SETAC PRESS

Environmental Toxicology and Chemistry, Vol. 19, No. 4, pp. 810–819, 2000
© 2000 SETAC
Printed in the USA
0730-7268/00 $9.00 + .00

EVALUATING MULTIMEDIA CHEMICAL PERSISTENCE: CLASSIFICATION AND REGRESSION TREE ANALYSIS

DEBORAH H. BENNETT,† THOMAS E. MCKONE,*†‡ and WILLIAM E. KASTENBERG§
†Environmental Energy Technologies Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA
‡School of Public Health, University of California at Berkeley, Berkeley, California 94720, USA
§Department of Nuclear Engineering, University of California at Berkeley, Berkeley, California 94720, USA

(*Received* 24 *February* 1999; *Accepted* 19 *July* 1999)

Abstract—For the thousands of chemicals continuously released into the environment, it is desirable to make prospective assessments of those likely to be persistent. Widely distributed persistent chemicals are impossible to remove from the environment and remediation by natural processes may take decades, which is problematic if adverse health or ecological effects are discovered after prolonged release into the environment. A tiered approach using a classification scheme and a multimedia model for determining persistence is presented. Using specific criteria for persistence, a classification tree is developed to classify a chemical as ''persistent'' or ''nonpersistent'' based on the chemical properties. In this approach, the classification is derived from the results of a standardized unit world multimedia model. Thus, the classifications are more robust for multimedia pollutants than classifications using a single medium half-life. The method can be readily implemented and provides insight without requiring extensive and often unavailable data. This method can be used to classify chemicals when only a few properties are known and can be used to direct further data collection. Case studies are presented to demonstrate the advantages of the approach.

Keywords—Classification tree Multimedia Persistence Characteristic time Decay rates

INTRODUCTION

Some chemical pollutants released into the environment degrade rapidly, while others persist for years or even decades [1]. Persistent pollutants have higher environmental concentrations per unit release and remain in the environment for extended periods even after use has ceased [2,3]. Concern arises because widely distributed persistent chemicals are impossible to remove from the environment and remediation by natural processes may take decades. Additionally, adverse health or ecological effects may only be discovered after a prolonged release into the environment. Hence, there is a need for simple but reliable methods for identifying the increasingly large number of potentially persistent pollutants.

For pollutants that partition primarily to a single medium, it is relatively easy to estimate the persistence—it is the pollutant half-life in the medium to which it is released (relatively easy assuming the half-life data is available). However, many persistent chemicals partition into multiple environmental media, and calculating the measure of persistence increases in complexity. This complexity occurs because degradation rates are media dependent, thus partitioning and transfer rates also play an important role in determining persistence [1–4].

The goal of this paper is to develop a two-tiered framework for classifying chemicals as persistent or nonpersistent in a multimedia environment. We employ a classification scheme in the first tier in order to reduce the need for explicitly modeling the persistence of numerous new and existing chemicals. This screening allows some chemicals to be classified as nonpersistent without extensive modeling and data collection. Those evaluated in the second tier require analysis with a multimedia model.

* To whom correspondence may be addressed (temckone@lbl.gov).

To develop the first tier classification method, we link the results of a simple multimedia model to a classification system, in this case the classification and regression tree (CART) methodology [5]. We then focus on sets of chemical properties rather than on specific chemicals, with the range of each chemical property falling within a plausible range for synthetic organic chemicals. We use Monte Carlo simulations to develop sets of properties for hypothetical chemicals, i.e., a set of chemical realizations, with each of the properties falling within the plausible range. We then use a multimedia model to calculate the persistence for each realization in a defined environmental system. We then apply the CART methodology to the results to determine the chemical parameter sets that are likely to result in a nonpersistent chemical. Chemicals that do not fall into the nonpersistent classification are not necessarily persistent. For these chemicals, a second tier of testing, using a full multimedia model and all chemical properties, is needed to evaluate persistence. The CART methodology, along with Monte Carlo simulations from a multimedia model, can be used as tools to develop effective strategies for the classification of chemicals without requiring an explicit simulation for each chemical.

We first present a unit world multimedia model and define our measure of persistence based on this model. Next, we discuss the CART system along with its application in the absence of complete data, and finally, a case study is used to demonstrate the approach.

