Optimized Resource Allocation for GIS-Based Model Implementation

Michele Crosetto, Francesco Crosetto, and Stefano Tarantola

Abstract
This paper focuses on a new procedure to support the planning and implementation of GIS-based models. Critical decisions are often based on the outputs of such models. A major goal of the GIS planning stage is to implement a model whose output can reliably support the decision process. The procedure allows the above goal to be achieved with an optimized allocation of resources for GIS data acquisition. It is based on two important modeling tools: uncertainty analysis and sensitivity analysis. An application of the procedure to a GIS-based hydrologic model for flood forecasting is discussed.

Introduction
This study focuses on a novel procedure to support the planning and implementation of GIS-based models, i.e., computational models linked with GIS databases. Model implementation, especially when it involves important technical and financial resources, includes many components and requires appropriate planning. The proposed procedure addresses two key aspects of the planning stage: the model output quality, and the quality of model input data.

Requirements related to output quality have a direct impact on many components of spatial models because critical and politically sensitive decisions are often based on the predictions of such models. In order to reliably support the decision process, the output quality is supposed to satisfy precise requirements, tailored to the needs of the application at hand. A major goal of the planning stage is to implement a model whose output meets the established quality requirements. This goal has to be attained with the minimum expenditure of resources. Because the most expensive component of a new GIS is often the data acquisition, the above goal has to be achieved with an optimized allocation of resources for data acquisition. This can be accomplished using the proposed procedure, which focuses on the impact of input data on the model output.

The core of the procedure is the analysis of the quality of (or, conversely, the uncertainty associated with) model output and input data, which is based on two useful modeling tools: uncertainty analysis (UA) and sensitivity analysis (SA). UA allows the assessment of the uncertainty associated with model output as a result of propagation through the model of errors in input data, and uncertainties in the model parameters. SA is useful for determining how much each individual source of uncertainty contributes to the model output uncertainty. It must be noted that, in the field of GIS modeling, different synonyms for UA and SA can be encountered (Crosetto and Tarantola, 2001). The definitions given in this paper refer to the convention employed in Saltelli et al. (2000), a comprehensive handbook on SA that incorporates a wide range of modeling applications.

A literature survey of UA and SA (UASA) techniques in the field of GIS modeling showed that several studies have been devoted to UA, often referred to as error propagation, while SA has received little attention. The majority of UA works are based on numerical techniques, in particular on Monte Carlo simulation (e.g., Heuvelink and Burrough, 1993; Veregin, 1994; Heuvelink, 1998). However, some analytical error propagation methods (usually restricted to particular GIS operations or particular types of data) have been investigated (e.g., Küpper, 1997; Arbia et al., 1998). SA works are almost exclusively based on the simple "one-at-a-time" method (Lodwick et al., 1990; McKenney et al., 1999). Furthermore, UA and SA are almost always treated separately in the GIS literature. A general procedure to perform jointly UA and SA with GIS-based models has not yet been proposed.

The procedure described in this work takes advantage of the synergetic use of UASA. Both these analyses are performed employing Monte Carlo (MC) simulation. In particular, SA is based on a flexible estimation procedure called Extended Fourier Amplitude Sensitivity Test, Extended FAST, (Saltelli et al., 1999). This technique, originally developed for non-spatial models, is in this work extended to GIS-based models and tailored to the specific characteristics of spatial data.

This paper begins with a description of an integrated UASA procedure for non-spatial models. This is followed by a section discussing the extension of the procedure to spatial data. The fourth section describes the main application of the procedure, i.e., its use in the planning stage of a new GIS-based model. The fifth section reviews error models for spatial data, and discusses a few practical examples related to the case study. The sixth section illustrates a comprehensive case study concerning the implementation of a new GIS-based hydrologic model for flood forecasting. Conclusions follow.

An Integrated UA and SA Procedure for Non-Spatial Models
This section concisely describes an integrated UASA procedure for non-spatial models. The procedure, based on MC simulation, treats the analyzed models as black boxes. Therefore, it
can be used with models characterized by any degree of complexity, i.e., without restrictions on the model form. The problem can be formulated as follows. Let us assume that a general model has to be analyzed: i.e.,

\[ Y = f(I_1, I_2, ..., I_k, p_1, p_2, ..., p_m) \]  

where the operator \( f(\cdot) \), which represents the computational model, maps the space of the input data \( I \), with \( i = 1, ..., K \), and model parameters \( p_i \), with \( j = 1, ..., M \), to that of the output variable \( Y \). In this section, \( I, p, \) and \( Y \) are considered scalar variables. Because it may happen that \( Y \) is not a scalar (e.g., it is a 2D field of values), a suitable summary variable needs to be identified. The procedure involves five basic steps, as detailed below.

**Step 1**
Identify all input data \( I_i \) and model parameters \( p_i \) affected by uncertainty that can affect the uncertainty of \( Y \). In this procedure, it is assumed that an appropriate model \( f(\cdot) \) has been selected to describe the relation between \( I, p, \) and \( Y \), i.e., that model error is negligible. This may often not be the case. However, model validation requires external information about the model output and therefore cannot be performed in the context of a simulation-based analysis.

