Efficient Combinatorial Optimization under Uncertainty. 2. Application to Stochastic Solvent Selection

Ki-Joo Kim and Urmila M. Diwekar*

CUSTOM (Center for Uncertain Systems: Tools for Optimization and Management) and Civil and Environmental Engineering, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213

Solvent selection is an important step in process synthesis, design, or process modification. The computer-aided molecular design (CAMD) approach, based on the reverse use of group contribution methods, provides a promising tool for solvent selection. However, uncertainties inherent in these techniques and associated models are often neglected. This paper, part 2 of the series, presents a new approach to solvent selection under uncertainty using the Hammersley stochastic annealing (HSTA) algorithm. A real world case study of acetic acid extraction from water, based on two stochastic CAMD models, namely, the infinite dilution activity coefficient model and the solubility parameter model, is presented. This example illustrates the importance of uncertainty in CAMD and demonstrates the usefulness of this HSTA approach to obtain robust decisions.

1. Introduction

Solvents are extensively used as process materials (e.g., extracting agent) or process fluids (e.g., chlorofluorocarbons, CFC) in chemical processing industries, pharmaceutical industries, and solvent-based industries (such as coating and painting). Because waste solvents are a main source of pollution to air, water, and soil, it is desirable to use reduced amounts of solvents and/or to use environmentally friendly solvents without sacrificing process performance. Further, there are some solvents that must be eliminated because of environmental health effects and regulatory requirements. For example, the 1987 Montreal protocol bans many chlorinated solvents.

Solvent selection, an approach used to generate candidate solvents with desirable properties, can help to handle these problems. Several methodologies have been developed for solvent selection over the years.[1] The first approach uses traditional laboratory synthesis and test methodology to find promising solvents. This method can provide reliable and accurate results but, in many cases, is limited by cost, time, and safety constraints. The second approach is to screen the property database. It is the most common and simple method, but this method is restricted by the size and accuracy of the database. Although these two approaches have been widely used, they may not provide the best solvent because of the sheer number of solvent molecules to be tested or searched. Finally, computer-aided molecular design (CAMD), based on the reverse use of group contribution methods (GCMs), can automatically generate promising solvents from their fundamental building blocks.[2,3] The GCM is a forward problem: if we know a molecule, we can estimate its physical, chemical, biological, and health-effect properties based on its groups or building blocks. In contrast, CAMD is a backward problem: if we know desired properties or regulation limits, we can find molecules that satisfy these properties or limits by combining groups. CAMD can also be applied to CFC substituents,[3,4] solvent blend designs,[5,6] polymer and drug designs,[7,8] and alternative process fluid designs.[9]

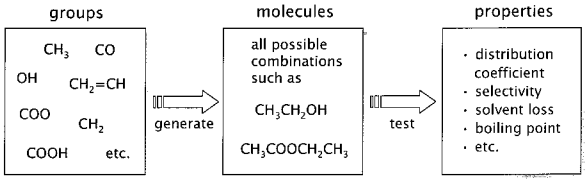

Figure 1. Basic diagram of CAMD based on GCMs.

A basic diagram of CAMD is shown in Figure 1, where there is a set of groups as a starting point. These groups are uniquely designed to generate all possible molecules by exploring all combinations. The properties of each group and/or the interaction parameters between groups can be theoretically calculated, experimentally obtained, or statistically regressed. From this set of groups, solvent molecules can be generated by group combinations. Once molecules are generated, the properties of the molecules are predicted based on the properties of their groups in order to determine whether they satisfy the specified criteria. This method can generate a list of candidate solvents with reasonable accuracy within a moderate time scale. However, CAMD is limited by the availability and reliability of property estimation methods.

There are three main CAMD approaches: generation-and-test, mathematical optimization, and combinatorial optimization approaches. The generation-and-test approach has been widely used since the 1980s.[2,3,10−12] This method usually implements heuristics or knowledge-based methods to combat combinatorial explosion of group combinations. This method, however, cannot guarantee the quality of the results, i.e., the best solvent for a given system. The mathematical optimization approach, which includes MINLP (mixed-integer nonlinear programming)[5,7,13,14] and MILP (mixed-integer linear programming)[8,9] problems, has been applied to the nonlinear or linear structure−property correlations. Even though this method can provide a great deal of

* To whom correspondence should be addressed. E-mail: urmila@cmu.edu. Phone: +1 412 268 3003. Fax: +1 412 268 7813.

10.1021/ie0101691 CCC: $22.00
Published on Web 02/01/2002

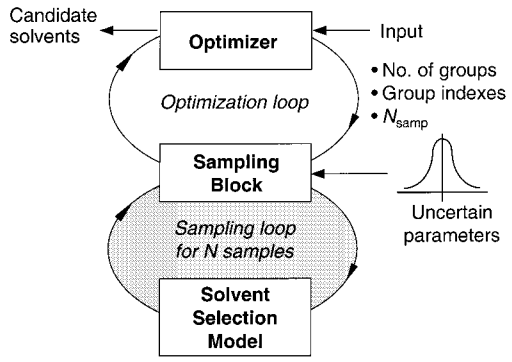

Figure 2. Stochastic optimization framework for solvent selection under uncertainty.

flexibility in problem formulation, the exact expression of the structure−property correlations may be very difficult or even impossible.[15,16] Further, the mentioned two approaches have difficulty deriving heuristics and knowledge-based constraints, dealing with combinatorial complexities, expressing the nonlinear structure−property correlations, and providing the global solvent. Thus, the third approach, combinatorial optimization such as genetic algorithm[15] and simulated annealing (SA),[4,16,17] is implemented in CAMD and can easily handle combinatorial complexity of molecular design.

All methodologies for solvent selection are exposed to uncertainties that arise from experimental errors, imperfect theories or models and their parameters, improper knowledge, or ignorance of systems. In addition, available group parameters may not be present, and current GCMs cannot estimate all necessary properties. Uncertainties can affect the real implementation of selected solvents; however, only a few papers have focused on uncertainties in CAMD. Maranas[18] presented a methodology for optimal polymer design under uncertainty, in which a chance-constrained MINLP problem is converted to a deterministic equivalent MINLP problem. Uncertainties of property prediction in this method should be *stable* distributions (e.g., normal distribution). Uncertainty on the optimizer, not the CAMD model, was described by Sinha et al.,[19] in which they presented a computationally efficient branch and bound algorithm for CAMD. Kim et al.[17] developed a new and efficient optimization approach, based on SA, for a solvent selection problem under uncertainty. Uncertainties on the solubility parameter were explicitly expressed with the probability distribution type, mean, and variance.

This paper, as an extension of the previous work, is a comprehensive paper to deal with uncertainties associated with the property estimation methods, which are key elements in solvent selection problems. Because it is emphasized that uncertainties are crucial in solvent selection problems, uncertainties in the property estimation methods should be identified and quantified.

Optimization under uncertainty, commonly known as *stochastic optimization* or *stochastic programming*, is then applied to solvent selection under uncertainty problems. Figure 2 shows a simple representation of this approach. This approach involves two recursive loops: the outer optimization loop and the inner sampling loop. The outer optimizer not only determines the discrete decision variables, such as the number of groups in a molecule and the group indices, but also determines the number of uncertain samples needed for the inner sampling loop. The inner sampling loop essentially converts the deterministic group contribution model into the stochastic model. Thus, the stochastic optimization framework involves four steps: (a) identifying and specifying key input parameter uncertainties in terms of probabilistic distribution functions, (b) sampling these distributions in an iterative fashion, (c) propagating the effects of these uncertainties through the model, and (d) analyzing the output using statistical techniques. Here in this paper we are using the new and efficient Hammersley stochastic annealing (HSTA) algorithm, described in part 1, which is designed to efficiently handle combinatorial optimization problems under uncertainty. Because the properties of the individual groups are inherently uncertain and there are a large number of molecular combinations of the groups, the HSTA algorithm can significantly improve the computational efficiency of the stochastic optimization algorithm.

In this study, two CAMD models, based on the infinite-dilution activity coefficient (γ^∞) and the solubility parameter (δ)[10,19] for solvent selection under uncertainty, are presented and applied to generate greener solvents for acetic acid extraction from water. Section 2 describes the deterministic solvent selection model based on γ^∞ and compares the results with other literature results in order to verify the new approach. Sections 3 and 4 present the stochastic solvent selection models based on γ^∞ and δ, respectively. Section 5 discusses the results from the two stochastic solvent selection models and is followed by conclusions.