METHODS

Defining the multimedia environment

To characterize chemicals that spread among multiple media, we need to employ a multimedia model, an approach often used for evaluating persistent chemicals [2–4]. Because many persistent organic pollutants are widespread across the earth,

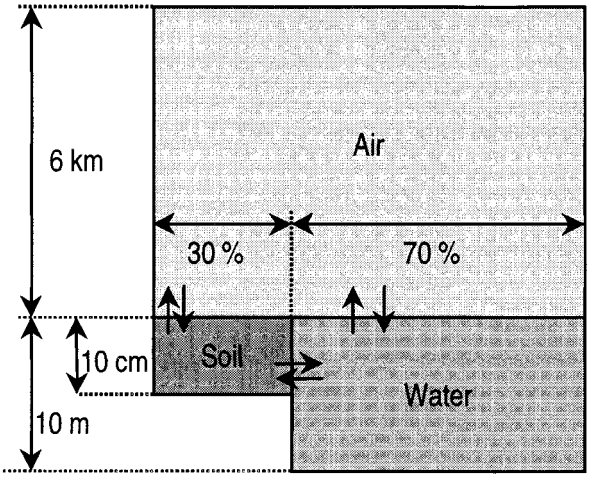


Fig. 1. Configuration of the evaluation unit used in the case studies.

we use the three-compartment closed unit world system presented in Figure 1. A unit world model approximates the actual distribution of environmental media found across the earth [6].

Interaction between the different media are treated using the principles of fugacity partitioning and mass transfer [7,8]. The fugacity capacities (the chemical concentration per unit chemical fugacity) and inventory-based mass transfer coefficients, T values, are listed in Table 1. The landscape property values as well as the calculated values needed in determining the mass distribution are listed in Table 2 with their corresponding references.

We use a steady-state mass distribution. This system replicates the eventual mass distribution in an environmental system with a steady source (after sufficient time such that the initial conditions no longer effect the system). The steady-state mass distribution accounts for equilibrium partitioning, the source term, and advective processes within the environment. Advective processes are active transport processes such as rainfall or runoff and may transfer mass from a region of lower fugacity, or chemical potential, to a region of higher fugacity. The steady-state mass distribution approximates the dynamic environmental mass distribution while retaining sufficient simplicity to complete calculations in a tractable form, such as a spreadsheet [2]. The equations that define the mass distribution between compartments are described in the Appendix.

Persistence in a multimedia environment

Persistence in the environment is calculated from the total mass in a steady-state system divided by the total mass loss rate from the system (based on reaction, there is no advection from the system). The user must know the mass distribution between environmental media and the chemical transformation rate in each medium. Persistence can be defined by the characteristic time, τ, calculated as [2]

Table 1. Fugacity capacities and fugacity mass transfer coefficients; notation and parameter values defined in Table 2

Fugacity capacities	
Air compartment	$Z_a = Z_{air} + Z_{ap}$
Pure air	$Z_{air} = \frac{1}{RT}$
Air particles	If $Tm < T$, $Z_{ap} = \frac{0.17 \times SA}{RT \times VP}$; If $Tm > T$, $Z_{ap} = \frac{0.17 \times SA}{RT \times VP \times e^{(-6.81(1-Tm/T))}}$
Water	$Z_{water} = 1/K_H$
Soil compartment	$Z_s = Z_{air}\alpha_s + Z_{water}\beta_s + (1 - \alpha_s - \beta_s)Z_{sp}$
Soil particles	$Z_{sp} = \frac{K_d \times \rho_s}{K_H \times 1{,}000}$
Fugacity mass transfer coefficients	
Air to soil	$T_{as} = \frac{(\gamma_{as} + rain \times Z_{water} + (v_d + Wr \times rain)Z_{ap}) \times (1 - fw)}{Z_a d_a}$
Transfer resistance, air and soil	$\gamma_{as} = \left(\frac{\delta_{as}}{Z_a D_{air}} + \frac{\delta_s}{Z_s D_s}\right)^{-1}$
Diffusion coefficient in soil	$D_S = \frac{\beta_s^{10/3} D_{water} Z_{water} + \alpha_s^{10/3} D_{air} Z_{air}}{Z_s(\alpha_s + \beta_s)^2}$
Soil to air	$T_{sa} = \frac{\gamma_{as} + v_d Z_{sp} \frac{\rho_{ba}}{\rho_s}}{Z_s d_s}$
Air to water	$T_{aw} = \frac{\gamma_{aw} fw + (rain \times Z_{water} + (v_d + Wr \times rain)Z_{ap}) \times fw}{Z_a d_a}$
Transfer resistance, air and water	$\gamma_{aw} = \left(\frac{\delta_{aw}}{Z_{air} D_{air}} + \frac{\delta_{wa}}{Z_{water} D_{water}}\right)^{-1}$
Water to air	$T_{wa} = \frac{\gamma_{aw}}{Z_{water} d_w}$
Soil to water	$T_{sw} = \frac{runoff \times Z_{water} + erosion \times Z_{sp}/\rho_s}{Z_s d_s}$