**Step 2**
Find appropriate tools to describe uncertainty, by associating an error model with each \( I_i \) and \( p_i \) identified in Step 1. An error model is a stochastic model capable of generating a population of error-corrupted versions (realizations) of \( I_i \) or \( p_i \) (Goodchild et al., 1992). Each realization is a sample from the same population, and differences between realizations represent the uncertainty present in \( I_i \) or \( p_i \). In the following, the term "input factor" is used to indicate a random variable (with a probability density function [PDF]) that is assumed known a priori, associated with a given error model in order to drive the generation of realizations. In the case of a scalar variable, the corresponding input factor represents the error associated with this variable. It is important to note that the PDF of the input factor reflects the degree of uncertainty associated with the corresponding \( I_i \) or \( p_i \) and has to be chosen using the best information available about the characteristics of \( I_i \) or \( p_i \).

**Step 3**
Propagate error through the model using MC simulation, i.e., evaluate the model (step 1) repeatedly, with randomly generated error-corrupted versions of \( I_i \) and \( p_i \) (note that nominal values are used for all \( I_i \) and \( p_i \) whose uncertainty can be disregarded). In MC-based UA, the random generation can be performed by using various procedures, e.g., simple random sampling, stratified sampling (McKay et al., 1979), etc. In this work, the integrated UASA procedure requires the specific Extended FAST sampling strategy.

**Step 4**
Perform uncertainty analysis. The output of the MC simulation is a sequence of \( Y_k \), with \( k = 1, ..., N_{MC} \) (where \( N_{MC} \) is the number of MC iterations), which represents the empirical PDF of \( Y \). The estimation of the PDF characteristics (e.g., variance and higher order moments) is straightforward. For instance, the variance of \( Y \) can be estimated by

\[ \sigma^2_Y = \frac{1}{N_{MC}-1} \sum_{k=1}^{N_{MC}} (Y_k - m_Y)^2 \]  

where \( m_Y \) is the sample mean. Additionally, a histogram of the output variable \( Y \) can be displayed, thus thoroughly describing the stochastic features of the model output.

**Step 5**
Perform sensitivity analysis. As mentioned above, SA is based on the Extended FAST through the estimation of the so-called sensitivity indices. In this technique, the input factors are assumed independent (note that this hypothesis only concerns the independence of errors associated with \( I_i \) and \( p_i \)). For a given input factor \( X_i \), the first-order sensitivity index \( S_i \) represents the fractional contribution to the variance of \( Y \), \( \text{Var}(Y) \), which is due to \( X_i \) and is defined as

\[ S_i = \frac{\text{Var}(Y|X_i = x_i)}{\text{Var}(Y)} \]  

where \( \text{Var}(Y|X_i = x_i) \) denotes the expectation of \( Y \) conditional on \( X_i \) having a fixed value \( x_i \), and the operator \( \text{Var}(\cdot) \) denotes conditional variance. The reason for this definition is intuitive: if the conditional mean \( \text{E}(Y|X_i = x_i) \) varies considerably with the value \( x_i \) of \( X_i \), while all the effects of the \( X_j \)'s, \( j \neq i \) are averaged, then surely \( X_i \) is an influential factor. In addition to the first-order indices \( S_i \), the Extended FAST provides the estimate of the total sensitivity index \( S_T \), defined as the sum of all the indices (first-order and higher orders) where \( X_i \) is included, thus concentrating in one single term all the interactions involving \( X_i \). An overview of the Extended FAST (sampling strategy and estimation of the indices) is given in the Appendix; full details are given in Saltelli et al. (1999). Advantages and disadvantages of this technique are discussed in Crosetto and Tarantola (2001). A limitation of the technique is related to its computational load. The accuracy of the sensitivity estimates depends on the complexity of the model under analysis and on \( N_{MC} \), which is proportional to the number of input factors. Dealing with models containing hundreds of input factors, or with very computationally expensive models, can make this SA technique unfeasible. In such cases, it is convenient to adopt more economical methods, such as the screening methods (e.g., Morris, 1991), that provide qualitative sensitivity measures (they rank the input factors in order of importance, without quantifying the relative importance).

**Extension of the Procedure to GIS-Based Models and Spatial Data**

The previous section describes a UASA procedure developed for non-spatial models, where, in particular, the input data \( I_i \) are considered scalar variables. The procedure can be easily extended to analyze general GIS-based models driven by any kind of spatial data (e.g., raster, vector, etc.). In the following, \( Y \) is considered a scalar variable. In case it is not, a summary variable has to be derived.

All steps but Step 2 of the above procedure remain the same. Step 2 includes more elaborated error models to generate error-corrupted versions of the spatial data \( I_i \) and model parameters \( p_i \), typically restricted to be scalar variables. Several error models have been proposed in the literature; a brief overview is given in the section on Error Models for Spatial Data. The proposed procedure can include most of these models, i.e., the state-of-the-art in this field.

