Journal of Hazardous Materials 176 (2010) 521–526

Contents lists available at ScienceDirect

Journal of Hazardous Materials

journal homepage: www.elsevier.com/locate/jhazmat

A stochastic optimization model under modeling uncertainty and parameter certainty for groundwater remediation design—Part I. Model development

L. He[a,*], G.H. Huang[b,c], H.W. Lu[b]

[a] Department of Civil Engineering, Faculty of Engineering, Architecture and Science, Ryerson University, 350 Victoria Street, Toronto, Ontario, Canada M5B 2K3
[b] Environmental Systems Engineering Program, Faculty of Engineering, University of Regina, Regina, Saskatchewan, Canada S4S 0A2
[c] College of Urban Environmental Sciences, Peking University, Beijing 100871, China

ARTICLE INFO

Article history:
Received 28 August 2009
Received in revised form 4 November 2009
Accepted 8 November 2009
Available online 14 November 2009

Keywords:
Groundwater remediation
Remediation design
Simulation
Optimization
Multivariate analysis
Modeling uncertainty

ABSTRACT

Solving groundwater remediation optimization problems based on proxy simulators can usually yield optimal solutions differing from the "true" ones of the problem. This study presents a new stochastic optimization model under modeling uncertainty and parameter certainty (SOMUM) and the associated solution method for simultaneously addressing modeling uncertainty associated with simulator residuals and optimizing groundwater remediation processes. This is a new attempt different from the previous modeling efforts. The previous ones focused on addressing uncertainty in physical parameters (i.e. soil porosity) while this one aims to deal with uncertainty in mathematical simulator (arising from model residuals). Compared to the existing modeling approaches (i.e. only parameter uncertainty is considered), the model has the advantages of providing mean-variance analysis for contaminant concentrations, mitigating the effects of modeling uncertainties on optimal remediation strategies, offering confidence level of optimal remediation strategies to system designers, and reducing computational cost in optimization processes.

1. Introduction

Currently thousands of aquifers have been contaminated due to leakage of petroleum hydrocarbons from underground storage tanks or pipelines in Canada [1]. This has posed significant risks to environmental quality and human health [2]. Therefore, a number of in situ remediation techniques have been applied for removing petroleum contaminants from soil and groundwater [3]. To facilitate remediation designs, researchers have developed a number of integrated simulation and optimization models to provide decision support for identifying the most cost-effective groundwater remediation strategies. Simulation models were used to forecast the fate of contaminants in subsurface environments under various conditions, while optimization ones were to screen an optimum design from a variety of potential alternatives.

Previously, a large number of studies were focused on developing simulation-based optimization approaches for design of groundwater remediation systems [4–14]. Some researchers attempted to directly incorporate a simulator into an optimization formulation, with the outputs from each simulation run being used to check whether or not the environmental and/or hydraulic constraints are satisfied [15,16]. Nevertheless, some preferred to replace the simulator with an approximated linear/quadratic transition/statistical one (referred to as proxy simulator) and then linked to the optimization formulation. Due to the reduced computational cost, many remediation systems were designed by solving optimization formulations based on the outputs from proxy simulators instead of original numerical simulators.

In detail, Cooper Jr. et al. [17] presented a simulation/regression/optimization (S/R/O) approach to predict, analyze, and optimize an oil recovery process. The application of the S/R/O approach to a simple representative problem revealed its capability in helping make cost-effective operation and management decisions. A number of power-form nonlinear regression equations were provided to describe relationships between system responses and time-varying water-pumping rates. McPhee and Yeh [18] presented an approach to solving groundwater management problems with reduced computational cost. In the approach, they used a simple model governed by an ordinary differential equation (i.e. proxy-model) to replace a groundwater flow model governed by a partial differential equation. Results from the models comparison showed that the proxy-model was able to reproduce head variations in the flow domain with good accuracy. He et al. [19] proposed a simulation-based fuzzy chance-constrained programming model based on the concept of possibility. Application of the model to a practical site showed that the model has the capabilities of handling fuzzy simulation and optimization problems, providing a possibility indicating how much degree one can believe the decision results,

* Corresponding author. Tel.: +1 416 979 5000x6459; fax: +1 416 979 5122.
E-mail address: li.he@ryerson.ca (L. He).

0304-3894/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhazmat.2009.11.060

and alleviating the computational burdens in finding the optimal solutions. He et al. [20] used a coupled simulation–optimization approach to supporting groundwater remediation design under parameter uncertainty. A remediation design model was then solved by introducing a set of proxy simulators. Results showed that the introduced approach was useful for alleviating computational cost in searching for optimal solutions and giving confidence levels for the obtained optimal remediation strategies.