Table 2. Properties and parameters used in calculations

Chemical or landscape property	Notation	Mean value/equation	Coefficient of variation	Reference
Depth of air (m)	d_a	6,000	none	[6]
Depth of water (m)	d_w	10	none	[6]
Depth of soil (m)	d_s	0.1	none	[6]
Fraction of area in surface water	*fw*	0.70	none	[6]
Universal gas constant (Pa·m³/mol·K)	*R*	8.31	none	
Ambient environmental temperature (K)	*T*	288	0.02	[19]
Surface area of particles (m²/m³)	*SA*	1.5×10^{-4}	0.1	[20]
Washout ratio	*Wr*	5.0×10^{4}	2	[20]
Atmospheric dust load (kg/m³)	ρ_{ba}	5.0×10^{-8}	0.2	[20]
Dry deposition velocity of air particles (m/d)	v_d	43.2	0.3	[20]
Boundary layer thickness in air above soil (m)	δ_{as}	0.005	none	[21]
Boundary layer thickness in air above water (m)	δ_{aw}	0.013	none	Calculated from [7]
Boundary layer thickness in water below air (m)	δ_{wa}	0.00013	none	Calculated from [7]
Boundary layer thickness in soil below air (m)	δ_s	$0.216D_s^{0.229}$		[22]
Annual average precipitation (m/d)	*rain*	0.002	1	[19]
Soil particle density (kg/m³)	ρ_s	2.6×10^{3}	0.05	
Water runoff rate (m/d)	*runoff*	9.4×10^{-4}	0.2	[7]
Soil runoff rate (kg/m²·d)	*erosion*	3.0×10^{-3}	0.2	[7]
Water content of root-zone soil (vol. fraction)	β_s	0.28	0.2	
Air content of root-zone soil (vol. fraction)	α_s	0.17	0.2	
Organic carbon fraction in upper soil zone	f_{oc}	0.02	1	
Partition coefficient in ground/root soil layer	K_d	$0.41 \cdot K_{ow} \cdot f_{oc}$		

$$\tau = \frac{M_1 + M_2 + M_3 + \ldots + M_j}{M_1k_1 + M_2k_2 + M_3k_3 + \ldots + M_jk_j} \quad (1)$$

where M_i is the chemical mass and k_i is the transformation rate constant in compartment i, $i = 1, 2, 3, \ldots, j$. This is equivalent to

$$\tau = \frac{M_1 + M_2 + M_3 + \ldots M_j}{M_1(1/\tau_1) + M_2(1/\tau_2) + M_3(1/\tau_3) + \ldots + M_j(1/\tau_j)} \quad (2)$$

where $\tau_i = 1/k_i$.

The characteristic time is used to classify a chemical as persistent or nonpersistent by comparison with a specified cutoff value of τ. The cutoff value defines the boundary between characteristic times that are considered non-persistent and those considered persistent. If six months were selected, all chemicals with a characteristic time of less than six months would be classified as non-persistent while all chemicals with a characteristic time of more than six months would be classified as persistent. Policy or appropriate regulations should determine the cutoff value. In the examples that follow, we use six months and five years as cutoff criteria for illustrative purposes.

The classification and regression tree approach

The CART algorithm is a nonanalytic, computationally intensive, statistical procedure that classifies data by producing a tree structure [5]. At each node in the tree, CART uses a parametric decision based on an inequality for a specific input parameter. Each inequality is chosen by the program to maximize the reduction in variance between the parent node and the resultant two subnodes. The CART algorithm compares all possible splits at each node, considering all parameters and all values for that parameter. The tree is grown until either there is no variance or the sample size is small at each node. When the tree is complete, CART assigns each terminal node a class outcome. The tree can be trimmed to optimize between the variance and the complexity of the tree [5].

The results of the CART approach are concise and easy to understand, which should assist decision making. The approach is similar to a response-surface model for the more complex multimedia model, providing an advantage by limiting the need to run the original model. The CART method has been available for almost 20 years and has been applied in many fields, including engineering, medicine, public health, and economics [9–15].