The key feature of Step 2 is that the generation of realizations of spatial data \( I_i(\cdot) \) is still controlled by a set of non-spatial input factors, thus allowing the SA to be performed with the same procedure used for non-spatial models. In case \( I_i(\cdot) \) is a scalar variable, the corresponding input factor represents its associated error. For more complex types of spatial data (e.g., quantitative raster data), the uncertainty can be decomposed into different components (e.g., systematic and stochastic errors). It is possible to treat these components separately using...
suitable models, as discussed in the section on Error Models for Spatial Data. In this case, the generation of error-corrupted versions of a single \( I_i \) can be driven by a set of input factors, each factor representing a particular aspect of uncertainty.

Performing the SA on input factors makes the procedure very effective. In fact, the control of complex spatial applications can be achieved by means of a limited set of input factors. Let us consider the case where \( I_i \) is a digital elevation model (DEM) with \( 10^6 \) grid points. The generation of DEM realizations can be controlled by only few input factors (e.g., a bias and a factor related to the stochastic error). A larger number of factors could be used, but the number will never be increased too much. For instance, it is theoretically possible to associate an input factor with each single DEM point. However, it will be practically impossible to analyze \( 10^6 \) input factors.

The SA is performed assuming the input factors are independent. There is no condition of independence in the spatial data (e.g., DEM, land-cover map) which enter the model. This assumption only concerns the mechanism of generating corrupted versions of \( I_i \) and \( p_j \) (it concerns their associated errors) and is often justified by the fact that the different spatial data come from different sources (different acquisition techniques, etc.).

### An Application: Support for the Planning Stage of GIS-Based Models

This section describes the main application of the above procedure, i.e., its use in the planning stage of new GIS-based models. The procedure addresses two complementary aspects. First, through UA, it addresses the following question: considering all sources of uncertainty simultaneously, does the model output support, reliably, the decision-making process? Then, through SA, it considers uncertainty in the opposite direction (from the output back to the input), determining how much each individual source of uncertainty contributes to the output uncertainty.

These two aspects can be exploited in order to implement a GIS-based model whose output can reliably support the decision process with an optimized allocation of resources for GIS data acquisition. Let us assume that a new Service, e.g., the flooding forecast for civil protection purposes, has to be implemented at the national level. This may involve the acquisition of a large GIS database and require important financial and technical resources. In the planning of such a system, the following low cost strategy, based on numerical simulations on a small prototype of the GIS-based model, can be employed.

### Step 1

Identify a test area suitable to implement a prototype representative of the entire GIS. Implement the GIS prototype, i.e., collect for this area all spatial data needed to run the GIS-based model.

### Step 2

Identify all spatial data \( I_i \) and model parameters \( p_j \) that can be affected by uncertainty that can affect the uncertainty of \( Y \).

### Step 3

Specify the project hypotheses about the quality of spatial data and model parameters, i.e., make assumptions about the level of uncertainty associated with each \( I_i \) and \( p_j \). It must be noted that these hypotheses will have a direct impact on the resources to be allocated for model implementation.

### Step 4

Perform the UA, i.e., evaluate the stochastic features of the model output, conditional on the project hypotheses of Step 3. If the output quality (e.g., measured by the variance) meets the established requirements, then skip to Step 7. This step guarantees that the model will reliably support the underlying decision process. It is possible that the requirements are exceeded. In this case, some hypotheses on input factors can be relaxed.

### Step 5

Perform the SA in order to assess the relative importance of the different sources of uncertainty. This step ensures an optimized allocation of resources for data acquisition. In fact, if the output quality is inadequate to the task, the quality of the input must be improved. The maximum benefit is obtained by acting on the subset of inputs with the highest sensitivity indices. In the opposite case (the output quality exceeds the requirements), the specifications on input quality levels can be relaxed, starting from the inputs with the lowest sensitivity indices.

### Step 6

Specify new project hypotheses about the quality of spatial data and model parameters. Iterate Steps 4, 5, and 6 until Step 4 is satisfied.

### Step 7

Use the project hypotheses, verified through UA on the prototype, as the technical specifications for the data acquisition procedure of the entire GIS.

The above strategy can be employed to answer in a scientifically sound way the following questions during the GIS planning stage:

- What is the required quality level for each spatial input data? The optimal level (in terms of cost-benefit analysis) is achieved through Steps 3 to 7. An advantage of the procedure is that the control of the characteristics of all spatial input data is achieved by means of a limited set of input factors, that have a practical meaning in the data acquisition (e.g., for a DEM the stochastic error, represented by the standard deviation, and the bias).
- What is the optimal level of resolution for spatial input data? Spatial resolution is an important characteristic of data quality. However, this aspect cannot be addressed in the procedure by generating error-corrupted versions of the data. An extension of the procedure has to be introduced. Let us assume that we can choose between two levels of resolution (\( R_a \) versus \( R_b \)): which is the best one in terms of cost benefit analysis? In this context, it is possible to treat a subjective decision of the modeler as an input factor. In addition to all input factors associated with \( p_j \) and \( I_i \), a switch between \( R_a \) versus \( R_b \) is added. For each MC iteration, the switch decides, randomly and with equal probability, which data resolution has to be employed. The sensitivity index associated with the switch measures how sensitive the model output is to the choice between \( R_a \) and \( R_b \). A low index (little influence on the output) means that the two resolution levels are practically equivalent. Hence, the cheaper one can be adopted. Otherwise, the more expensive level (higher resolution) has to be chosen.