Although the above efforts were useful for increasing computational efficiency, a major concern still needs to be addressed regarding the mathematical-modeling uncertainty arising from residuals introduced by proxy simulators. While considerable efforts have been made to improve proxy-simulator precision, residuals unavoidably exist resulting from model assumptions, statistical samples, model structures, and computational errors. One can hardly conclude whether or not this uncertainty can be ignored due to propagation of residuals in optimization processes. Virtually even if the residuals were small enough, the obtained optimal solutions probably deviate from the "true" ones of the problems to a large extent, due to the approximation errors. However, the previous studies seldom attempted to investigate and mitigate the impact of modeling uncertainty on optimization results.

Taguchi [21] proposed a dual-phase response analysis (DPRA) method to handle modeling uncertainty in production management systems. The principal idea was to fit two proxy simulators respectively for the mean and variance of the response variable. DPRA was capable of providing a good estimate for solving optimization problems with the requirement of simultaneously achieving a target value and keeping the variance small; therefore, it has many applications to industrial management [22–24]. In practice, however, one could be faced with the need of determining the weights required for the resulting multi-objective optimization problem [23], or of quantifying the allowable variances for the converted single-objective one [22]. Besides, DPRA cannot provide some essential information such as a confidence level to reflect how much one can believe the generated optimal remediation strategies.

To address the abovementioned concerns, this study attempts to develop a stochastic optimization model under modeling uncertainty and parameter certainty (SOMUM). In the model, the uncertainty to be addressed is associated with the residuals generated by the introduced proxy simulators; the parameter certainty indicates that the potential errors in parameter estimations are not taken into account (investigation of both modeling and parameter uncertainties will be conducted in the ongoing study). This is a new attempt different from the previous efforts [19,20]. The previous ones focused on addressing individual uncertainty in physical properties (e.g., soil porosity), while this one aims to particularly deal with uncertainty in mathematical modeling.

The study is introduced in two parts, with each one occupying a full paper. In this first part, formulation of the model is presented, where uncertainties stemming from simulator residuals are regarded as random variables and then incorporated. It is expected that such a model can support optimal design of groundwater remediation systems under modeling uncertainty. A solution method and discussion are also provided in this part. In the second companion paper, the model will be applied to a practical petroleum-contaminated site in western Canada. Results from hypothetical tests, optimal design and models comparison will also be presented.

2. Modeling formulation

The goal of the design model is to determine optimal remediation strategies, subjected to constraints imposed by the physical nature of the problem [25]. While the objective of the model could be the minimization of total remediation cost, this model used the total pumping rate of all injection and extraction wells considering the difficulty in obtaining the unit cost for well installation, operation and maintenance [15]. The model is subject to a number of technical, water-balance and environmental constraints, and can be formulated as follows [26,27]:

$$\text{Minimize} \quad TR = \sum_{i=1}^{I+J} q_i \tag{1a}$$

$$\text{s.t.} \quad 0 \le q_i \le q_{i,\max} \text{ for all } i = I+1, I+2, \ldots, I+J \tag{1b}$$

$$\sum_{i=1}^{I} q_i = \sum_{i=I+1}^{I+J} q_i \tag{1c}$$

$$c_k(q_1, q_1, \ldots, q_{I+J}) \le MCL \text{ for all } k = 1, 2, \ldots, K \tag{1d}$$

where TR is total pumping rate for all injection/extraction wells; I and J are the number of injection and extraction wells, respectively; q_1 to q_I are decision variables, indicating the pumping rate at injection wells, respectively; q_{I+1} to q_{I+J} are decision variables, indicating the pumping rate at the extraction wells, respectively; $q_{i,\max}$ is maximum pumping rate for the ith well; c_k is contaminant concentration of well k after a period of remediation, which is computed through a three-dimensional multiphase multi-component simulator; MCL is maximum contaminant level which is determined in terms of the given environmental standard.

The objective function may also be total pumping rate, and total injection (or extraction) rate. This mainly depends on the requirement of system designers and the available data information. The sum of injection (or extraction) rates could be a good alternative in formulating this problem. However, in terms of the requirement of system designers, the optimal injection and extraction rates should be simultaneously obtained. If the total injection rate is used as the objective function, then only the optimal pumping rates at injection wells can be identified, with those at extraction wells being unknown. Similarly, the total extraction rate was not selected as the objective function.

Constraints (1b) and (1c) are technical restrictions regulating the injection and extraction pumping rates to be limited within specified practical operating intervals; the lower bound is allowed to be zero, while the upper one is determined according to technical alternatives and site characteristics. Eq. (1d) is proposed to guarantee all extracted groundwater are treated and re-injected into the aquifer [7,28]. Constraint (1e) is initiated to satisfy the environmental restriction, which requires the contaminant concentrations at all monitoring wells should be less than environmental standard. The concentrations are predicted through a three-dimensional multiphase multi-component simulator.