2. Case I: Deterministic CAMD Based on γ^∞

This section presents the deterministic solvent selection model for acetic acid extraction from water, which is one of the widely studied applications in CAMD, and compares the results with other literature results.[3,12,13,20]

2.1. Solvent Selection Model. To replace the current solvent or to design a new one, there are several criteria to be considered, such as (a) distribution coefficient (m), (b) solvent selectivity (β), (c) solvent loss (S_L), (d) physical properties such as boiling point, flash point, density, and viscosity, (f) toxicology, (g) environmental properties such as LC_{50} (lethal concentration at 50% mortality), LD_{50} (lethal dose at 50% mortality), BCF (bioconcentration factor), and persistence, and (h) cost. For extraction processes, the final selection of solvents will generally be dominated by m and β. The distribution coefficient (m), a measure of solvent capacity, is the most important factor and represents the solute distribution between the solvent and the raffinate phases, as shown in the following equation:

$$m = \frac{x_{B,S}}{x_{B,A}}\frac{MW_A}{MW_S} \simeq \frac{\gamma^\infty_{B,A}}{\gamma^\infty_{B,S}}\frac{MW_A}{MW_S} \qquad (1)$$

where the symbols A, B, and S represent the raffinate, solute, and solvent phases, respectively. MW is molecular weight, and x is mole fraction. A high value of m reduces the size of the extracting equipment and the amount of recycling solvent. Because this parameter neglects the solubility of the raffinate A in the solvent phase S, this approximation can be a good estimation of the m *only* at low concentrations of solute B.

Solvent selectivity (β), equivalent to the relative volatility in distillation, is the ratio between the distri-

Table 1. m, β, and S_L of Typical Solvents for Acetic Acid Extraction from Water

solvent	m	β	S_L
ethyl acetate	0.3156	14.63	0.0560
isopropyl acetate	0.2027	16.66	0.0226
isoamyl acetate	0.1950	20.20	0.0034

bution coefficients of solute and raffinate and is defined by

$$\beta = \frac{m_B}{m_A} \simeq \frac{\gamma^{\infty}_{A,S}}{\gamma^{\infty}_{B,S}} \frac{MW_B}{MW_A} \tag{2}$$

Solvent selectivity estimates the ability of the solvent to selectively dissolve a solute, and a high β value thus reduces the cost of solute recovery. Another important criterion is solvent loss (S_L), which is expressed by the following equation:

$$S_L = \frac{1}{\gamma^{\infty}_{S,A}} \frac{MW_S}{MW_A} \tag{3}$$

Low S_L means high selectivity toward solute and determines the immiscibility between solvent and raffinate.

Acetic acid is commonly used as a process solvent or is produced as a byproduct. Because acetic acid can be a pollutant as well as a valuable solvent, it is desirable to minimize the discharge of acetic acid to the environment. To recycle or remove acetic acid from waste process streams or units, extraction is commonly applied. For the extraction process, one can either use high-boiling solvents[12,20] or low-boiling solvents[3] depending on the process considered. Ethyl acetate, isoamyl acetate, and isopropyl acetate are widely used in industries to extract acetic acid. Table 1 shows m, β, and S_L values of some typical commercial solvents for this extraction. The fifth-revised original UNIFAC equation[21] is used to estimate infinite-dilution activity coefficients. All three solvents in this table have high β values and their typical lower limit of liquid−liquid extraction systems is 7. Ethyl acetate, which is one of the common solvents for acetic acid extraction, has high m, but it unfortunately also has high S_L. This mandates process engineers to use more powerful and greener solvents.

In this section, high-boiling solvents are generated by the efficient simulated annealing (ESA) algorithm that is described in part 1 of this series and compared with other literature results. High-boiling solvents can be easily separated from the extract stream and then recycled to extraction equipment. The deterministic solvent selection model based on γ^{∞} is given below:

$$\min_{N_1, N_2^{(i)}} -m \tag{4}$$

s.t.

$$\beta \geq 7$$

$$S_L \leq 0.01$$

$$148 \leq T_{BP}\,(°C) \leq 268$$

$$2 \leq N_1 \leq 10$$

$$1 \leq N_2^{(i)} \leq 24, \quad \forall\, i \in N_1$$

Table 2. Set of Discrete Decision Variables for UNIFAC GCM[a]

i	$N_2^{(i)}$	i	$N_2^{(i)}$	i	$N_2^{(i)}$	i	$N_2^{(i)}$
1	CH_3-	7	$CH_2{=}C<$	13	CH_3CO-	19	CH_3O-
2	$-CH_2-$	8	$-CH{=}C<$	14	$-CH_2CO-$	20	$-CH_2O-$
3	$-CH<$	9	$>C{=}C<$	15	$-CHO$	21	$>CH-O-$
4	$>C<$	10	$-OH$	16	CH_3COO-	22	$-COOH$
5	$CH_2{=}CH-$	11	CH_3OH	17	$-CH_2COO-$	23	HCOOH
6	$-CH{=}CH-$	12	H_2O	18	HCOO−	24	$-COO-$

[a] −, >, and < represent the connecting nodes.

In this problem, the discrete decision variables are the number of groups (N_1) in a solvent molecule and the group index ($N_2^{(i)}$, $i = 1, \ldots, N_1$) of that molecule.

This group index can then build a unique solvent molecule which represents a *configuration* in the ESA algorithm. For structural feasibility of the configuration, the octet rule relates the total number of free attachments of groups with the number of groups. For acyclic groups in this paper, the octet rule becomes

$$\sum_{i}^{N_1} b_i = 2(N_1 - 1) \tag{5}$$

where b_i is the number of free attachments in a group index i. Some additional constraints such as the maximum number of branches in a group and the total number of functional groups in a molecule can be applied to remove nonplausible molecules.

The UNIFAC model then is used to predict the objective function (i.e., m) and other solvent selection criteria such as β and S_L. The bounds on the constraints in this optimization problem are mainly taken from Pretel et al.[12] To estimate boiling points (T_{BP}), the boiling point prediction GCM (eq 6) has t_a and t_b as parameters:[22]

$$T_{BP} = \sum_{i=1}^{N_1} t_a(N_2^{(i)}) + t_b \text{ (units: °C)} \tag{6}$$

To move to a new configuration (i.e., group combination) from the current configuration in the ESTA algorithm (see step 2.1 in Table 3 in part 1 of this series), there are three processes used: addition, contraction, and random bump. In the addition process ($N_1 = N_1 + 1$), the number of groups (N_1) in a solvent molecule is increased and a random group index is assigned to that increased group. In the contraction process ($N_1 = N_1 - 1$), one group is randomly deleted. In random bump ($N_1 = N_1$), the number of groups in a molecule is unchanged. Instead, an arbitrarily selected group index ($N_2^{(i)}$) is randomly bumped up or down. The magnitudes of these bumps are also random. The probabilities for these three processes are specified at 30%, 30%, and 40%, respectively. Large random bump probability is guaranteed to span all of the group indexes. Besides these basic probabilities, there are also several probabilities for configurational moves.

The group index used in this case study is summarized in Table 2. Because the total number of groups is 24 and a maximum of 10 groups per molecule is allowed, the total combinatorial space is 24^{10} (6.34×10^{13}) combinations. This problem represents the deterministic IP (integer programming) problem and can be efficiently solved by the ESA algorithm.

Table 3. High-Boiling Candidate Solvents (Deterministic Case)

no.	molecules	m	β	typical solvent
1	CH_3, $6CH_2$, OH	0.6074	19.77	1-heptanol
2	$2CH_3$, $4CH_2$, CH, OH	0.6073	19.78	2-heptanol
3	CH_3, $7CH_2$, OH	0.5478	21.73	1-octanol
4	$2CH_3$, $5CH_2$, CH, OH	0.5477	21.74	3-octanol
5	$2CH_3$, $4CH_2$, CH, HCOO	0.5331	30.80	isoheptyl formate
6	CH_3, $8CH_2$, OH	0.4998	23.66	1-nonanol
7	$2CH_3$, $6CH_2$, CH, OH	0.4998	23.68	3-nonanol
8	$2CH_3$, $5CH_2$, CH, OH	0.4698	32.77	isooctyl formate
9	$2CH_3$, $7CH_2$, CH, OH	0.4603	25.60	2-decanol
10	$2CH_3$, $6CH_2$, CH, HCOO	0.4194	34.65	isononyl formate
11	CH_3, $5CH_2$, CH_3CO	0.3765	47.68	2-octanone
12	$2CH_3$, $3CH_2$, CH, CH_3CO	0.3764	47.70	3-methyl-2-heptanone
13	CH_3, $6CH_2$, COOH	0.3599	12.04	octanoic acid
14	$2CH_3$, $4CH_2$, CH, COOH	0.3599	12.05	2-ethylhexanoic acid
15	CH_3, $6CH_2$, CH_3CO	0.3361	52.93	2-nonanone
16	$2CH_3$, $4CH_2$, CH, CH_3CO	0.3361	52.96	methyl 4-propylbutyl ketone
17	$2CH_3$, $4CH_2$, CH_2CO	0.3167	61.53	3-octanone
18	$3CH_3$, $2CH_2$, CH, CH_2CO	0.3166	61.57	butyl isopropyl ketone
19	$4CH_3$, 2CH, CH_2CO	0.3166	61.60	isopropyl isobutyl ketone
20	CH_3, $7CH_2$, CH_3CO	0.3032	58.01	2-decanone

Table 4. Candidate Solvents from Literature

reference	solvents	m	method
Odele and Macchietto[13]	$5CH_3$, $2CH_2$, 3CH	0.81	MINLP
Pretel et al.[12] and Hostrup et al.[20]	1-nonanol	0.50	generation-and-test
	decanoic acid	0.29	
	5-decanone	0.28	
	diisobutyl ketone	0.28	

2.2. Results and Discussion. The ESA algorithm generated a set of top 40 solvents which have high m and satisfy given constraints. The first top 20 solvents are summarized in Table 3. Most of the highly promising candidate solvents in this table are alcohols, and some of the highly promising solvents are formates. Ketones are also highly ranked candidate solvents.