### Error Models for Spatial Data

This work combines UA techniques, originally developed for non-spatial data, with suitable tools for describing and simulating errors in spatial data. Error modeling represents an important research topic in the field of GIS. In the following sections, some strategies to model errors in quantitative raster, qualitative raster, and vector data are outlined. For each data type an example related to the case study is discussed.
Quantitative Raster Data

Fields of attribute data (raster data) represent an important input to GIS-based models. Raster data can be either quantitative (e.g., terrain heights) or qualitative, also called categorical (e.g., landcover types). They are usually 2D fields, but 3D and 4D fields are employed in some applications. An example of 3D input is the rainfall intensity field described further in this section.

Much research has been focused on error modeling in quantitative raster data. The proposed models mainly differ in the way the correlation structure of the error is addressed. Nackaerts et al. (1999) consider a DEM and model the associated error assuming no spatial correlation (white noise). Davis and Keller (1997) introduce error spatial correlation by using a smoothing filter over an error-corrupted version of a DEM. A discussion of error modeling and techniques to generate spatially correlated error fields is given in Heuvelink (1998). Arbia et al. (1998) use a general model capable of describing attribute error as well as location error.

The model discussed here allows the description of the error correlation structure. Let us assume that a quantitative field \( A(\mathbf{x}) \), where \( \mathbf{x} \) is the domain of \( A \) defined in the space \( R^n \), constitutes the input to a given model. Usually a particular realization \( S(\mathbf{x}) \) of the field \( A(\mathbf{x}) \), available in the GIS database, is used. A suitable error model is given by

\[
Z(\mathbf{x}) = S(\mathbf{x}) + N(\mathbf{x})
\]

where \( S(\mathbf{x}) \) is the deterministic field, \( Z(\mathbf{x}) \) is the corresponding error-corrupted random field, and \( N(\mathbf{x}) \) is the error random field with mean \( m_N \) and variance \( \sigma_N^2 \). This model can describe two important features of uncertainty. A systematic attribute error (bias) is introduced by fixing non-zero values for \( m_N \), while a stochastic attribute error is introduced by setting \( N(\mathbf{x}) \) to be random with a given variance-covariance matrix \( C_{NN} \). This matrix permits the description of the spatial correlation structure of \( N(\mathbf{x}) \). The generation of spatially correlated error fields can be computationally expensive, and suitable techniques have to be employed. One of these techniques, based on the Cholesky decomposition, is described in Crosetto et al. (2001).

With the above model, the generation of \( Z(\mathbf{x}) \) can be driven by two input factors representing the systematic and stochastic error respectively, i.e., the uncertainty can be decomposed in two components. There are important differences in the way these components are treated in the analysis. The systematic error \( m_N \) is per se a scalar variable, and its associated input factor coincides with the error \( m_N \) itself. On the other hand, the stochastic error has an intrinsically multi-dimensional nature. In fact, assuming a 2D field with \( n \) rows and \( m \) columns and \( C_{NN} = \sigma_N^2 \cdot I \), a single parameter \( \sigma_N^2 \) controls the extraction of \( n \times m \) different values. That causes problems in the SA, i.e., in the output variance decomposition. The same applies in the case of non-diagonal matrices \( C_{NN} \). An empirical way to treat these kinds of parameters is to use a switch between two configurations as an input factor: in the first one, the stochastic error is generated and added to the 2D field while, in the second one, no stochastic error is added. In practice, the switch is realized by extracting \( e \) from a uniformly distributed interval \([0,1]\). If \( e \leq 0.5 \), the first configuration is used; otherwise, the second one is employed. With this stratagem, the sensitivity index associated with such an input factor measures how sensitive the model output is to the switch, and hence it gives an indication of how sensitive it is to the stochastic error.

In the following, an application of the above error model is illustrated. The main input to the hydrologic model described in the case study is the rainfall intensity, a 3D quantitative field which varies spatially (2D in space) and temporally, has an original spatial resolution of 2 km (then upsampled to 50 m) and a temporal resolution of 30 minutes (then upsampled to 15 minutes). It consists of 325 and 316 samples in the west-east direction and south-north direction, respectively, and 11 temporal samples, for a total of 1,129,700 voxels (volume elements). For the MC analysis, a rainfall scenario was assumed, with total rainfall duration of 2.5 hours and an average intensity of 16.1 mm/hour. This scenario was employed as nominal realization (deterministic field) in the MC simulation (Figure 1). The associated error field was built assuming a systematic error \( m_N \), uniformly distributed in the range \([-4, +4 \text{ mm/h}] \), and a stochastic error, normally distributed, spatially correlated, and temporally uncorrelated (Figure 1).

At each MC iteration, an error-corrupted realization of the rainfall field was generated extracting \( m_N \) from the uniform distribution, adding \( m_N \) to the rainfall intensity of all voxels, and generating, for each of the 11 temporal steps \( t \), a normally distributed and spatially correlated stochastic error field. The random generation was performed through Cholesky decomposition of the \( C_{NN} \), built using the autocovariance function depicted in Figure 1. The eleven 2D error fields (325 by 316 pixels) were generated separately because the error was assumed to be temporally uncorrelated.