In the simulation, the flow of multiple (e.g., water and residual) phases, mass transfer of species between the phases, and transport of species in each of the phases should be addressed [29,30]. However, the processes can hardly be described by a conventional solute transport simulator. A three-dimensional multiphase multicomponent simulator was thus used to predict the concentrations of dissolved contaminants in the groundwater. The previous studies have demonstrated the effectiveness of this simulator in representing the complex processes involved in multiphase flow and transport of multi-components in subsurface environments [20,31]. The basic mass conservation equation for components in the subsurface can be written as [29]:

$$\frac{\partial}{\partial t}(\phi \tilde{C}_m \rho_m) + \vec{\nabla} \cdot \left[\sum_{l=1}^{n_p} \rho_m (C_{ml} \vec{u}_l - \phi S_l \vec{\bar{D}}_{ml} \cdot \vec{\nabla} C_{ml}) \right] = R_m \tag{2}$$

Table 1
Part of modeling parameters input to the simulator.

Parameter	Value	Unit
Residual water saturation	0.10	–
Residual oil saturation	0.20	–
Residual gas saturation	0.10	–
Permeability of sandy soil in *x*, *y*, and *z* direction	2900	MD
Permeability of clay till in *x*, *y*, and *z* direction	195	MD
Permeability of silty clay in *x*, *y*, and *z* direction	380	MD
Porosity of sandy soil	0.35	–
Porosity of till	0.30	–
Porosity of silty clay	0.53	–
NAPL/water interfacial tension	45	Dyne/cm
NAPL density	0.713	g/cm^3
Longitudinal dispersivity of sandy soil	5	m
Longitudinal dispersivity of clay till	5	m
Longitudinal dispersivity of silty clay	5	m
Transverse dispersivity of sandy soil	0.5	m
Transverse dispersivity of clay till	0.5	m
Transverse dispersivity of silty clay	0.5	m
Hydraulic gradient	0.003	m/m
NAPL/water partition coefficient of benzene	0.00203	–
Benzene solubility	1750	mg/L
Time step at $t = 0$	0.101	Day
Maximum time step size	100	Day
Tolerance for concentration change	0.001	–

where m is component index; l is phase index; ϕ is porosity; $\tilde{C}_m$ is overall concentration of component k (volume of component m per unit pore volume); ρ_m is density of component k [M L^{-3}]; n_p is number of phases; C_{ml} is concentration of component k in phase l (volume fraction); $\vec{u}_l$ is Darcy velocity of phase l [L T^{-1}]; S_l is saturation of phase l (volume of phase l per volume of pores); R_m is total source/sink term for component m (volume of component m per unit volume of porous media per unit time); $\tilde{C}_m$ is volume of component m summed over all phases. S_l = saturation of phase l (volume of phase l per volume of pores); R_m = total source/sink term for component m (volume of component m per unit volume of porous media per unit time) [L^3 L^{-3} T^{-1}]; $\vec{\vec{D}}_{ml}$ = dispersion tensor.

The simulator can be solved numerically with a block-centered finite difference model [31]. The solution procedures are solving the pressure equation implicitly using a Jacobi conjugate gradient solver to yield the water phase pressure in all grid blocks; using capillary pressures from the previous time step to determine the pressure of other phases in each grid block once the water phase pressure is known; determining phase velocities through the Darcy's law; yielding the concentration of each component in each grid block by explicitly solving the mass conservation equations; identifying phase concentrations and saturations through flash calculations; determine new capillary pressures from the new saturations; repeating the procedures for each time step until simulation ends. A third-order finite-difference method that can greatly reduce numerical dispersion effects is used to solve these equations. Aquifer boundaries were modeled as either constant potential surfaces or closed surfaces. Table 1 shows part of parameters input to the simulator [14,31,40]. More details regarding the full model are provided in the section of numerical simulator in the supplementary material. Other fundamental parameters (like hydraulic conductivity) was estimated by $K = \rho g k/\mu$ where k is intrinsic permeability, μ is water viscosity, K is hydraulic conductivity, ρ is water density, and g is acceleration due to gravity [41].

3. Multivariate analysis

Fig. 1 shows the flowchart of the proposed optimization method for solving model (1). The first step is numerical experiment design, which mainly includes selection and identification of the statistical samples involving explanatory and response variables. In this study, the explanatory variables are injection and extraction rates of pumping wells while the response variables are contaminant concentrations. The statistical samples were obtained by computer-assisted random sampling within the ranges of pumping rates. The second step is to create a set of proxy simulators through stepwise response surface analysis (SRSA) to capture the relationships between explanatory and response variables. The third step is to test the hypotheses of normality and zero-means for the residuals generated by proxy simulators. If the hypotheses can be statistically accepted, then a direct bridge between the simulation and optimization processes would be created; based the bridge, model (1) can be converted to model (17) which will be introduced later. Finally, this problem is transformed to an equivalent deterministic problem and then solved by nonlinear optimization solvers like Lingo 8.0. When the proxy simulators are generated through multivariate analysis, they can be represented as a set of Lingo codes (referred to as simulation module). To seek the optimal solution of the SOMUM model, an optimization module (also written as another set of Lingo codes) needs to be run to check the environmental constraints based on the outputs from the simulation module. Note that other conventional optimization algorithms than Lingo can also be used such as gradient algorithm and genetic algorithm. The following sections describe the detailed procedures of SRSA and statistical tests, as well as the formulation of the SOMUM model.

