formation of MSA are of more relevance in the hydrogen-abstraction pathway than in the alternative pathway. Furthermore, the same result was found by Campolongo et al. [13] when studying the sensitivity of the KIM-III model with the old value of k_{21}.

The second most important factor is WATLIQ, explaining about (7–11)% of D. The other 7 factors included in the analysis account for (4–13)% of the total output variance D.

As mentioned above, the apparent discrepancy between the two pie charts in Fig. 3, concerning the possible role of interactions, is due to the numerical error of the method. In any case it can be said that the interactions, if present, do not play a predominant role.

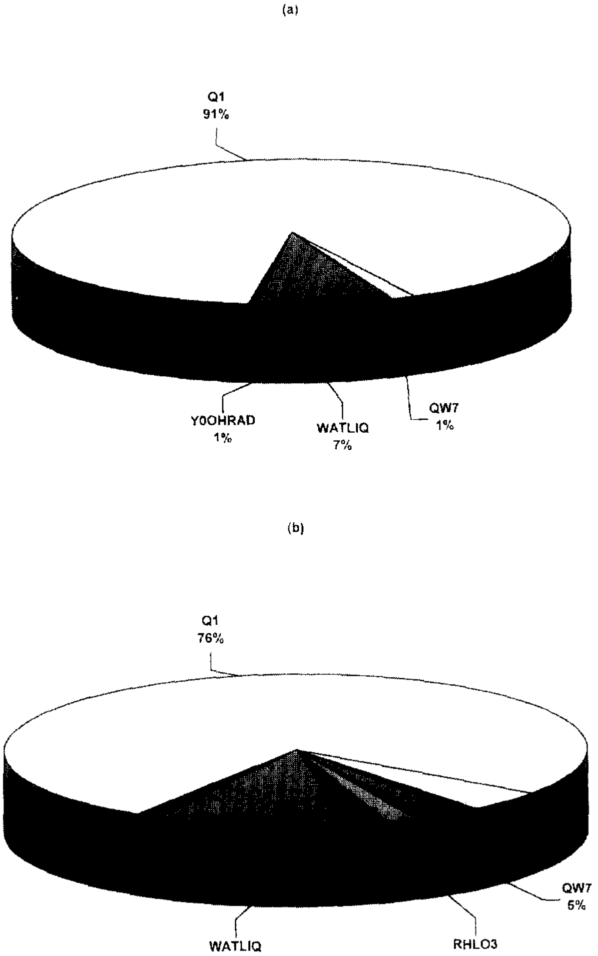

Fig. 3. Results of the FAST quantitative measure. A sample size $N = 1026$ has been used for the calculation of each pair of indices (S_i, S_{T_i}), with $M = 4$, $\omega_i = 64$, and $N_r = 2$. The total cost of the analysis is $C = N \times k = 10260$, being $k = 10$. The pie chart (a) represents the S_i's and the pie chart (b) denotes the S_{T_i}'s.

This conclusion, valid for the 10-factor model where 58 out of 68 uncertain factors were kept fixed to their nominal values, cannot be extended to the original version of the model, where 68 input factor are uncertain. In fact, possible interactions between the elements of the 10-factor model and others that were kept fixed in this analysis are not accounted for by the FAST results. The results of the simplified model are deemed to be more additive than the original one.

The same consideration explains the minor differences in the ranking of the importance of the factors provided by Morris and FAST.

The reliability of the FAST results is conditional upon the choice of the subgroup of factors to which FAST is applied. The choice of the factors to be forwarded to FAST is a subjective and very delicate step.

Typically a SA exercise may fail by two different types of error,

- (Type I) a factor which is not important is erroneously identified; or
- (Type II) a factor which is important is not identified.

While an error of Type I made by Morris would be of no concern, because the factor would simply be assigned a zero or low percentage of variance by the consequent FAST analysis, a Type II error would not be corrected by FAST. Is there a risk to cut off from the analysis a factor that is actually important?

For sure, such a risk cannot be ruled out outright. Previous exercises [7], and iterations of the present one, seem to indicate that Morris does not make Type II errors. This is reasonable, since the influence of factors in models follows – according to our experience – a Pareto-like distribution, with few factors accounting for most of the variance, and most of the factors taking up the remaining bit. One has to build a model artificially, in order to get a case where the influence of factors is more uniformly balanced.

For a model with about 70 factors it is very well likely that less than 10 have some sizeable influence, and in this case two of them clearly eat up more than 90% of the variance. In other words, although it is an arbitrary choice, by selecting the first 10 factors identified by Morris for the FAST analysis we feel reasonably protected from an error of Type II.

Acknowledgements

This work has been partially funded by the Statistical Office of the European Union (Eurostat), contract 13592-1998-02 A1CA ISP LU, Lot 14, SUP COM 97.

References

[1] R. Rosen, Life Itself. A Comprehensive Inquiry into the Nature, Origin, and Fabrication of Life (Columbia Univ. Press, New York, 1991).