These results generally coincide with other literature results[12,13,20] but provide an optimized set of candidate solvents. Table 4 shows some of literature results for acetic acid extraction. The solvent generated by the MINLP approach[13] has a very high m as compared to other results included in this paper. However, the proposed solvent is a highly branched alkane, and thus its real application may be suspicious because of its availability. The chemical feasibility constraints in this study prohibit the generation of highly branched chemicals. Pretel et al.[12] and Hostrup et al.[20] have similar results even though constraints are slightly different. The candidate solvents generated by the generation-and-test method in the literature are easily available, but the distribution coefficients are relatively low (except 1-nonanol) compared to the result in this paper. Note that the m of the 20th solvent in Table 3 of this study is 0.30, which is higher than those of the typical solvents in Table 4. Thus, only 1-nonanol with an m of 0.50 can appear in the top 20 solvents.

From these two tables, it can be seen that the ESA algorithm can generate a *set* of candidate solvents with higher accuracy (i.e., better m). For actual implementation, these solvents need to be further investigated in areas such as uncertainty, cost, safety, and environment factors.

Table 5. Errors in γ^∞ Estimations

reference	family[a]	model	N_{data}	mean dev (%)
Gmehling[26]	total	UNIFAC (original)	1773	21.1
		UNIFAC (Dortmund)	1773	5.7
	alkanes/alkanes	UNIFAC (Dortmund)	374	24.9
	alkanes/alcohols	UNIFAC (Dortmund)	216	30.5
	alcohols/alcohols	UNIFAC (Dortmund)	18	6.6
Bastos et al.[23]	total	UNIFAC (original)	4860	28.1
		UNIFAC−γ^∞	8300	20.2
	organic/water	UNIFAC−γ^∞	65	429
Zhang et al.[25]	organic/water	UNIFAC (original)	416	53.2
		UNIFAC (Dortmund)	417	52.5
		UNIFAC−γ^∞	107	72.2
		UNIFAC (improved)	132	15.3

[a] Family A/B means solute A and solvent B.

3. Case II: Stochastic CAMD Based on γ^∞

3.1. Property Prediction Errors. The (original) UNIFAC equation has three terms: the surface area (R_k) and volume (Q_k) of each group k and the interaction parameters (a_{mn}) between groups m and n. The surface area and volume terms of each group are constants because they are calculated from atomic and molecular structure data. However, the interaction parameters are obtained from the regression of experimental data and thus are subject to uncertainty due to experimental and regression errors. Further, activity coefficients (γ_i) at a finite condition (e.g., $C_i = 0.000\,01$) are, by definition, extrapolated to infinite-dilution activity coefficients ($\gamma_i^\infty = \lim_{C_i \to 0} \gamma_i$) in which large discrepancies between experimental and calculated values can be observed.

The estimation errors in γ^∞ are usually expressed as the mean deviation, which is defined as

$$\text{mean deviation (\%)} = \frac{1}{N_{data}} \sum^{N_{data}} \frac{|\gamma^\infty_{exp} - \gamma^\infty_{cal}|}{\gamma^\infty_{exp}} \times 100 \quad (7)$$

Several UNIFAC equations have been developed to improve the accuracy of γ^∞ estimation. UNIFAC−γ^∞ [23] is specially designed for the estimation of γ^∞. It is based on the original UNIFAC equation, but the group interaction parameters are obtained by fitting the experimental γ^∞ data. Dortmund UNIFAC[24] is a modified UNIFAC model in which the temperature-dependent interaction parameters are used. In this model, the surface area (R_k) and volume (Q_k) are also adjustable parameters from experimental data, not from the atomic and molecular structure data. An *improved* UNIFAC,[25] based on the modified Dortmund UNIFAC, has additional mixture-type groups to account for the special hydrophobic effects in the organic/water systems.

The mean deviations of various UNIFAC equations are summarized in Table 5. Generally speaking, all of the models in this table provide relatively poor γ^∞ estimations, whose errors span 15−429% (the value 5.7% in the Dortmund UNIFAC model in the original

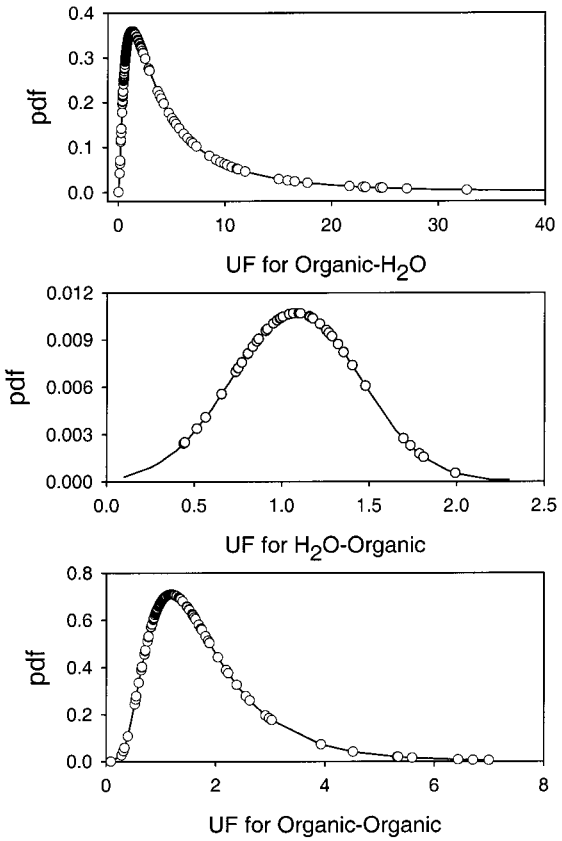

Figure 3. pdf's of UFs for organic/water, water/organic, and organic/organic families.

source[26] is questionable). Even though several variants of the UNIFAC equation are used for analysis, the results are not as accurate as expected. Because this table shows large errors in the γ^{∞} estimation, one has to incorporate uncertainties in solvent selection models.

3.2. Uncertainty Identification and Quantification. To handle uncertainties in the solvent selection models, uncertainty identification and quantification are necessary. Uncertainty quantification involves finding the type of probability distribution and quantifying the first and second moments (i.e., mean and variance). Ideally, all uncertainties on a_{nm} can be explicitly expressed. However, because the number of interaction parameters (i.e., n^2) increases with the number of (main) groups in the CAMD formulation, it may not be sound to derive all n^2 uncertainty information. Furthermore, a_{nm} is used in the fundamental or lower level of the UNIFAC equation, and hence, imposing n^2 uncertainty information will be highly complicated and computationally expensive.

To avoid these problems, uncertainty information is derived for γ^{∞} and not for the group interaction parameters. In addition, to identify or characterize uncertainties in γ^{∞}, γ^{∞} can be divided into three categories based on its type of family (γ^{∞} of organic/water, water/organic, and organic/organic families) because the properties of water are quite different from those of organic chemicals. A new uncertainty quantification term called the uncertainty factor (UF) is introduced, which is the ratio of the experimental γ^{∞} to the calculated γ^{∞}. UF, defined in eq 8, shows how much the calculated γ^{∞} is deviated from the true γ^{∞}.

$$\mathrm{UF} = \gamma^{\infty}_{\mathrm{exp}}/\gamma^{\infty}_{\mathrm{cal}} \tag{8}$$

Note that a UF of 1.0 means the calculated value is exactly equal to the experimental value.

For the organic/water family, a total of 227 binary γ^{∞} data obtained from Gmehling and co-workers[27,28] are used. The total numbers of data for the water/organic and organic/organic families are 41 and 161, respectively. Figure 3 shows the probability density functions (pdf) of the UFs of $\gamma^{\infty}_{\mathrm{organic,water}}$, $\gamma^{\infty}_{\mathrm{water,organic}}$, and $\gamma^{\infty}_{\mathrm{organic,organic}}$. For the ogranic/water family, the type of distribution is a log-normal distribution with an arithmetic mean of 2.92 in UF and a standard deviation of 5.94 in UF. From this pdf, one can expect that uncertainties exert a great impact on the γ^{∞} because of their large mean and wide standard deviation in UF. It is also found that the pdf's of UF of the other two infinite-dilution activity coefficients ($\gamma^{\infty}_{\mathrm{water,organic}}$ and $\gamma^{\infty}_{\mathrm{organic,organic}}$) are normally distributed ($N(1.08, 0.37)$) and log-normally distributed (log $N(1.42, 1.14)$), respectively. The values of $\gamma^{\infty}_{\mathrm{water,organic}}$ tend to be less affected by uncertainties because nonideality caused by water is small in this family. It is interesting to see that all of the distributions are shifted to the right or are positively skewed. Therefore, uncertainties not only perturb γ^{∞} but also increase the output values of γ^{∞}.