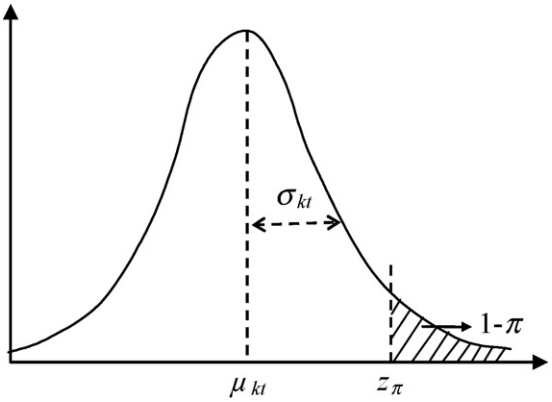

Fig. 1. Scheme of the stochastic optimization approach.

3.1. Stepwise response surface analysis

Response surface analysis (RSA) aims at empirically quantifying the relationship between response variables and explanatory variables [32,33]. RSA needs to establish two empirical models—one for the mean and one for the standard deviation [34]. The least squared fitting is a general means of determining coefficients in the two models, according to which the optimal setting is thus characterized for the explanatory variables that maximize (or minimize) the response. Consider one response variable (c_k^p) representing the contaminant concentration predicted by the proxy simulators for well k. Then c_k^p is assumed to be a polynomial function (i.e. proxy simulator) of a set of explanatory variables ($q_1, q_2, \ldots, q_{I+J}$), which can be formulated as:

$$c_k^p = a_{0,k} + \sum_{i=1}^{I+J} a_{i,k} q_i + \sum_{i=1}^{I+J} \sum_{j=1}^{I+J} a_{ij,k} q_i q_j (i \neq j) + \sum_{i=1}^{I+J} a_{ii,k} q_i^2 + e_k \quad (3)$$

where $a_{0,k}$, $\sum_{i=1}^{I+J} a_{i,k}q_i$, $\sum_{i=1}^{I+J}\sum_{j=1}^{I+J} a_{ij,k}q_iq_j(i \neq j)$, $\sum_{i=1}^{I+J} a_{ii,k}q_i^2$ and e_k are intercept, linear, interaction, quadratic and residual terms of the proxy simulator for well k, respectively. Model (3) can be used to capture the relationships between contaminant concentrations and operating conditions. Two concerns should be considered in this step: one is the increased complexity of the proxy simulator due to a portion of statistically insignificant terms being introduced needlessly; the other is the decreased accuracy of the proxy simulator caused by another portion of statistically significant terms being ignored arbitrarily. Stepwise RSA (SRSA) is a useful technique for tackling these concerns, since it supports automatic selection of models in cases where a large number of potential explanatory variables exist and no prior knowledge on which to base the selection of proxy simulators.

SRSA includes forward selection and backward elimination [22,34]. The forward selection starts with a proxy simulator without any explanatory variable (or the coefficients of all variables are zero). In the first step, the variable with the smallest p-value is added in the proxy simulator; then each following step supplements the variable that has the smallest p-value in the presence of the variables already in the proxy simulator. Variables are added one-at-a-time as long as their p-values are smaller than a given critical value (which was determined to be 0.05 in this study). In each step of forward selection, the model coefficients can be obtained through least square method. Comparatively, backward elimination begins with all explanatory variables presented in the proxy simulator. In each step, the variable with the least significant level (i.e. the largest p-value) is eliminated and the model is refitted using least square method. Each subsequent step removes the least significant variable in the proxy simulator until all remaining variables have an individual p-value larger than a provided criterion (e.g., 0.95).

In terms of the algorithm for SRSA, the p-value can be defined as the probability of the F-value greater than the F-statistic:

$$F = \frac{\left(\sum_{u=1}^{N} c_{num,u} - \bar{c}_{pro}\right)/(N' - 1)}{\left(\sum_{u=1}^{N} c_{num,u} - c_{pro,u}\right)/(N - N')} \tag{4}$$

where $c_{num,u}$ and $c_{pro,u}$ are concentrations predicted through numerical and proxy simulators, respectively, for the uth sampling; $\bar{c}_{pro}$ is the average of concentrations predicted through proxy simulators; N' is the number of sample groups; N is the number of samples. Based on the F-value, the p-value can be calculated from an F-distribution table with the degrees of freedom being $m-1$ for the numerator and $N-m$ for the denominator. In this study, N' is equal to 2 as only two groups of sampling data (one is obtained from the numerical simulator and the other one is from the proxy simulators) are used.