Qualitative or Categorical Raster Data

Error modeling for qualitative raster data is not straightforward. For those data coming from remotely sensed images, usually processed with image classification techniques, suitable error models can be based on the information contained in the so-called confusion or error matrix. This matrix can be used to estimate the percentage of correctly classified data and other information about the nature of errors. Using this information, it is possible to generate error-corrupted versions of this type of fields.

An interesting error model for qualitative raster data is proposed in Davis and Keller (1987), where a fuzzy spatial model,
called "corridor of transition model," is employed to represent both attribute error and attribute boundary uncertainty. A model for simulating inclusions of unmapped units within areas delineated on the map as uniform is described in Fisher (1991). Veregin (1994) studies the propagation of error in categorical attributes values through a particular type of operation (the buffer operation), while a model for a particular type of raster data (containing the so-called mixed pixels) is proposed in Goodchild et al. (1992).

In the following, an example related to the case study is discussed. An input to the hydrologic model is the Manning map, derived from a land-cover map by assigning to each cover class a nominal Manning's roughness coefficient. The uncertainty associated with the Manning map was decomposed into three components: two components related to the land-cover map and the third one related to the Manning's roughness coefficients. The latter one was modeled adding to the nominal coefficients an error, uniformly distributed in the range [−20%, +20%] of the nominal Manning's values. The land-cover map was produced in vector form and then rasterized by IPLA (Istituto Piante da Legno e per l'Ambiente, Turin, Italy) using 1:100,000-scale maps as a topographic base and large-scale aerial images and field work for the thematic data acquisition. This map consists of large portions of homogeneous classes. Classification errors in the map would affect any large portions of the covered area and, considering the map quality level, were excluded. It was assumed that the main sources of uncertainty in the map are represented by attribute boundary uncertainty and by inclusions of unmapped units within uniform areas.

Adopting the epsilon band approach (see, for instance, Shi (1998)), the boundary uncertainty was modeled as a 100-m band on both sides of the cover-class boundaries. The idea of the epsilon band model is that boundary location is only known within a certain confidence region. By randomly extracting boundary positions within this region, it is possible to generate error-corrupted versions of the map at hand. From the nominal map, a set of error-corrupted maps was generated by changing the class boundaries. Changes were made considering both the band width and the land-cover classes (the algorithm includes some heuristic rules, e.g., change a boundary only if neighbor classes have sensibly different Manning's roughness coefficients). The input factor associated with the boundary error was a "multipletag way switch." At each iteration a variable e, uniformly distributed in the interval [0, 1], is extracted and the kth band is chosen and used in the model if (k − 1)/m ≤ e/k/m, where m is the total number of maps.

Error in categorical raster data, in the form of map-unit inclusions, may represent an important component of uncertainty. In order to model this type of error in the land-cover map (a sample map is presented in Figure 2), an algorithm for simulating inclusions, similar to that described in Fisher (1991), was employed. In the simulation, different aspects were considered. Some inclusions were a priori excluded considering natural constraints. For instance, in the area beech trees do not grow above an elevation of 1600 m; hence, "Beech wood" inclusions above this limit were not allowed. Considering the four land-cover classes included in Figure 2, the following assumptions were made to control the simulation: no inclusions in "Bare soil" units which cover areas located over 2000 m above sea level, i.e., outside the elevation range for all other classes; 5 percent of "Bare soil" in "Coniferous" units; 4 percent of "Pastures" in "Beech wood" units; and 4 percent of "Bare soil" plus 6 percent of "Coniferous" in "Pastures" units.

In Figure 2 (right), a realization of the above inclusions is presented. The effect of these inclusions on the model output is model and context dependent, and can only be assessed through SA. In the case study described further in this paper, the map-unit inclusions proved to have very little impact on the output of the hydrologic model. This was verified through a preliminary screening method. Therefore, this type of uncertainty was not considered in the UASA of the case study.

Vector Data
Vector data have received great attention in the field of error modeling. A simple approach consists of modeling the positional uncertainties in vector data by random errors in the constituent points. General models for positional errors in vector data are described in Kiiveri (1987), Shi (1998), and Shi and Liu (2000).

Uncertainty identification, the first step in UASA, is intimately related to the role that each input plays in the model. This is illustrated in the following example, focused on the area subject to flooding risk considered in the case study. The vector data, derived by topographic survey, describe two main zones: the river bed and a recreation center located near the river. The uncertainty associated with the vector data of the two zones has different characteristics, and must be represented by two different input factors. The recreation center consists of geometrically well defined features whose height is used to derive the warning and flooding river stages. The positional uncertainty of these features, due to errors in the topographic survey, was described by a normally distributed zero-mean error with standard deviation of 5 mm in the three coordinates.

In the hydrologic model, the river geometry plays a key role in the computation of the forecasted river level. Uncertainty in the river geometry reconstruction depends both on the topographic measurement errors and on the river bed spatial variability compared to the measurement density. Because the river bed is characterized by a high spatial variability, reconstruction uncertainty is mainly due to the measurement density. The uncertainty was modeled by a normally distributed error with zero mean and standard deviation of 10 cm in the three coordinates. The error in the three coordinates was assumed to be uncorrelated. In the case study, the uncertainty associated with the recreation center features proved, through a screening SA, to have a negligible impact on the model output. Therefore, it was not considered in the analysis.