3.3. Stochastic Solvent Selection Model. In this stochastic solvent selection problem (and case III), we are interested in low-boiling solvents because one of the main interests is to replace the current solvent, ethyl acetate, for acetic acid extraction. The discrete stochastic optimization problem for generating a set of candidate solvents can be formulated as follows:

$$\min_{N_1, N_2^{(i)}} -\frac{1}{N_{\mathrm{samp}}} \sum_{j=1}^{N_{\mathrm{samp}}} \left[\frac{\xi_1^j \gamma^{\infty}_{B,A}}{\xi_3^j \gamma^{\infty}_{B,S}}\right] \frac{\mathrm{MW_A}}{\mathrm{MW_S}} \tag{9}$$

s.t.

$\xi_1 \sim \log N(2.92, 5.94)$ for the organic/water family

$\xi_2 \sim N(1.08, 0.37)$ for the water/organic family

$\xi_3 \sim \log N(1.42, 1.14)$ for the organic/organic family

$$\beta = \frac{1}{N_{\mathrm{samp}}} \sum_{j=1}^{N_{\mathrm{samp}}} \left[\frac{\xi_2^j \gamma^{\infty}_{\mathrm{A,S}}}{\xi_3^j \gamma^{\infty}_{\mathrm{B,S}}}\right] \frac{\mathrm{MW_B}}{\mathrm{MW_A}} \geq 7$$

$$S_{\mathrm{L}} = \frac{1}{N_{\mathrm{samp}}} \sum_{j=1}^{N_{\mathrm{samp}}} \left[\frac{1}{\xi_1^j \gamma^{\infty}_{\mathrm{S,A}}}\right] \frac{\mathrm{MW_S}}{\mathrm{MW_A}} \leq 0.058$$

$$47 \leq T_{\mathrm{BP}}\,(^{\circ}\mathrm{C}) \leq 108$$

$$2 \leq N_1 \leq 10$$

$$1 \leq N_2^{(i)} \leq 24, \quad \forall\, i \in N_1$$

where ξ_i is an uncertain parameter of UF_i and is imposed on the output γ^{∞}. Discrete decision variables are the number of groups (N_1) in a solvent molecule and the group index ($N_2^{(i)}$, $i = 1, \ldots, N_1$) of that molecule that can build a unique solvent molecule.

To reduce the computational burden and guarantee best candidate solvents, the HSTA is implemented,[29] as shown in Figure 2. The optimizer in the upper loop

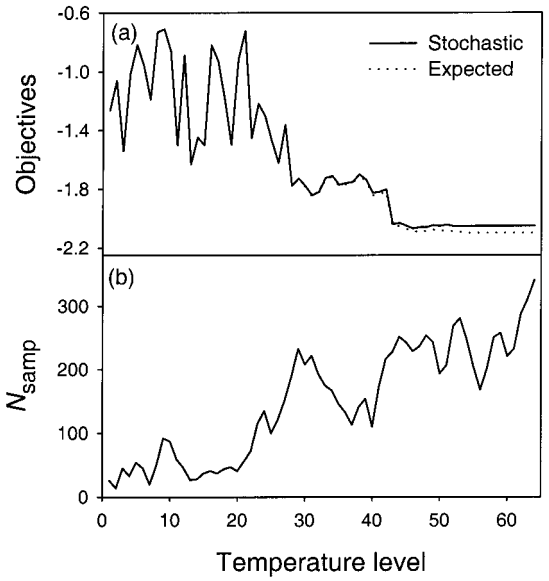

Figure 4. Changes of the objective values and the number of samples with respect to the annealing temperature level.

determines the number of groups (N_1) in a solvent molecule, group indexes ($N_2^{(i)}$, $i = 1, ..., N_1$) that tell which group is present in a proposed solvent, and the number of uncertain samples (N_{samp}). The information about the distribution functions, expressed as 0.1% and 99.9% quantiles, is supplied to the inner sampling loop. This sampling loop then uses an efficient sampling method[30] to generate the uncertain samples. Each sample is propagated through the model based on the GCM to evaluate the expected values for the distribution coefficient, solvent selectivity, and boiling points of each component. This information is transferred to the optimization loop as the objective function and constraints. The optimizer determines whether the probabilistic results of the given molecule are optimal.

3.4. Results and Discussion. Figure 4a shows the progress of the HSTA algorithm. Two objective values are plotted with respect to the annealing temperature level: the true (expected) objective and the objective with a penalty term. At the beginning of annealing, exploring the configuration surface in order to locate local optima is more important than the solution accuracy of these local optima. Therefore, the two objectives are very close because of a small weighing function $b(t)$ that results in a small penalty term. As annealing proceeds, solution accuracy becomes more important than solution efficiency and $b(t)$ increases very rapidly because it is not precise. To maintain or enhance the solution accuracy by reducing the penalty, the trend is to increase the number of samples N_{samp}, as shown in Figure 4b.

Table 6 shows the optimal solvent candidates for the stochastic (HSTA) and deterministic (STA) cases. Among 40 candidate solvents generated in this study, the first top 20 solvents of both cases are ranked with respect to the order of m, and only 8 solvents appear in both cases (see bold solvents in this table). This implies that the deterministic case does not generate several or many promising solvents, which appeared in the stochastic case. As expected, distribution coefficients for the stochastic case are greater than those for the deterministic case because of the positively skewed uncertainty factor, mainly $\gamma^{\infty}_{organic,water}$, and because increased m values become closer to the experimental values.

Different sets of solvents in Table 6 require further screening of solvents in terms of design, safety, health, and environmental constraints. Solvent availability, toxicity evaluation, and safety consideration are some valuable criteria that may be used. Finally, experimental verifications should be followed. After screening, if the first top three solvents in Table 6 are not found to be useful, then different candidate solvents between the stochastic and deterministic cases can lead to big differences in the whole solvent selection process. Thus, it is clear from this result that the implementation of CAMD without considering uncertainties may fail to find the best solution.

Uncertainty factors also affect the constraints, β and S_L in eq 8. Figure 5a shows probability density functions of β and S_L at both cases. It is observed that the probability density of the stochastic β is lower at low β values and higher at high β values than the probability density of the deterministic β. This means that the uncertainty factor positively affects β. However, the probability density of the stochastic S_L in Figure 5b is higher toward the S_L constraint limit. This results in a large solvent loss of the stochastic solvents ranked 1, 3, 7, 8, and 13 in Table 6. Thus, in this case study, the S_L constraint is more tight than the β constraint.

The value of stochastic solution (VSS)[31] can be used to quantify the effects of uncertainty. The difference between taking the average value of the uncertain variable as the solution and using the stochastic analysis is defined as VSS. Thus, VSS represents the loss by not considering the uncertainties. Because the uncertainty factors, represented as ξ over γ^{∞}, are implemented in the objective function, we can assume that the expected value of the stochastic problem with the average ξ be 2.06 (=2.92/1.42) times the deterministic m shown in Table 6. For the first set of solvent molecules in this table, the VSS of this case study is estimated as 1.16 (=2.95 − 2.06 × 0.87), and the stochastic optimization, therefore, increases the performance (distribution coefficient in this study) by 65%. Other sets of solvent molecules have a similar VSS.

Further, we can also observe differences in the probability density function of the solvents. Figure 6 shows the pdf's of distribution coefficients of the top 40 solvents for each case. The pdf of the deterministic case looks like a narrow log-normal distribution with a small standard deviation and has a strong peak at the distribution coefficient of 0.53. On the contrary, the pdf of the stochastic case is a wide normal distribution with a mean of 1.34 due to a positive skewness of the uncertainty factors. The narrow deterministic log-normal pdf is changed to a wide normal pdf and shifted positively under the stochastic case, and the stochastic case can thus cover a wider range of the configuration space. In addition, the types of proposed solvents are different from each other. Although most of the solvents are in one of the types of formates, alcohols, esters, or ethers, the stochastic case can provide additional types of solvents that are alkanes and alkenes. From this case study, we can see the importance of uncertainties and the usefulness of the HSTA algorithm for large-scale combinatorial stochastic programming problems.