In the field of multimedia modeling, CART has been used to evaluate parameters that determine the relationship between dioxin emissions and soil concentration levels by Eisenberg et al. [9]. Eisenberg and McKone [10] have applied CART to the output from a multimedia exposure model to determine which chemical properties result in high levels of human exposure for a unit input soil concentration. They used multiple realizations of chemical property sets, covering a broad range of property values [10]. This paper uses a similar approach to the one developed by Eisenberg and McKone [10], applying it to a different environmental model and evaluating a different environmental outcome (in this case, persistence).

Using the unit world multimedia model, we employ CART to assess the link between an output classification of ''persistent'' or ''nonpersistent'' and the range of input values that give rise to each classification. The steps needed to develop the classification tree from a multimedia model are summarized as follows: (1) Complete a standard Monte Carlo simulation to develop realizations of hypothetical chemical property sets and landscape parameters. (2) Calculate the persistence in the unit world environment for each realization. This process is completed 10,000 times in the calculations used in this paper. (3) Enter the output from each Monte Carlo simulation, the calculated persistence, into a spreadsheet with the corresponding chemical and landscape input-parameter values. (4) Determine the value of τ defining the boundary between nonpersistent and persistent chemicals. (5) Compare the calculated persistence for each realization to this value and add the classification to the spreadsheet. Enter the input-parameter values and corresponding classification from each of the simulations into the CART program. A classification tree is generated based on this data. The tree might use several chemical properties in reaching the final endpoints. Branches not yield-

ing a ''nonpersistent'' result for more than 0.5% of simulations are trimmed. (6) Read the tree by asking a question regarding the inequality listed (e.g., is the value of x greater than a?). If the answer is yes, the user follows the left branch, and if the answer is no, the user follows the right branch. The user follows this procedure until a terminal node is reached. The terminal node indicates whether the chemical can be classified as ''nonpersistent'' or the chemical needs to be evaluated in a second tier evaluation.

A ''nonpersistent'' classification may often be reached using only a few parameters. These input parameters can be used to expedite the classification of large numbers of chemicals. Ideally, these input parameters should be evaluated for all chemicals before they are released into the environment such that use of persistent chemicals could be minimized.

The classification ''nonpersistent'' is assigned to a terminal node when a specified percent of the chemicals in that node are defined as nonpersistent. The percentage required to make the ''nonpersistent'' classification can be varied in order to minimize the false negatives and hence the number of persistent chemicals misclassified as ''nonpersistent.'' One could even allow for 99% nonpersistent chemicals; however, as the specified percentage increases, the number of chemicals that need to be tested in the second tier increases, thus increasing the time required to classify a large number of chemicals. We have selected 95% as the minimum percent of chemicals classified as ''nonpersistent'' that must be nonpersistent at a given node; thus, if we select a chemical at random from this node, there is at least a 95% probability it will be a nonpersistent chemical. For some nodes, the likelihood will be higher than 95%. This classification threshold favors false positives; thus, a persistent result does not mean a chemical will be persistent but rather identifies the chemical as one that needs to be evaluated in a second tier.

Second tier testing

If the classification system is not used, the persistence of all chemicals would need to be evaluated using a multimedia model. The classification method is very effective at screening out nonpersistent chemicals, eliminating the need for further testing on a large number of chemicals. Using a classification tree, only those chemicals classified as ''persistent'' must have the characteristic time calculated with a multimedia model, such as the model in Figure 1. At each persistent node, there are chemicals that are both persistent and nonpersistent as a result of our preference for false positives over false negatives. A chemical may be classified as ''persistent'' using the classification tree even though the same multimedia model used to generate the CART tree will determine that the chemical is nonpersistent. The chemical may have been classified as ''persistent'' because many of the chemicals with some similar properties are persistent. Once the complete set of properties is entered into the model, the resulting characteristic time may result in a classification of persistent or nonpersistent.

Approach in absence of complete data

Many of the chemicals in use today have an incomplete set of data on environmental chemical properties. In other words, research has not been completed to determine one or more of the chemical properties needed for modeling the transport and transformation of the chemical. For example, all that may be known about a chemical are two partition factors and the half-life in one medium. The CART system can be used to classify chemicals that lack a complete set of property values. A tree can be generated that uses only a selected set of property values, called predictor variables, to classify a chemical. Ultimately, a portfolio of classification trees can be generated, each created with a different set of predictor variables, thus providing the ability to classify chemicals with a range of known property values.