Case Study: Implementation of a GIS-Based Flooding Forecasting Model

The use of the UASA procedure as support for the planning stage of GIS-based models is illustrated in the following implementation of a hydrologic model to be integrated in a system for near real-time flood forecasting. The system is conceived to receive as input the forecasted rainfall over a given area, to simulate the...
hydrologic process and to forecast the magnitude of peak flood discharges at given locations.

The model was implemented employing the GRASS GIS software, a public domain software that includes programs for hydrologic simulations (information can be found at www.baylor.edu/~grass). The main part of the model was based on the CASC2D program, a physically based, distributed, raster model, which simulates the hydrologic response of a watershed subjected to a given rainfall field (Saghafian, 1996). Major components of the model include interception (i.e., the process whereby rainfall is retained by vegetation), infiltration, and surface runoff routing (see for details the site www.geog.uni-hannover.de/grass/modelintegration.html).

Let us assume that the system has to be implemented for a large portion of the Italian Alps, thereby requiring important financial resources. The planning of the system, based on the procedure proposed in this work, allows the optimization of such resources to be achieved, as detailed below.

**Step 1**
Identification of a test area. The chosen test area covers a little alpine valley (Gesso Valley) included in the Regional Park of the Maritime Alps that is a part of the Western Italian Alps (Figure 3). The bottom valley hosts a popular recreation center, located near the Gesso River in an area subjected to flooding risk. The model inputs were derived from existing cartographic maps and geographical databases.

**Steps 2 and 3**
Identification of all input data and model parameters that can be affected by uncertainty, and specification of the project hypotheses about their quality. Input data and associated uncertainties are concisely listed below.

- Rainfall intensity (Figure 1). Details were reported in the previous section.
- Vector data of the recreation center. Details were reported in the previous section.
- Digital terrain model (DTM). A 50-m DTM, by-product of the 1:10,000-scale maps produced by the Cartographic Service of Piedmont Region, Italy, was employed. The DTM is one of the most important inputs for the hydrologic model. The errors of the DTM (e.g., unreal pits) usually create problems for the numerical techniques used in physically based, distributed models. In order to avoid numerical instability of the model, the DTM underwent a time-consuming pre-processing (DTM filtering and manual editing) and all simulations were performed with very short computational time steps. The last aspect resulted in a high computational load of the model simulations, while the need of a pre-processing made impossible to use error-corrupted realizations of the DTM in the UASA. Therefore, in the UASA the DTM errors were not considered.
- Map of Manning's roughness coefficient. Details were reported in the previous section.
- Storage capacity map and interception coefficient map. These maps were derived by re-classification of the land-cover map (obtaining the interception map), and assigning to each class the capacity and interception coefficient values. The uncertainty associated with these maps was represented by a class boundary error in the interception map (200-m band) and an error (normally distributed with zero mean and standard deviation that equals the 9 percent of the corresponding nominal values) in the capacity values and the interception coefficients.
- Infiltration maps: hydraulic conductivity, pressure, and porosity maps, derived by re-classification of a soil texture map. A 200-m band was used to model boundary positional uncertainty in the soil texture map. To the conductivity, pressure, porosity values an error uniformly distributed in the range $[-20\%, +20\%]$ of the corresponding nominal values was added.
- Soil moisture map. A scenario, antecedent the beginning of the rainstorm, was simulated, assuming a constant moisture content of 15 percent in the permeable terrain. This simple approximation, which was chosen for both the project purposes, could be improved in a further analysis using more refined wetness indices. Moisture measurement uncertainty was modeled by adding to the moisture content an error uniformly distributed in the range $[-20\%, +20\%]$ of the nominal value.

**Steps 4 and 5**
Uncertainty and sensitivity analysis. The complete list of the 13 input factors is reported in Table 1. Because the number of input factors is quite high, a preliminary SA, based on the Morris technique, was performed. According to the results published in literature (Morris, 1991), 56 samples (four samples for each factor) were chosen. In Table 2, the input factors are ranked according to their relative importance on the output uncertainty. Due to the important difference between the top-ranked five factors and the other ones, it was decided to analyze the first five factors using the Extended FAST. Following the rule of thumb of about 100 MC iterations for each input factor, 485 model evaluations were performed. From the PDF of the forecasted river level $H$, the standard deviation of $H$ was estimated: i.e.,

$$\sigma_H = 0.49 \text{ m}$$

The standard deviation $\sigma_H$ gives an assessment of the model output quality. Assuming such quality to be inadequate, SA was performed in order to determine which input factors mostly drive the output uncertainty. The Extended FAST total sensitivity indices $S_H$ are reported in Table 3 (First Analysis).

**Step 6**
Specification of new project hypotheses. The bias on the rainfall intensity represents the major source of uncertainty ($S_H = 0.84$). Aiming at reducing the prediction uncertainty, the best way is to improve the accuracy in rainfall forecasting. A second analysis was therefore performed by reducing the range of the systematic error in the rainfall intensity from $[-4, +4 \text{ mm/h}]$ to $[-2.5, +2.5 \text{ mm/h}]$. This, of course, represented a more severe design constraint in the rainfall forecasting system.