[2] R.I. Cukier, C.M. Fortuin, K.E. Shuler, A.G. Petschek, J.H. Schaibly, Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I Theory, J. Chem. Phys. 59 (1973) 3873–3878.

[3] M.D. Morris, Factorial sampling plans for preliminary computational experiments, Technometrics 33 (1991) 161–174.

[4] F. Campolongo, R. Braddock, The use of graph theory in the sensitivity analysis of the model output: a new screening method, Reliability Eng. & System Safety 64 (1997) 1–12.

[5] I.M. Sobol', Sensitivity estimates for nonlinear mathematical models, Matematicheskoe Modelirovanie 2 (1990) 112–118 [in Russian].

[6] A. Saltelli, S. Tarantola, K. Chan, A quantitative, model independent method for global sensitivity analysis of model output, Technometrics 41 (1) (1998) 39–56.

[7] F. Campolongo, A. Saltelli, Sensitivity analysis of an environmental model; a worked application of different analysis methods, Reliability Eng. & System Safety 57 (1997) 49–69.

[8] R.J. Charlson, S.E. Schwartz, J.M. Hales, R.D. Cess, J.A. Coakley, Jr., J.E. Hansen, D.J. Hofmann, Climate forcing by anthropogenic aerosols, Science 255 (1992) 423–430.

[9] R.J. Charlson, J.E. Lovelock, M.O. Andreae, S.G. Warren, Sulfur phytoplankton, atmospheric sulfur, cloud albedo and climate, Nature 326 (1987) 655–661.

[10] A. Saltelli, J. Hjorth, Uncertainty and sensitivity analyses of OH-initiated dimethylsulphide (DMS) oxidation kinetics, J. Atmos. Chem. 21 (1995) 187–221.

[11] J.M. Remedio, A. Saltelli, J. Hjorth, J. Wilson, KIM. A chemical kinetic model of the OH-initiated oxidation of DMS in air: a Monte Carlo analysis of the latitude effect, EUR Report (1994) EN.

[12] S. Koga, H. Tanaka, Numerical study of the oxidation process of dimethylsulfide in the marine atmosphere, J. Atmos. Chem. 17 (1993) 201–228.

[13] F. Campolongo, A. Saltelli, N.R. Jensen, J. Wilson, J. Hjorth, The role of multiphase chemistry in the oxidation of dimethylsulphide (DMS). A latitude dependent analysis, J. Atmos. Chem. (1998), in press.

[14] G.P. Ayers, J.M. Cainey, H. Granek, C. Leck, Dimethylsulfide oxidation and the ratio of methanesulfonate to non-sea-salt sulfate in the marine aerosol, J. Atmos. Chem 25 (1996) 307–325.

[15] E.S. Saltzman, D.L. Savoie, J.M. Prospero, R.G. Zika, Methane sulfonic acid and non-sea-salt sulfate in Pacific air: Regional and seasonal variations, J. Atmos. Chem. 4 (1996) 227–240.

[16] D.L. Savoie, J.M. Prospero, Comparison of oceanic and continental sources of non-sea-salt sulphate over the Pacific Ocean, Nature 339 (1989) 685–689.

[17] T.S. Bates, J.A. Calhoun, P.K. Quinn, Variations in the methanesulfonate to sulfate molar ratio in submicrometer marine areosol particles over the South Pacific Ocean, J. Geophys. Res. 97 (1992) 9859–9865.

[18] N.R. Draper, H. Smith, Applied Regression Analysis (Wiley, New York, 1981).

[19] A. Ray, I. Vassalli, G. Laverdet, G. LeBras, Kinetics of the thermal decomposition of the CH3SO2 radical and its reaction with NO2 at 1 Torr and 298K, J. Phys. Chem. 100 (1996) 8895–8900.

[20] A. Mellouki, J.L. Jourdain, G. LeBras, Discharge flow study of the reaction mechanism using as the source, Chem. Phys. Lett. 2 (1988) 231–236.

[21] R.I. Cukier, J.H. Schaibly, K.E. Shuler, Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. III. Analysis of the approximations, J. Chem. Phys. 63 (1975) 1140–1149.

[22] J.K. Schaibly, K.E. Schuler, Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. Part II. Applications, J. Chem. Phys. 59 (1973) 3879–3888.

[23] R.I. Cukier, H.B. Levine, K.E. Shuler, Nonlinear sensitivity analysis of multiparameter model systems, J. Comput. Phys. 26 (1978) 1–42.

[24] G.E.P. Box, W.G. Hunter, J.S. Hunter, Statistics for Experimenters. An Introduction to Design, Data Analysis and Model Building (Wiley, New York, 1978).

[25] H. Weyl, Mean motion, Am. J. Math. 60 (1938) 889–896.

[26] S. Tarantola, Analysing the efficiency of quantitative measures of importance: The improved FAST, Proc. SAMO 98: Second International Symposium on Sensitivity Analysis of Model Output, Venice, April 19–22 1998, EUR Report 17758 EN (1998).

[27] T. Homma, A. Saltelli, Importance measures in global sensitivity analysis of model output, Reliability Eng. & System Safety 52 (1996) 1–17.