3.2. Tests of the residuals' normality

The uncertainty associated with proxy-simulator residuals may arise from inappropriate simulator forms, missing variables, biased parameters or sampling errors. Before incorporating the residuals into the optimization formulation, statistical analysis should be undertaken to estimate the probability distribution functions, means, and standard deviations. In the analysis, the residual of the proxy simulator for well k (e_k) is assumed to be (i) independent random variables (i.e. $\mu_k = 0$) and (ii) normally distributed with mean zero and unknown standard deviations (σ_k^2). Therefore, two hypotheses are used: one is for goodness-of-fit test of distributions, and the other is for the zero-mean hypothesis test. In this study, the Lilliefors and Jarque–Bera tests are both applied to assess whether the normality hypothesis can be accepted. The null and the alternative hypotheses for the *JB*-test are

$$H_0 : e_k \sim N(\mu_k, \sigma_k^2), \text{ and } H_1 : \text{not } H_0 \tag{5}$$

3.2.1. Lilliefors test

The Kolmogorov–Smirnov test was originally proposed to check whether a sample comes from a population with a specified distribution. This test is not appropriate for the samples with unknown population parameters such as the average and deviation. Therefore, Lilliefors [35] improved the test under the normality hypothesis when the average and deviation are not specified. The null and the alternative hypotheses for the Lilliefors test (L-test) are the same as those for the *JB*-test. This test is similar to but improves the Kolmogorov–Smirnov test by adjusting the parameters with normal distributions that are estimated from e_k rather than specified in advance [35]. Let $\bar{e}_k$ be the unbiased population parameter, μ_k, which is calculated by

$$\bar{e}_k = \sum_{u=1}^{N} \frac{e_{k,u}}{N} \tag{6}$$

where $e_{k,u}$ is the residual of the proxy simulator for well k for sample u, and N is the number of samples. Form the order statistics (from smallest to largest) $e_{k(1)}, e_{k(2)}, \ldots, e_{k(N)}$ and let s_k^2 denote the unbiased population parameter, σ_k^2, which is calculated by

$$s_k^2 = \sum_{u=1}^{N} \frac{(e_{k,u} - \bar{e}_k)^2}{N-1} \tag{7}$$

Standardize each sample with

$$e'_{k,u} = \frac{e_{k,u} - \bar{e}_k}{s_k} \tag{8}$$

Let $F(z)$ be the empirical distribution function of the population, which is equal to the number of $e'_{k,u}/N$ for every z, and $\Phi(z)$ be the standard normal cumulative distribution function. The Lilliefors test statistic (defined as L) can then be calculated with the maximum vertical distance between $F(z)$ and $\Phi(z)$, which is [36]:

$$L = \sup_z \{|F(z) - \Phi(z)|, -\infty \leq z \leq \infty\} \tag{9}$$

For a population with N samples, reject H_0 in favor of H_1 at a significance level (α) when and only when L exceeds the upper percentage point $L_{(\alpha)}(N)$. The percentile points $L_{(\alpha)}(N)$ is approximated using Monte Carlo simulation for each sample size N.

3.2.2. Jarque–Bera test

The Jarque–Bera test (*JB*-test), proposed by Bera and Jarque [37], can be used to evaluate the hypothesis that whether a random variable is normally distributed. This test is based on the difference between the skewness and kurtosis of statistical samples and those of normal distributions [38]. The skewness is calculated to determine if the distribution is symmetric and the kurtosis is used to evaluate how fat the tails of the distribution are. The *JB* statistic is given by

$$JB_k = N\left[\frac{sk_k^2}{6} + \frac{(ku_k - 3)^3}{24}\right] \tag{10}$$

where sk_k and ku_k are skewness and kurtosis of residuals of the proxy simulator for well k, respectively, which are calculated by

$$sk_k = \frac{\sum_{u=1}^{N}(e_{k,u} - \bar{e}_k)^3/N}{\sigma_k^3} \tag{11}$$

$$ku_k = \frac{\sum_{u=1}^{N}(e_{k,u} - \bar{e}_k)^4/N}{\sigma^4} \quad (12)$$

$$\sigma_k^2 = \sum_{u=1}^{N}(e_{k,u} - \bar{e}_k)^2/N \quad (13)$$

The *JB* statistic has an asymptotic distribution that follows a Chi-squared distribution ($\chi^2_{(2)}$) under the null hypothesis of normality. It should be noted that this test is merely effective for big samples, with the population size over 30.

3.2.3. The t-test

The *t*-test is used to evaluate the hypothesis that whether a sample from a normal distribution could have an average equal to a constant under unknown deviation. The null (H_0) and alternative (H_1) hypotheses for the *t*-test are

$$H_0 : \mu = \mu_k = 0, \quad \text{and } H_1 : \mu \neq \mu_k = 0 \quad (14)$$

The *t*-statistic is calculated by

$$t_k = \frac{\bar{e}_k - \mu_k}{s_k/\sqrt{N}} \sim t(N-1) \quad (15)$$

where $t(N-1)$ is statistic value under the student distribution with a degree of freedom being $N-1$. For a two-sided *t*-test under the significance level of α, the rejection region is

$$W = (-\infty, -t_{\alpha/2}(N-1)] \cup [t_{\alpha/2}(N-1), \infty) \quad (16)$$

where $t_{\alpha/2}(N-1)$ is critical value.