4. Case III: Stochastic CAMD Based on δ

This section describes another stochastic solvent selection model based on Hansen's three-dimensional

Table 6. Top 20 Candidate Solvents for Both Cases

	deterministic case			stochastic case		
no.	solvents	m	β	solvents	m	β
1	$2CH_3$, CH_2, CH, HCOO	0.87	24.1	$2CH_3$, CH_2, CH, HCOO	2.95	30.1
2	CH_3, $3CH_2$, HCOO	0.87	24.1	CH_3, CH_2, CH=CH, HCOO	2.60	20.8
3	CH_3, CH_2, CH=CH, HCOO	0.76	16.7	CH_3, CH_2, CH_2=C, HCOO	2.55	19.0
4	CH_3, CH_2, CH_2=C, HCOO	0.75	15.2	CH_3, CH_2=CH, $2CH_2O$	2.27	9.46
5	$2CH_2$, CH_2=CH, HCOO	0.72	15.0	CH_3, CH_2=CH, CH_3O, CH−O	2.27	9.49
6	CH_3, CH, CH_2=CH, HCOO	0.72	15.0	CH_3, CH, CH_2=CH, CH_3O, CH_2O	2.15	9.02
7	CH_3, CH_2=CH, CH_3O, CH−O	0.66	7.60	CH_3, CH_2=C, CH_3CO	2.09	18.2
8	CH_3, CH,CH_2=CH, CH_2O, CH_3O	0.63	7.21	CH_3, CH=CH, CH_3CO	2.08	20.1
9	$2CH_2$, CH_2=CH, CH_2O, CH_3O	0.63	7.21	CH_3, CH_2, CH_2=C, CH_3O, CH_2O	2.04	8.12
10	CH_3, CH_2=C, CH_3CO	0.61	14.6	CH_3, $2CH_2$, CH_3CO	1.96	38.5
11	CH_3, CH=CH, CH_3CO	0.61	16.1	CH_3, CH_2, CH_2=C, CH_3O	1.84	9.11
12	CH_2, CH_2=CH, CH_3CO	0.60	14.1	$2CH_2$, CH_2=CH, CH_2O, CH_3O	1.55	7.81
13	CH_3, CH, CH_2=CH, CH_3O	0.58	8.22	CH_3, CH=CH, CHO	1.51	16.3
14	$2CH_2$, CH_2=CH, CH_3O	0.58	8.20	CH_3, CH_2, CH, CH_2=CH, CH_3O	1.49	9.92
15	CH_3, $2CH_2$, CH_3CO	0.57	30.9	$2CH_3$, CH, CH_2=C, CH_3O	1.41	9.00
16	$2CH_3$, CH, CH_3CO	0.57	30.9	CH_3, $2CH_2$, CH_2=C, CH_3O	1.40	8.98
17	CH_3, CH_2, CH_2=C<, CH_3O	0.54	7.30	CH_3, CH_2, CH_2=C, CHO	1.33	17.9
18	CH_3, CH_2, CH=CH, CH_3O	0.54	7.77	$2CH_2$, CH_2=CH, CHO	1.32	16.6
19	CH_2, CH_2=CH, CH_2CO	0.53	16.1	CH_2, CH, CH_2=CH, CHO	1.32	16.6
20	CH_3, CH_2, CH_2=CH, $2CH_2O$	0.52	7.36	$2CH_3$, CH_2, CH, CH_2=C, CH_3O	1.13	8.96

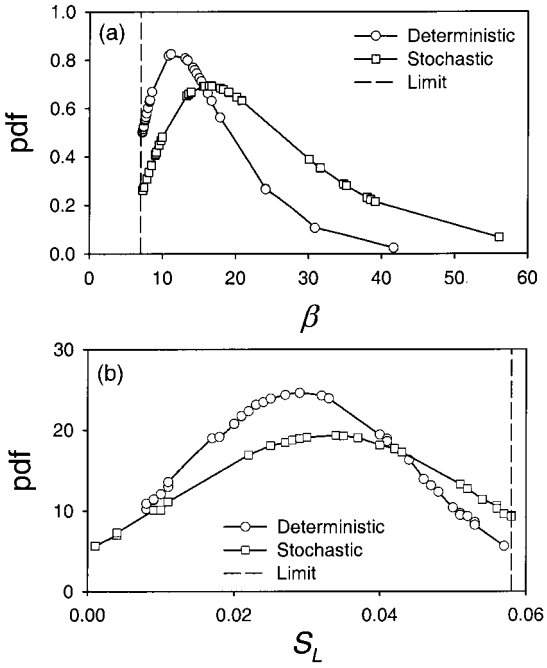

Figure 5. pdf's of constraints: (a) β; (b) S_L.

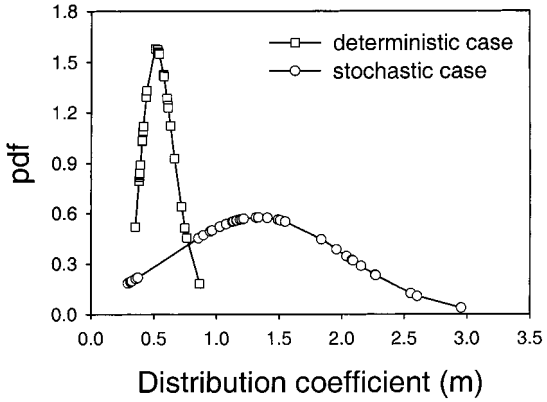

Figure 6. pdf's of distribution coefficients for the deterministic and stochastic cases based on γ^{∞}.

solubility parameters and will be used for comparison of model uncertainty with case II.

4.1. Solvent Selection Model. The solubility parameter, δ, is one of the most important parameters in physical chemistry and thermodynamics of solutions. It can serve as a key parameter for solvent selection,[10,19] solubility estimation, and the estimation of polymer swelling.[32] Though it was originally introduced by Hildebrand and Scott,[33] the most common form of the solubility parameter is Hansen's three-dimensional solubility parameter[34] which is given by

$$\delta = \sqrt{\delta_d^{\,2} + \delta_p^{\,2} + \delta_h^{\,2}}\ (\text{units: MPa}^{1/2}) \qquad (10)$$

where δ_d is the dispersive term, δ_p is the polar term, and δ_h is the hydrogen-bonding term. The solubility parameter (δ) and its three terms (δ_d, δ_p, and δ_h) can be determined by a semiempirical method and are tabulated by Barton for most common liquids.[35]

Miscibility of the two liquids i and j depends on the heat of mixing ΔH_{mix}, and ΔH_{mix} in the Hansen theory is determined by the following equation:

$$\Delta H_{mix} = (n_i V_i + n_j V_j)[(\delta_d^{\,i} - \delta_d^{\,j})^2 + (\delta_p^{\,i} - \delta_p^{\,j})^2 + (\delta_h^{\,i} - \delta_h^{\,j})^2]\Phi_1\Phi_2 \qquad (11)$$

where n is the number of moles, V is the molar volume, and Φ is the volume fraction. When the heat of mixing approaches zero, the two liquids i and j are soluble or miscible with each other. Hence, the three solubility parameter terms should be close in order to minimize the heat of mixing.

Solubility parameters for solvent selection are, in general, not as accurate as other property estimation methods such as infinite-dilution activity coefficients (γ^{∞}). However, the solubility parameter method is universal and simple to apply and thus can be used for guiding and screening candidate solvents with relatively acceptable accuracy.

The criteria for solvent selection used in cases I and II can be similarly defined. The distribution coefficient (m), a measure of the solvent capacity, is defined as[36]

$$m \propto \left(\frac{r_{BA}}{r_{BS}}\right)^2 \frac{MW_A}{MW_S} \qquad (12)$$

where A, B, and S are raffinate, solute, and solvent, respectively. r_{ij} is the Euclidean distance metric between

Table 7. Hansen's Three-Dimensional Solubility Parameters and Solvent Properties of the Given System

solvent	δ_d	δ_p	δ_h	property	value
acetic acid	13.9	12.2	18.9	m	1.043
water	12.2	22.8	40.4	β	35.20
ethyl acetate	13.4	8.60	8.90	S_L	0.0041

Table 8. Solubility Parameters of Groups (Units: $MPa^{1/2}$)

group	dispersive	polar	hydrogen-bonding
$-CH_3$	0.344	−0.591	−0.848
$-CH_2-$	0.268	−0.377	−0.595
>CH−	−0.142	−0.801	−1.172
>C<	−1.163	−1.039	−2.496
CH_2=CH−	−1.163	−1.039	−2.496
CH_2=C<	−0.243	0.275	−3.542
−CH=CH−	−0.566	−0.034	−0.0776
−CH=C<	−0.695	−0.529	−3.175
>C=C<	−0.823	−1.025	−5.574
−OH	−0.648	5.548	10.630
−O−	−0.638	2.315	1.804
>C=O	−1.145	4.670	4.486
O=CH−	−1.114	5.922	5.256
−COOH	1.068	6.942	11.120
−COO−	−0.862	4.729	4.012
>NH	−1.074	3.875	2.772
−CN	−1.628	6.904	8.317
intercept	13.290	5.067	7.229

two molecules in the three-dimensional space as shown in the following equation:

$$r_{ij} = [(\delta_d{}^i - \delta_d{}^j)^2 + (\delta_p{}^i - \delta_p{}^j)^2 + (\delta_h{}^i - \delta_h{}^j)^2]^{1/2} \quad (13)$$

Liquids having similar solubility parameters are soluble or miscible with each other. The distance between solute and solvent (r_{BS}) should be small, while the distance between solute and raffinate (r_{BA}) is fixed for a given solute/raffinate system. Thus, solvents having a smaller r_{BS} can increase m, and a high m reduces the size of the extracting equipment and the amount of recycling solvent.