**Steps 4 and 5 (Second Iteration)**
Uncertainty and sensitivity analysis. A new set of 485 model evaluations was performed, obtaining a new estimate of the standard deviation of the forecasted river level $H$: i.e.,

$$\sigma_H = 0.35 \text{ m}$$

The histogram of $H$ is shown in Figure 4. In the same figure, the warning stage for the analyzed area, which equals 3.74 m, is indicated. The distribution of $H$ can be compared, with suitable
TABLE 1. INPUT DATA, INPUT FACTORS, AND ASSOCIATED DISTRIBUTIONS \{UNIFORM, U[\text{range}] \}; NORMAL, G[\text{mean, stdev}]; n.v. STANDS FOR NOMINAL INPUT VALUE

<table>
<thead>
<tr>
<th>Input Data</th>
<th>Input Factor Description</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector data</td>
<td>Error on point coordinates</td>
<td>U: [-4; 4 mm/h]</td>
</tr>
<tr>
<td>Rainfall intensity</td>
<td>Bias</td>
<td>([-100; 100 \text{ m}])</td>
</tr>
<tr>
<td>Rainfall intensity</td>
<td>Stochastic error field</td>
<td>([-0.2; 0.2 \ast \text{n.v.}])</td>
</tr>
<tr>
<td>Manning map</td>
<td>Error on Manning’s coefficients</td>
<td>([-100; 100 \text{ m}])</td>
</tr>
<tr>
<td>Manning map</td>
<td>Error on Manning’s coefficients</td>
<td>([-0.2; 0.2 \ast \text{n.v.}])</td>
</tr>
<tr>
<td>Storage capacity map and interception coeff. map</td>
<td>Error on boundaries of the interception map</td>
<td>([-100; 100 \text{ m}])</td>
</tr>
<tr>
<td>Storage capacity map</td>
<td>Error on interception coefficients</td>
<td>([-0.2; 0.2 \ast \text{n.v.}])</td>
</tr>
<tr>
<td>Interception coeff. map</td>
<td>Error on capacity values</td>
<td>([-0.2; 0.2 \ast \text{n.v.}])</td>
</tr>
<tr>
<td>Conductivity map, Pressure map, and Porosity map</td>
<td>Error on conductivity values</td>
<td>([-0.2; 0.2 \ast \text{n.v.}])</td>
</tr>
<tr>
<td>Conductivity map</td>
<td>Error on pressure values</td>
<td>([-0.2; 0.2 \ast \text{n.v.}])</td>
</tr>
<tr>
<td>Capillary pressure map</td>
<td>Error on porosity values</td>
<td>([-0.2; 0.2 \ast \text{n.v.}])</td>
</tr>
<tr>
<td>Porosity map</td>
<td>Error on pressure values</td>
<td>([-0.2; 0.2 \ast \text{n.v.}])</td>
</tr>
<tr>
<td>Soil moisture map</td>
<td>Error on porosity values</td>
<td>([-0.2; 0.2 \ast \text{n.v.}])</td>
</tr>
</tbody>
</table>

TABLE 2. INPUT FACTOR RANKING ACCORDING TO THE MORRIS’S SCREENING ANALYSIS

<table>
<thead>
<tr>
<th>Input Factor</th>
<th>Relative Importance</th>
<th>Input Factor</th>
<th>Relative Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias on rainfall intensity</td>
<td>1.74</td>
<td>Error on pressure values</td>
<td>0.10</td>
</tr>
<tr>
<td>Error in the point coordinates</td>
<td>0.66</td>
<td>Error on Manning coefficients</td>
<td>0.09</td>
</tr>
<tr>
<td>Error on porosity values</td>
<td>0.33</td>
<td>Error on interception coefficients</td>
<td>0.06</td>
</tr>
<tr>
<td>Error on capacity values</td>
<td>0.28</td>
<td>Error on capacity values</td>
<td>0.04</td>
</tr>
<tr>
<td>Error on soil moisture values</td>
<td>0.28</td>
<td>Error on Manning map</td>
<td>0.01</td>
</tr>
<tr>
<td>Error on conductivity values</td>
<td>0.16</td>
<td>Error field on rainfall intensity</td>
<td>0.01</td>
</tr>
<tr>
<td>Error on soil texture map</td>
<td>0.11</td>
<td></td>
<td>-</td>
</tr>
</tbody>
</table>

TABLE 3. RESULTS OF THE TWO SENSITIVITY ANALYSES PERFORMED WITH THE EXTENDED FAST

<table>
<thead>
<tr>
<th>Input Factors</th>
<th>Total Sensitivity Indices</th>
<th>First Analysis</th>
<th>Second Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias on rainfall intensity</td>
<td>0.84</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td>Error in the point coordinates</td>
<td>0.06</td>
<td>0.32</td>
<td></td>
</tr>
<tr>
<td>Error in Manning map</td>
<td>0.08</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>Error on capacity values</td>
<td>0.05</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>Error on porosity values</td>
<td>0.04</td>
<td>0.10</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4. Histogram of the forecasted river level \(H\) obtained with the Second Analysis. The distribution of \(H\) can be compared with the warning stage (3.74 m).