4. Modeling formulation under uncertainty

If the statistical tests show that the two hypotheses can be accepted, then the residuals are linked to the deterministic optimization formulation. Correspondingly, violations of environmental constraints (1b) are expressed as a probability of the contaminant concentrations exceeding the environmental standard. Based on this consideration, model (1) is formulated as the following SOMUM model:

$$\text{Minimize} \quad TR = \sum_{i=1}^{I+J} q_i \quad (17a)$$

$$\text{s.t.} \quad 0 \le q_i \le q_{i,\max} \text{ for all } i = I+1, I+2, \ldots, I+J \quad (17b)$$

$$\sum_{i=1}^{I} q_i = \sum_{i=I+1}^{I+J} q_i \quad (17c)$$

$$Pr\{c_k^p(q_1, q_1, \ldots, q_{I+J}) + e_k \le MCL\} \ge \pi \text{ for all } k = 1, 2, \ldots, K \quad (17d)$$

where *Pr* represents probability of constraint satisfaction and π is confidence level. In terms of Fig. 2, constraint (17e) can be converted to a deterministic expression as follows [39]:

$$EV[c_k^p(q_1, q_2, \ldots, q_{I+J}) + e_k] - \Phi^{-1}(\pi) \cdot \sigma_k \ge c_{\max} \quad (18)$$

where *EV* is expected value, and $\Phi^{-1}(\pi)$ is value of the standard normal cumulative distribution corresponding to a confidence level of π. Thus, the SOMUM model is transformed to its equivalent deterministic model composed of (17a)–(17d) and (18), which can be solved by Lingo 8.0. The features of the model include: although all modeling parameters are deterministic, problem (17) is stochastic because the incorporated residuals are given by probability distributions; additional outputs of the problem can be yielded as the introduced control variable (π) is used to represent the confidence level of optimal solutions. Moreover, the model has two assumptions. One is that the correlations among residuals are small enough to be ignored, which would be investigated in future studies. The other is that violations of environmental constraints due to the use of proxy simulators are allowed to some extent, indicating the risk of system failure (i.e. constraint violation) may exist (equal to $1-\pi$).

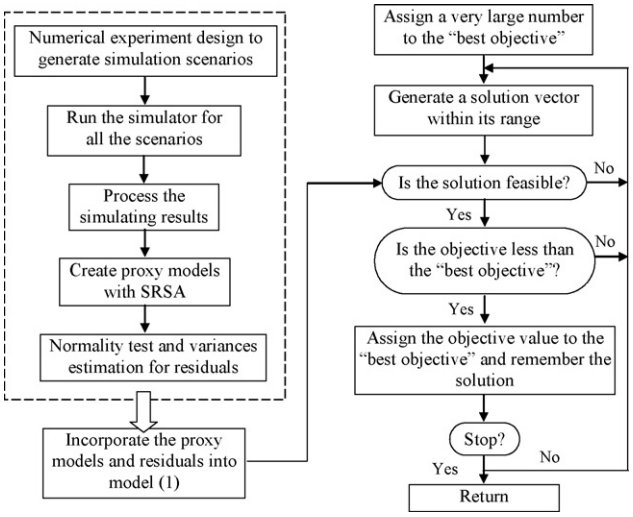

Fig. 2. Flowchart of the solution method.

5. Discussion and conclusions

Solving optimal design problems based on proxy simulators would yield optimal solutions that differ from the "true" ones of the problems due to the approximation errors. The main concern then is the approximation errors (even if they are small), which may lead to large deviations between solutions; otherwise, extra efforts need to be made for obtaining good approximations in order to arrive at solutions with high precision. If this challenge cannot be well handled, a risk of system failure could be raised, leading to a decrease in the confidence level of optimal design strategies [42]. By introducing proxy-model residuals into the groundwater management formulation, the risk of system failure originating from proxy-model residuals can be effectively controlled.

Based on this consideration, a set of proxy simulators were created by SRSA to replace the initial numerical simulator. The residuals generated by these proxy simulators are then incorporated into optimal groundwater remediation design, thus formulating a stochastic optimization problem under modeling uncertainty (arising from modeling residuals) but parameter certainty. SRSA was selected to create proxy simulators and exclude as many explanatory variables (e.g., pumping rates) as possible to mitigate computational time in the subsequent optimization process. Through this algorithm, the variables without significant contribution to modeling outputs (e.g., contaminant concentrations) are removed step by step. In general, SRSA has the advantage of supporting automatic selection of models in the cases where a large number of potential explanatory variables exist and no prior knowledge on which to base the model selection is available [18,39]. Another advantage is its capability in separately examining the effects of each linear, interactive, or quadratic term on response variables.