Solvent selectivity (β), the ratio between distribution coefficients of solute and raffinate, is given by

$$\beta = \frac{m_B}{m_A} \propto \left(\frac{r_{AS}}{r_{BS}}\right)^2 \frac{MW_B}{MW_A} \quad (14)$$

where r_{AS} is defined in a similar way. Solvent loss (S_L) can be expressed by the following equation:

$$S_L \propto \left(\frac{1}{r_{SA}}\right)^2 \frac{MW_S}{MW_A} \quad (15)$$

Low S_L means high selectivity toward the solute and determines the immiscibility between solvent and raffinate. Table 7 shows Hansen's three-dimensional solubility parameters for the acetic acid extraction case study and calculates solvent properties (m, β, and S_L) for this case study.

A total of 17 groups and their three-dimensional solubility parameter terms are shown in Table 8. Joback[10] developed a linear GCM for the three solubility parameter terms using the least-squares method from the literature data.[35] In this table, each column consists of group-specific solubility parameters and the intercept value. The solubility parameter is then estimated by linearly adding group properties and the intercept. For example, ethyl acetate ($CH_3-COO-CH_2-CH_3$) has three distinctive groups, and its dispersive solubility parameter term is estimated to 13.38 $MPa^{1/2}$ ($2 \times 0.344 + 0.268 - 0.862 + 13.290$). Its polar and hydrogen-bonding terms are 8.24 and 8.95 $MPa^{1/2}$, respectively. The literature values for δ_d, δ_p, and δ_h are 13.40, 8.60, and 8.90, respectively, so one can see that the estimated and literature data are very close.

Table 9. Example Calculations of the Hansen's Solubility Parameters

	ethanol			ethyl acetate			diisobutyl ketone		
	lit.	est	Δ (%)	lit.	est	Δ (%)	lit.	est	Δ (%)
δ_d	12.6	11.7	−7.4	13.4	13.4	−0.1	14.5	13.8	−5.0
δ_p	11.2	9.7	−13.9	8.6	8.2	−4.2	6.8	5.0	−26.2
δ_h	20.0	16.4	−17.9	8.9	9.0	0.6	3.9	5.2	32.0
δ	26.2	22.3	−14.5	18.2	18.1	−0.9	16.5	15.5	−5.8

A set of groups in Table 8 is designed only for linear or branched hydrocarbons and not for the aromatic, cyclic, and/or halogenated compounds because of environmental concerns. As described earlier, one of the features of the solubility parameter method is universality: we can use the same groups for estimating other properties such as Joback's boiling point method,[3] and thus there is no need to have another group set.

4.2. Uncertainty Identification and Quantification. The Hansen solubility parameters of liquid molecules are estimated by semiempirical methods, and the three solubility parameter terms for each group are regressed using the least-squares method. Table 9 shows some examples of estimation errors of the Hansen's three-dimensional solubility parameters. For example, ethanol shows a 15% mean deviation in the (total) solubility parameter, mainly because of the error in the δ_h term. As explained earlier, the estimated solubility parameters of ethyl acetate are quite close to the literature values. For diisobutyl ketone, one can find an interesting result. Though each solubility parameter term has a large discrepancy, the resulting total solubility parameter is closer to the value reported in the literature. However, what is important is to quantify an individual uncertainty for each solubility parameter term and not the uncertainty in the total solubility parameter.

The GCM based on δ has 17 groups as shown in Table 8, and each group has three solubility parameter terms. Because the total uncertainty distributions are 51, it is impractical and statistically insignificant to figure out each uncertainty distribution. Like the solvent selection model based on γ^∞, the uncertainties of the Hansen solubility parameters (δ_d, δ_p, and δ_h) of the solute/solvent systems are analyzed and quantified in terms of the uncertainty factor (UF). Here, UF is the ratio of the literature solubility parameter[36] to the calculated solubility parameter.

$$UF = \delta_{lit}/\delta_{cal} \quad (16)$$

where UF can be applied to dispersive, polar, or hydrogen-bonding terms. Also note that a UF of 1.0 means that the estimated value is exactly equal to the literature value.

To elicit the UFs, the estimated solubility parameters of 66 noncyclic and nonaromatic compounds are compared with the literature values. The probabilistic distributions of the three UFs associated with the three solubility parameter terms are shown in Figure 7. The UF of the dispersive term (δ_d) is normally distributed with a 1.05 mean and a 0.08 standard deviation. The UFs of δ_p and δ_h are normally distributed with a 1.21

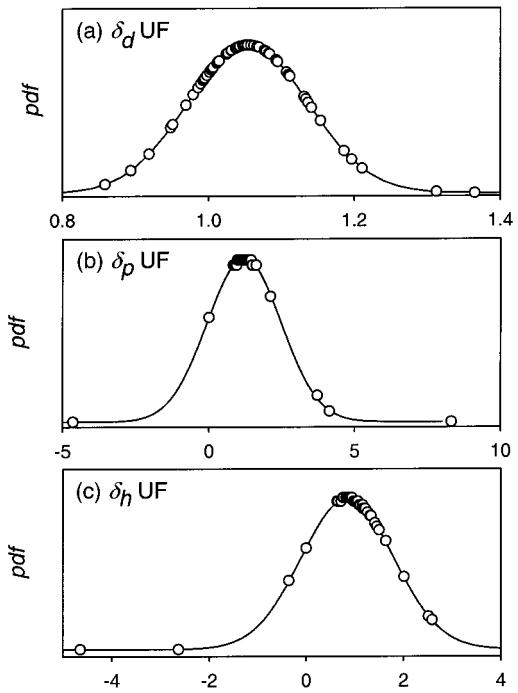

Figure 7. UFs for dispersive, polar, and hydrogen-bonding terms.

Table 10. Constraints for Solvent Selection under Uncertainty Based on Hansen's Solubility Parameters

parameter	bounds
N_1 (no. of groups in a molecule)	2−12
β	≥17.43
S_L	≤0.0045
T_{BP} (°C)	47−108

mean and a 1.28 standard deviation and with a 0.85 mean and a 0.96 standard deviation, respectively. From the figure, we can see that the effect of uncertainty on δ_d is not significant as compared to the effects on the polar and hydrogen-bonding terms. It is equivalent to the fact that the variations in δ_d for most solvents are small and thus a two-dimensional plot of δ_p and δ_h is commonly used to correlate solvent properties. It can be said that UF increases an estimated δ_p but decreases an estimated δ_h, and the resulting δ, m, β, and S_L are subject to change.

4.3. Results and Discussion. The generation-and-test CAMD approach commonly used with the solubility parameter model is computer-intensive, because this method tries to generate all possible molecular combinations. If the maximum number of groups in a molecule, for example, is 12, then the total number of possible molecular combinations is 17^{12} (5.8×10^{14}). In addition, if we consider uncertainties in the solubility parameter terms of groups, this problem becomes computationally very expensive. The HSTA algorithm is applied to this stochastic solvent selection model.

The main priority (objective function) is the distribution coefficient, while the other properties are used as constraints, which are summarized in Table 10. The number of groups (N_1) in a solvent molecule spans from 2 to 12. The bounding values for β, S_L, and the boiling points are based on the values of ethyl acetate, one of common solvents for acetic acid extraction.

Table 11 presents the first top 20 candidate solvents with and without consideration of uncertainty. For the deterministic case, the first 10 solvents are alcohols and the remaining solvents are mostly aldehyde. Some of the promising solvents in this case are ethyl alcohol (no. 1), propyl alcohol (no. 3), isopropyl alcohol (no. 5), acetone (no. 12), and methyl ethyl ketone (no. 16). The alcohol function group has the largest hydrogen-bonding term and also the second largest polar term. This feature of the OH group decreases r_{BS}, resulting in an increase of m. Similarly, the CHO group has the second largest hydrogen-bonding term and the largest polar term, which also decreases r_{BS}. Large differences in hydrogen-bonding terms between the OH and CHO groups (10.63 vs 5.26) and small differences in the polar terms (5.55 vs 5.92) make alcohols preferred solvents for acetic acid extraction.

However, the stochastic case provides a different set of candidate solvents. Only 13 of the solvents generated in the deterministic case appear on the list of the stochastic case. Some of promising solvents at the stochastic case are isopropyl acetate (no. 1), isopropyl alcohol (no. 3), acetone (no. 5), and propyl alcohol (no. 6). Isopropyl acetate, which is not listed in the top 20 solvents for the deterministic case, is one of the common industrial solvents for acetic acid extraction and has proven to be highly selective for this extraction purpose. Acetone, which is appeared in both cases, was also reported as the best solvent for this purpose by Joback and Stephanopoulos.[3] They used a different solubility parameter method but similar constraints even though their CAMD approach was the generation-and-test method. The combinatorial optimization methods used in this study provide more promising solvents than acetone. Ethyl acetate, one of the most common industrial solvents, is generated outside the top 20 candidate solvents at both cases because the m for ethyl acetate (1.04) is relatively low.

Because the mean of UF for δ_h is 0.85, the contribution of the hydrogen-bonding term is decreased in the stochastic case, resulting in different functional groups in the solvent. In addition, the reduced δ_h term also decreases the resulting distribution coefficients. If one looks at the distribution coefficients of solvents generated in both cases, the expected value of m under uncertainty is slightly smaller than that of the deterministic case.