Conclusions

In this paper, a novel procedure to support the planning and implementation of GIS-based models has been proposed. The procedure takes advantage of the synergetic use of UA and SA. An integrated UASA procedure for non-spatial models has been described. An extension of the procedure to GIS-based models and spatial data has been proposed.

The main application of the procedure, i.e., its use in the planning stage of new GIS-based models, has been discussed. Following a low cost strategy, based on numerical simulations on a small model prototype, important benefits in the implementation of new models can be obtained.

In the last part of the paper, a comprehensive case study has been described. It concerns the implementation of a new GIS-based hydrologic model for flood forecasting, which receives several types of input data. The following aspects have been highlighted:

- The importance of evaluating the output quality of a model for risk assessment. Through UA it is possible to support in a sound way a decision process which can have an important impact on the life of many people.
- The effectiveness of the UASA procedure to support an optimized implementation of expensive systems.
- The importance of error modeling for spatial data. This is the most difficult aspect of the entire procedure. In fact, the error modeling is very context dependent and the modeler must choose the error models tailored to the input data and the model at hand.

In this work the output of the analyzed model has been assumed to be a scalar variable. This requires adopting a suitable summary variable in case the output is not a scalar. A future development of this work will be the extension of the procedure to non-scalar output variables, such as 2D fields of values.
Acknowledgments
This work has been partially funded by the Statistical Office of the European Union (EUROSTAT), contract 14756-1999-01 AICA ISPLU, Lot 20, SUP COM 98.

References

Appendix
The FAST, proposed by Cukier et al. (1973), was employed to investigate the sensitivity of large sets of coupled chemical reaction systems to uncertainties in kinetic rate coefficients. In a further article, the method was reviewed and re-interpreted so as to fit into an ANOVA setting (Cukier et al., 1978). FAST supplies estimates of the first-order sensitivity indices $S_i$. A curve $\Gamma$ is defined over the space $\Omega$ of the input factors $X_i$, with $i = 1, \ldots, k$ by using a set of parametric equations: i.e.,

$$X_i(s) = G_i(s, \omega_i), \quad i = 1, \ldots, k.$$  \hfill (A1)

The $\omega_i$ are integer numbers and the operators $G_i$ are chosen in order to achieve the desired probability distributions for $X_i$. As the parameter $s$ is varied over $(-\pi_1, \pi_1)$, each $X_i$ varies over its range of uncertainty making $\omega_i$ oscillations. A careful choice of $\omega_i$ allows $\Gamma$ to systematically explore $\Omega$, and it represents the strategy to split, through Fourier transform, the output variance $V(Y)$ according to its components. The output $Y = f(X_1(s), X_2(s), \ldots, f(s))$ is expressed as function of $s$. The Fourier integrals

$$A(\omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(s) \cos \omega s \, ds, \quad B(\omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(s) \sin \omega s \, ds \hfill (A2)$$

are computed, and the spectral components $A^2(\omega_i)$ of $f(s)$ are then obtained at each frequency $\omega$ as


The sensitivity index $S_i$ for each $X_i$ is obtained as $V_i / V(Y)$. $V_i$ is estimated by collecting the information over the frequency domain at $\omega_i$ and higher harmonics: i.e.,

$$V_i = 2 \sum_{p=1}^{+\infty} A^2(\omega_i), \quad i = 1, \ldots, k \hfill (A3)$$

The choice of the frequencies represents a very delicate step, which requires avoiding interferences to occur among the higher harmonics of $\omega_i$. Technical details are given in Cukier et al. (1973). The quadratures in Equation (A2) are calculated numerically. A sample of size $N$ is generated along $\Gamma$ by selecting evenly spaced values of $s$ over $(-\pi_1 + \pi_1)$. $N$ is the total number of model evaluations that are required to estimate the whole set of $S_i$.

An extension of the FAST method was developed by Saltelli et al. (1999) to estimate also the total effects $S_{ij}$. The curve $\Gamma$ is defined by Equation (A1) where a “high” value of $\omega_i$ is selected for a given $X_i$ and a set of “low” values are assigned to the other frequencies, $\omega_{i-j}$, that correspond to the remaining factors. A sample of size $N_i$ is generated along $\Gamma_i$ and the indices $S_i$ and $S_{ij}$ for $X_i$ can be estimated after evaluating the model at the $N_i$ points. $S_i$ is obtained using Equation (A3), but only for $X_i$, whereas $S_{ij}$ is given by $1 - V_i / V(Y)$, where

$$V_{ij} = 2 \sum_{j=1}^{\infty} A^2(\omega_i), \quad i = 1, \ldots, k \hfill (A5)$$

A different curve $\Gamma_i$, and a new sample (e.g., of size $N_i$), is required to estimate the pair of $S_i$ and $S_{ij}$ for another factor. The total cost of the analysis is then $N = k \cdot N_i$. (Received 26 January 2001; accepted 01 June 2001; revised 23 August 2001)