It should be mentioned that many other learning algorithms such as linear and piecewise (or locally) linear regression can also be used to create proxy simulators. Linear regression may cause significant errors although it is computationally efficient; in comparison, piecewise (or locally) linear regression may need high computational support though they can produce results with low errors. This study use SRSA where only linear, quadratic and interactive

terms were considered. Future studies thus need to be undertaken to investigate the prediction performance of proxy simulators with more terms (e.g., cubic terms) and the uncertainty in the resulting proxy-models.

The SOMUM model has significant distinctions from previous efforts [25]. Firstly, this model assumed that the errors of proxy simulators exist and should be accounted for; the errors are then treated as stochastic variables. In comparison, the previous depended on a hypothesis that the proxy simulators could perfectly approximate to the original simulator. Secondly, this study attempted to test the normality of residuals of proxy simulators and then incorporated the residuals into the optimization model, while the previous efforts ignored the effect of such residuals on optimal solutions. Thirdly, different algorithms were used; this one selected a type of parametric statistical method, while the previous ones used nonparametric ones. Difference of this effort from the previous ones in modeling implications will be detailed in the following companion paper.

Acknowledgements

This research was supported by the Major State Basic Research Development Program of MOST (2005CB724200 and 2006CB403307), the Canadian Water Network under the Networks of Centers of Excellence (NCE), and the Natural Science and Engineering Research Council of Canada. The authors would like to thank the editor and the anonymous reviewers for their helpful comments and suggestions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jhazmat.2009.11.060.