As seen in Figure 8, the mean UF of δ_h also reduces the stochastic β and S_L values. Thus, the β constraint approaches the constraint limit while the S_L constraint goes away from the limit. This phenomenon is opposite to that of case II.

Figure 9 shows the frequency of the functional groups in the top 40 solvents in both cases. For the deterministic case, as expected, OH and CHO groups are the most common types in the candidate solvents. However, for the stochastic case, other functional groups also have high frequencies, and alkanes and alkenes are in the list of optimal solvents. This means that the stochastic case provides a wider range of solvents.

Probability density functions (pdf's) of distribution coefficients for both cases are shown in Figure 10. The pdf for the deterministic case is log-normally distributed, while the one for the stochastic case is governed by the Weibull distribution with a shape parameter of 1.51 and a scale parameter of 2.58. The Weibull distribution is one of the asymptotic distributions of general extreme value theory; hence, this distribution can approximate the extremely high value of m of isopropyl acetate under uncertainty. We can also conclude from this figure that the pdf of the stochastic case is narrower

Table 11. Top 20 Candidate Solvents

	deterministic case			stochastic case		
no.	solvents	*m*	*b*	solvents	*m*	*b*
1	CH_3, CH_2, OH	17.2	190.6	$3CH_3$, CH, COO	161	835.1
2	CH_3, CH=CH, OH	12.6	175.9	$2CH_3$, CH_2=C, CO	20.1	78.3
3	CH_3, $2CH_2$, OH	9.50	144.4	$2CH_3$, CH, OH	5.95	22.2
4	CH_2=CH, OH	9.34	105.9	CH_3, CH_2, CH_2=CH, COO	4.91	30.9
5	$2CH_3$, CH, OH	6.06	99.0	$2CH_3$, COO	4.26	18.2
6	CH_2=CH, $2CH_2$, OH	5.83	91.4	CH_3, $2CH_2$, OH	3.90	15.5
7	CH_3, CH_2=C, OH	5.20	85.3	CH_3, CH_2, OH	3.53	11.4
8	CH_2=CH, $2CH_2$, OH	3.86	72.8	$2CH_3$, CO	3.51	11.4
9	CH_3, CH_2=C, CH_2, OH	3.41	60.7	$2CH_2$, CH_2=CH, OH	3.38	12.9
10	CH_2, CH_2=CH, CH, OH	2.78	47.7	CH_2, CH_2=CH, OH	3.34	12.2
11	CH_2, CH=CH, CHO	2.07	42.3	$2CH_3$, CH=C, CHO	2.85	13.7
12	$2CH_3$, CO	2.06	45.4	CH_2=CH, OH	2.77	8.82
13	CH_3, $2CH_2$, CHO	1.83	43.45	CH_2, CH=CH, OH	2.58	10.4
14	CH_3, CH_2, CH=CH, CHO	1.50	38.8	CH_3, $2CH_2$, CHO	2.40	10.6
15	$2CH_3$, CH, CHO	1.47	38.4	$2CH_3$, CH_2, CO	2.16	10.2
16	$2CH_3$, CH_2, CO	1.44	36.4	CH_3, CH_2=C, OH	2.10	8.56
17	CH_2=CH, CH_2, CHO	1.42	38.1	$2CH_3$, CH_2, CH=CH, O	2.09	11.6
18	$2CH_3$, COO	1.39	41.0	$2CH_3$, CH=CH, CO	2.01	10.6
19	CH_3, $3CH_2$, CHO	1.33	32.6	$2CH_3$, CH_2=C, COO	1.86	8.89
20	CH_3, CH_2=C, CHO	1.22	37.4	$2CH_3$, CH_2, CH, CHO	1.85	13.7

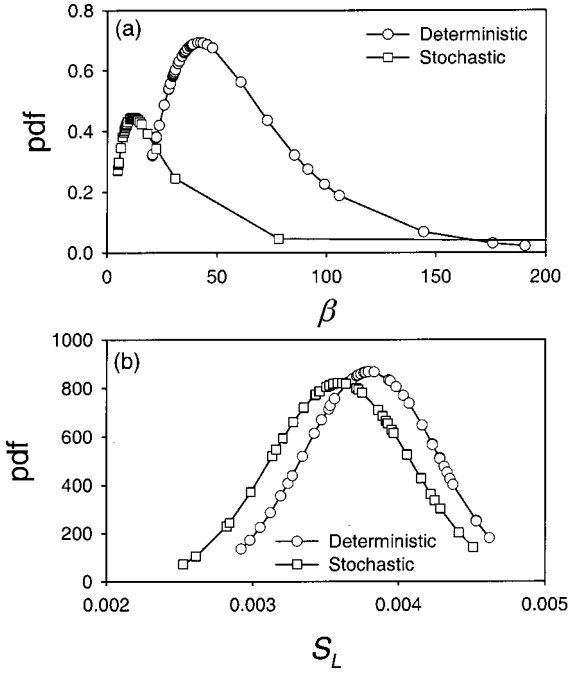

Figure 8. pdf's of constraints: (a) β; (b) S_L.

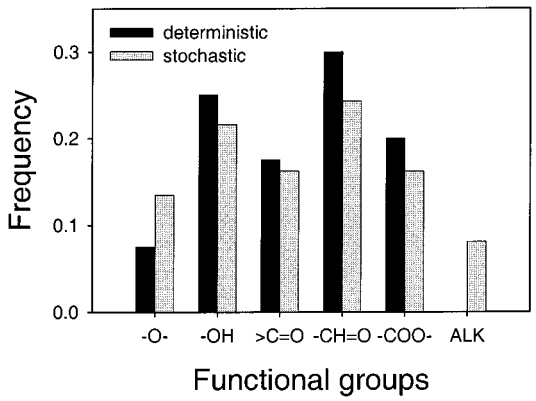

Figure 9. Frequencies of functional groups (ALK means alkanes and alkenes).

and shows robustness in the solvent selection process. The reduced δ_p and the resulting smaller *m* can be attributed to this narrow peak of candidate solvents.

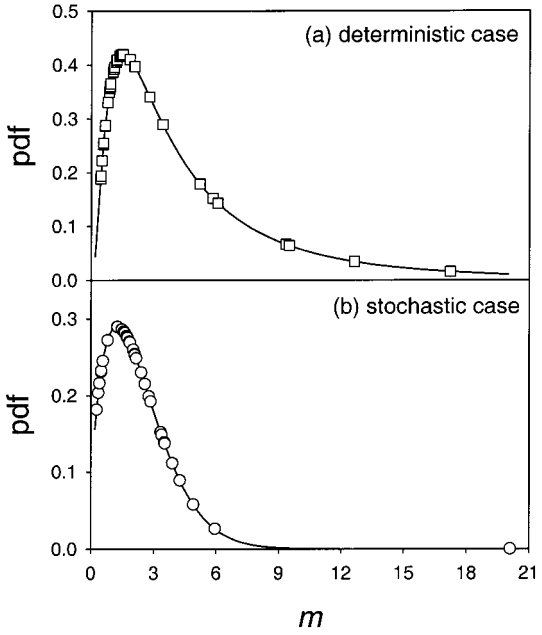

Figure 10. Probabilistic density functions of the distribution coefficients *m*.

5. Comparison between Cases II and III

Two different solvent selection models, the infinite dilution activity coefficient model and the solubility parameter model, are used to generate candidate solvents. The results from the two models have similarities as well as differences.

In both models, acetate solvents are eliminated from the top solvents because acetate solvents have low *m* values (e.g., ethyl acetate, $m = 0.32$). Ethyl acetate, one of the most widely used commercial solvents, is listed at 27th and 21st in the stochastic γ^∞- and δ-based models, respectively. Thus, current acetate solvents such as ethyl acetate, isopropyl acetate, and isoamyl acetate may not be good choices in the context of solvent properties.

Ketone solvents, except acetone, are generated by both solvent selection models, and hence ketones can have a high chance of being good candidate solvents in real implementations. Because acetone, also suggested by a different δ-based model,[3] has a relatively high

solvent loss based on γ^{∞}, it is not listed from the γ^{∞}-based model.

Differences between the two models may arise from different model accuracy, different expressions of m, β, and S_L and hence different constraint values, availability of groups, and different levels of uncertainties of model parameters.

The most promising solvents generated from each model are different. The γ^{∞}-based model generates formates (HCOO−) as the best solvents, while the δ-based model provides alcohols as the best solvents in both the deterministic and stochastic cases. The formate group, although not present in the current δ-based model, has a great impact on solvents to have a high m. However, the S_L values of formate solvents are generally close to the given constraint (i.e., $S_{L,max} = 0.058$), and thus formate solvents are very sensitive to the solvent loss criterion. The alcoholic solvents definitely have a high m in both solvent selection models, but the S_L values are significantly different. The γ^{∞}-based S_L of (low-boiling) alcohols cannot satisfy the S_L constraint, while the δ-based S_L does. As seen in case I, high-boiling alcohols can be good candidate solvents (for a process which prefers high-boiling-point solvents) because they have a low S_L as well as a high m. The relatively low S_L values in the δ-based model may arise from the equivalence in the Euclidean distance metric between liquids i and j (i.e., $r_{ij} = r_{ji}$), which is used in the expression of S_L. Because we did not include uncertainties (errors) associated with S_L in the analysis, the results from the two cases are different.

Solvent selection models can generate different results because of model limitations as described. Thus, it is important to look at the basic assumptions (e.g., availability of groups) of the models and possible errors in the problem design (e.g., $S_{L,max}$) in order to achieve the better quality of the candidate solvents.

6. Conclusion

This paper presented the application of the new and efficient HSTA algorithm for solvent selection under uncertainty. Model uncertainties, expressed by different models, and model parameter uncertainties, expressed by UF, are considered here. Case I shows the usefulness of the new combinatorial optimization algorithm for deterministic solvent selection problems and compares this with literature results. The new ESA algorithm can generate similar results as compared to literature but provide a more robust set of candidate solvents. Two stochastic solvent selection models based on γ^{∞} and δ are used to generate sets of candidate solvents. The stochastic solvent selection model in case II is based on γ^{∞}, while case III is based on δ. The uncertainties in the model parameters are represented by the type of probability distribution and UF. The γ^{∞}-based model supports formate solvents as candidate solvents, whereas the δ-based model supports alcoholic solvents. Ketones are generated in both models, and acetates are listed outside the top 20 solvents because of low m. The difference between the two solvent selection models can be attributed to different model accuracy, different expressions of solvent criteria, availability of groups, and hence different levels of uncertainties. Therefore, once a solvent selection model is determined, an HSTA algorithm can efficiently explore the problems of the solvent selection under uncertainty and can provide reliable candidate solvents. There are several good papers[5,20] which evaluate these candidate solvents in chemical processes, and *simultaneous* integration of solvent selection and process synthesis is one of the open areas in solvent selection problems.

Acknowledgment

We greatly appreciate the National Science Foundation for funding this research (Goali Project CTS-9729074).

Literature Cited

(1) Zhao, R.; Cabezas, H. Molecular thermodynamics in the design of substitute solvents. *Ind. Eng. Chem. Res.* **1998**, *37*, 3268.

(2) Gani, R.; Brignole, E. A. Molecular design of solvents for liquid extraction based on Unifac. *Fluid Phase Equilib.* **1983**, *13*, 331.

(3) Joback, K. G.; Stephanopoulos, G. Designing molecules possessing desired physical property values. *Foundations of Computer-Aided Process Design*; Snowmass Village, CO, 1989; pp 363−387.

(4) Marcoulaki, E. C.; Kokossis, A. C. On the development of novel chemicals using a systematic synthesis approach. Part I. Optimisation framework. *Chem. Eng. Sci.* **2000**, *55*, 2529.

(5) Buxton, A.; Livingston, A. G.; Pistikopoulos, E. N. Optimal design of solvent blends for environmental impact minimization. *AIChE J.* **1999**, *45*, 817.

(6) Cabezas, H.; Zhao, R.; Bare, J. C. Designing environmentally benign solvent substitutes. In *Tools and Methods for Pollution Prevention*; Sikdar, S. K., Diwekar, U. M., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1999; pp 317−331.

(7) Vaidyanathan, R.; El-Halwagi, M. Computer-aided synthesis of polymers and blends with target properties. *Ind. Eng. Chem. Res.* **1996**, *35*, 627.

(8) Maranas, C. D. Optimal computer-aided molecular design: A polymer design case study. *Ind. Eng. Chem. Res.* **1996**, *35*, 3403.

(9) Constantinou, L.; Jaksland, C.; Bagherpour, K.; Gani, R. Application of the group contribution approach to tackle environmentally related problems. In *Pollution Prevention via Process and Product Modifications*; El-Halwagi, M. M., Petrides, D. P., Eds.; AIChE Symposium Series; AIChE: New York, 1994; pp 105−116.

(10) Joback, K. G. Solvent substitution for pollution prevention. *Pollution Prevention via Process and Product Modifications*; AIChE Symposium Series; AIChE: New York, 1994; pp 98−104.

(11) Gani, R.; Nielsen, B.; Fredenslund, A. A group contribution approach to computer-aided molecular design. *AIChE J.* **1991**, *37*, 1318.

(12) Pretel, E. J.; López, P. A.; Bottini, S. B.; Brignole, E. A. Computer-aided molecular design of solvents for separation processes. *AIChE J.* **1994**, *40*, 1349.

(13) Odele, O.; Macchietto, S. Computer aided molecular design: A novel method for optimal solvent selection. *Fluid Phase Equilib.* **1993**, *82*, 47.

(14) Churi, N.; Achenie, L. E. Novel mathematical programming model for computer aided molecular design. *Am. J. Math. Manage. Sci.* **1996**, *35*, 3788.

(15) Venkatasubrananian, V.; Chan, K.; Caruthers, J. Computer-aided molecular design using genetic algorithms. *Comput. Chem. Eng.* **1994**, *18*, 833.

(16) Marcoulaki, E. C.; Kokossis, A. C. Molecular design synthesis using stochastic optimization as a tool for scoping and screening. *Comput. Chem. Eng.* **1998**, *22*, s11.

(17) Kim, K.-J.; Diwekar, U. M.; Joback, K. G. Greener solvent selection under uncertainty. *Clean Solvents: Alternative Media for Chemical Reactions and Processing*; ACS Sumposium Series 819; Abraham, M., Moens, L., Eds.; ACS Publishing: Washington, DC, 2002.

(18) Maranas, C. D. Optimal molecular design under property prediction uncertainty. *AIChE J.* **1997**, *43*, 1250.

(19) Sinha, M.; Achenie, L. E. K.; Ostrovsky, G. M. Environmentally benign solvent design by global optimization. *Comput. Chem. Eng.* **1999**, *23*, 1381.

(20) Hostrup, M.; Harper, P. M.; Gani, R. Design of environmentally benign processes: Integration of solvent design and separation process synthesis. *Comput. Chem. Eng.* **1999**, *23*, 1395.

(21) Hansen, H. K.; Rasmussen, P.; Fredenslund, A.; Schiller, M.; Gmehling, J. Vapor−Liquid equilibria by UNIFAC group contribution. 5. Revision and Extension. *Ind. Eng. Chem. Res.* **1991**, *30*, 2352.

(22) Joback, K. G.; Reid, R. C. Estimation of pure-component properties from group-contributions. *Chem. Eng. Commun.* **1987**, *57*, 233.

(23) Bastos, J. C.; Soares, M. E.; Medina, A. G. Infinite dilution activity coefficients predicted by UNIFAC group contribution. *Ind. Eng. Chem. Res.* **1988**, *27*, 1269.

(24) Gmehling, J.; Li, J.; Schiller, M. A modified UNIFAC model. 2. Present parameter matrix and results for different thermodynamic properties. *Ind. Eng. Chem. Res.* **1993**, *32*, 178.

(25) Zhang, S.; Hiaki, T.; Hongo, M.; Kojima, K. Prediction of infinite dilution activity coefficients in aqueous solutions by group contribution models. A critical evaluation. *Fluid Phase Equilib.* **1998**, *144*, 97.

(26) Gmehling, J. Group contribution methods for the estimation of activity coefficients. *Fluid Phase Equilib.* **1986**, *30*, 119.

(27) Tiegs, D.; Gmehling, J.; Medina, A.; Soares, M.; Bastos, J.; Alessi, P.; Kikic, I. *Activity Coefficients of Infinite Dilutions*; DECHEMA: Frankfurt, Germany, 1986; Vol. 6, Part 1.

(28) Gmehling, J.; Menke, J.; Schiller, M. *Activity Coefficients of Infinite Dilutions*; DECHEMA: Frankfurt, Germany, 1994; Vol. 6, Part 3.

(29) Kim, K.-J.; Diwekar, U. M. Efficient Combinatorial Optimization under Uncertainty. 1. Algorithmic Development. *Ind. Eng. Chem. Res.* **2002**, *41*, xxxx.

(30) Kalagnanam, J. R.; Diwekar, U. M. An efficient sampling technique for off-line quality control. *Technometrics* **1997**, *39*, 308.

(31) Birge, J. R.; Louveaux, F. *Introduction to Stochastic Programming*; Springer: New York, 1997.

(32) Panayiotou, C. Solubility parameter revisited: An equation-of-state approach for its estimation. *Fluid Phase Equilib.* **1997**, *131*, 21.

(33) Hildebrand, J.; Scott, R. *Regular Solutions*; Prentice-Hall: Englewood Cliffs, NJ, 1962.

(34) Hansen, C. M. The universality of the solubility parameter. *Ind. Eng. Chem. Prod. Res. Dev.* **1969**, *8*, 2.

(35) Barton, A. *CRC Handbook of Solubility Parameters and Other Cohesion Parameters*; CRC Press: Boca Raton, FL, 1983.

(36) Lo, T. C., Baird, M. H. I., Hanson, C., Eds. *Handbook of Solvent Extraction*; John Wiley & Sons: New York, 1983.

Received for review February 21, 2001
Revised manuscript received September 6, 2001
Accepted November 29, 2001

IE0101691