References

[1] NRTEE (National Round Table on the Environment and the Economy), Contaminated Site Issues in Canada: Backgrounder. http://www.nrtee-trnee.ca/Publications/HTML/BK_Contaminated-Site-Issues_E.htm, 1997.
[2] Z. Chen, Integrated environmental modeling and risk assessment under uncertainty, PhD Thesis, University of Regina, Regina, Saskatchewan, 2000.
[3] K. Soga, J.W.E. Page, T.H. Illangasekare, A review of NAPL source zone remediation efficiency and the mass flux approach, J. Hazard. Mater. 110 (2004) 13–27.
[4] B.J. Wagner, S.M. Gorelick, Optimal groundwater quality management under parameter uncertainty, Water Resour. Res. 23 (1987) 1162–1174.
[5] L.L. Rogers, F.U. Dowla, V.M. Johnson, Optimal filed-scale groundwater remediation using neural networks and the genetic algorithm, Environ. Sci. Technol. 29 (1995) 1145–1155.
[6] D.C. McKinney, M.D. Lin, Pump-and-treat groundwater remediation system optimization, J. Water Res. Pl. - ASCE 122 (1996) 128–136.
[7] F.P. Espinoza, B.S. Minsker, D.E. Golberg, Adaptive hybrid genetic algorithm for groundwater remediation design, J. Water Res. Pl. - ASCE 131 (2005) 14–24.
[8] T.B. Culver, C.A. Shoemaker, Dynamic optimal groundwater reclamation with treatment capital costs, J. Water Res. Pl. - ASCE 123 (1997) 23–29.
[9] S. Maskey, A. Jonoski, D.P. Solomatine, Groundwater remediation strategy using global optimization algorithms, J. Water Res. Pl. - ASCE 128 (2002) 431–440.
[10] A.E. Mulligan, D.P. Ahlfeld, A new interior-point boundary projection method for solving nonlinear groundwater pollution control problems, Oper. Res. 50 (2002) 636–644.
[11] A.S. Mayer, C.T. Kelley, C.T. Miller, Optimal design for problems involving flow and transport phenomena in subsurface systems, Adv. Water Resour. 25 (2002) 1233–1256.
[12] D.A. Baú, A.S. Mayer, Stochastic management of pump-and-treat strategies using surrogate functions, Adv. Water Resour. 29 (2006) 1901–1917.
[13] D.A. Baú, A.S. Mayer, Data-worth analysis for multiobjective optimal design of pump-and-treat remediation systems, Adv. Water Resour. 30 (2006) 1815–1830.
[14] L. He, G.H. Huang, H.W. Lu, Health-risk-based groundwater remediation system optimization through clusterwise linear regression, Environ. Sci. Technol. 42 (2008) 9237–9243.
[15] J.B. Guan, M.M. Aral, Optimal design of groundwater remediation systems using fuzzy set theory, Water Resour. Res. 40 (2004), doi:10.1029/2003WR002121.
[16] C. Zheng, P.P. Wang, A field demonstration of the simulation–optimization approach for remediation system design, Ground Water 40 (2002) 258–266.
[17] G.S. Cooper Jr., R.C. Peralta, J.J. Kaluarachchi, Optimizing separate phase light hydrocarbon recovery from contaminated unconfined aquifers, Adv. Water Resour. 21 (1998) 339–350.
[18] J. McPhee, W.W.-G. Yeh, Groundwater management using model reduction via empirical orthogonal functions, J. Water Res. Pl. - ASCE 134 (2008) 161–170.
[19] L. He, G.H. Huang, H.W. Lu, A simulation-based fuzzy chance-constrained programming model for optimal groundwater remediation under uncertainty, Adv. Water Resour. 31 (2008) 1622–1635.
[20] L. He, G.H. Huang, H.W. Lu, A coupled simulation–optimization approach for optimal design of groundwater remediation under uncertainty: an application to a petroleum-contaminated site in Canada, Environ. Pollut. 157 (2009) 2485–2492.
[21] G. Taguchi, Introduction to Quality Engineering: Designing Quality into Products and Processes, Kraus International Publications, White Plains, NY, 1986.
[22] E. del Castillo, S-K.S. Fan, Calculation of an optimal region of operation for dual response systems fitted from experimental data, J. Oper. Res. Soc. 50 (1999) 826–836.
[23] R. Ding, D.K.J. Lin, D. Wei, Dual-response surface optimization: a weighted MSE approach, Qual. Eng. 16 (2004) 377–385.
[24] G. Miró-Quesada, E. del Catillo, Two approaches for improving the dual response method in robust parameter design, J. Qual. Technol. 36 (2004) 154–168.
[25] W.F. Ramirez, Application of Optimal Control Theory to Enhanced Oil Recovery, Elsevier Science Publishing Company Inc, York, NY, USA, 1987.
[26] S.M. Gorelick, C.I. Voss, P.E. Gill, W. Murray, M.A. Saunders, M.H. Wright, Aquifer reclamation design: the use of contaminant transport simulation coupled with nonlinear programming, Water Resour. Res. 20 (1984) 415–427.
[27] D.P. Ahlfeld, J.M. Mulvey, G.F. Pinder, E.F. Wood, Contaminated groundwater remediation design using simulation, optimization, and sensitivity theory. 2. Analysis of a field site, Water Resour. Res. 24 (1988) 443–452.
[28] S.Y. Yan, B. Minsker, Optimal groundwater remediation design using an adaptive neural network genetic algorithm, Water Resour. Res. 42 (2006), 10.1029/2005WR004303.
[29] M. Delshad, G.A. Pope, K. Sepehrnoori, A compositional simulator for modeling surfactant enhanced aquifer remediation. 1. Formulation, J. Contam. Hydrol. 23 (1996) 303–327.
[30] P.C. de Blanc, Development and demonstration of a biodegradation model for non-aqueous phase liquids in groundwater, Ph.D. Dissertation, The University of Texas at Austin, USA, 1998.
[31] J.B. Li, Development of an inexact environmental modeling system for the management of petroleum-contaminated sites, Ph.D. Thesis, Faculty of Engineering, University of Regina, Regina, Saskatchewan, Canada, 2003.
[32] M. Tir, N. Moulai-Mostefa, Optimization of oil removal from oily wastewater by electrocoagulation using response surface method, J. Hazard. Mater. 158 (2008) 107–115.
[33] B.H. Hameed, I.A.W. Tan, A.L. Ahmad, Preparation of oil palm empty fruit bunch-based activated carbon for removal of 2,4,6-trichlorophenol: optimization using response surface methodology, J. Hazard. Mater. 164 (2009) 1316–1324.
[34] R. Ding, D.K.J. Kin, D. Wei, Dual-response surface optimization: a weighted MSE approach, Qual. Technol. 34 (2002) 437–447.
[35] H. Lilliefors, On the Kolmogorov-Smirnov test for normality with mean and variance unknown, J. Am. Stat. Assoc. 62 (1967) 399–402.
[36] M.E. Calzada, S.M. Scariano, Visual EDF software to check the normality assumption, in: P. Bogacki, E.D. Fife, L. Husch (Eds.), The Electronic Proceedings of the 15th International Conference on Technology in Collegiate Mathematics, From the 2002 meeting in Orlando, FL, 2004.
[37] A. Bera, C. Jarque, Efficient tests for normality, heteroscedasticity, and serial independence of regression residuals, Econ. Lett. 6 (1980) 255–259.
[38] L.B. Dong, D.E.A. Giles, An empirical likelihood ratio test for normality, in: Econometrics Working Paper EWP0401 (1485-6441), Department of Economics, University of Victoria, Canada, 2004.
[39] C. Tiedeman, S.M. Gorelick, Analysis of uncertainty in optimal groundwater contaminant capture design, Water Resour. Res. 29 (1993) 2139–2153.
[40] Energy Environment Program (EEP), Numerical simulation for contaminant flow and transport in subsurface—a study of soil and groundwater contamination at the Coleville Site, in: Process Report, University of Regina, Regina, Saskatchewan, Canada, 2005.
[41] L. He, G.H. Huang, H.W. Lu, G.M. Zeng, Optimization of surfactant-enhanced aquifer remediation for a laboratory BTEX system under parameter uncertainty, Environ. Sci. Technol. 42 (2008) 2009–2014.
[42] H.W. Lu, G.H. Huang, Y.P. Lin, et al., A two-step infinite α-cuts fuzzy linear programming method in determination of optimal allocation strategies in agricultural irrigation systems, Water Resour. Manag. 23 (2009) 2249–2269.