The steps for generating a classification tree in the absence of complete data are the same as those for generating a tree using a complete data set except for the fifth step. In this case, only the selected predictor variables and the corresponding classification calculated from the multimedia model are input into the CART program (all of the randomly selected, uncorrelated parameter values are used in the initial Monte Carlo simulation). A classification tree is generated based on only the selected predictor variables, which define the pathways for reaching the terminal nodes.

Because the full multimedia model with all of the chemical and landscape properties varied over their full range of values was used to generate the input set for the CART analysis, this tree can be used for tier 1 evaluation with any chemical in any location. The effectiveness of the tree in classifying chemicals as ''persistent'' or ''nonpersistent'' is decreased as fewer data are used in creating the tree. As a result, either the 95% confidence level of nonpersistence would need to be reduced or the number of chemicals leading to a classification of nonpersistence would be reduced or a combination of the two would occur. For this analysis, we have kept the cutoff criterion at 95%, thus reducing the number of chemicals that can be classified. However, if too few of the chemical properties are made available, CART will be unable to generate a tree that will provide useful results.

Many of the published measured values for a given chemical property differ significantly. Thus, uncertainty about these parameters can be large. In the first tier, if the range of uncertainty falls within the range of parameter values defining a nonpersistent terminal node, the chemical can be classified. Otherwise, it cannot be reliably classified. If a chemical cannot be classified using the first tier, it must be evaluated in the second tier, where persistence is determined with a Monte Carlo simulation of the uncertain parameters using a multimedia model. Only if the entire range of outcome persistence from the Monte Carlo simulation is nonpersistent (or persistent) can the chemical be reliably classified. Otherwise, more data should be collected on the uncertain parameters.

The CART trees can also be used to help prioritize data gathering. For new chemicals, a decision maker can examine the CART trees to determine which chemical properties are the most important for classifying a chemical as ''persistent'' or ''nonpersistent'' and thus which ones to evaluate first. The known chemical properties may be enough to classify the chemical. If not, the CART trees can help determine which chemical properties to evaluate next until the chemical can be adequately classified. The same process can be used with existing chemicals with unknown property values to assist in further data collection.

CASE STUDY

To demonstrate the approach, we consider five cases. In the first case, we consider a three-compartment unit world system with an emission to air and define the cutoff point between persistent and nonpersistent chemicals as a characteristic time of more than six months. We present the classification tree

that results from using all the chemical properties as predictor variables. In the second case, we consider the same scenario and limit the assumed available predictor variables to the Henry's law constant, the octanol–water partition coefficient, and the decay rate in air. The third case is another example of limited data, this time using the decay rates in air and water and the Henry's law constant. We compare the ranges of chemical properties defining nonpersistent regions between these three cases and demonstrate how this system can also be used to direct data gathering. In the fourth case, we again consider a unit world system with an emission to air, this time with a different decision rule, using a cutoff value of five years as opposed to six months. Finally, in the fifth case, we consider a source emission to water in a unit world environment, defining the cutoff point as six months and using all the chemical properties as predictor variables.

DEFINING THE CHEMICAL PROPERTIES

The chemical parameter value ranges used in the simulations are presented in Table 3. Also listed are examples of chemicals with property values near the minimum and maximum of each range. Most of the distributions are log uniform (a uniform distribution in base 10 logarithmic space), yielding the same number of simulations in each decade of the range. Using the Monte Carlo package Crystal Ball [16], 10,000 chemical realizations are developed and the persistence is calculated for each using the unit world model. These 10,000 outputs and the associated inputs were input to the CART analysis program. A commercial implementation of the CART program was used [17]. Multiple sets of 10,000 parameters yielded similar results, and thus 10,000 simulations was determined to be sufficient. Approximately 30% of the simulations yielded persistent chemicals, while 70% yielded nonpersistent chemicals.

RESULTS

In the first case, CART was applied to the results from emissions to air in a unit world environment using all the chemical and landscape properties as predictor variables, with the resulting tree shown in Figure 2. The CART algorithm and splitting rules first found that splitting the data based on whether the half-life in air was more or less than 26 d maximizes the variance reduction between the whole data set and the two resultant subsets. To read the tree, the user first asks whether the half-life in air is less than 26 d. If the answer is yes, the user follows the left branch of the tree and asks whether Henry's law constant is less than 0.41 Pa m^3/mol. If the answer is no, the user follows along the right branch and learns that 98% of the chemicals within the specified value ranges for these two properties are nonpersistent. The number of realizations in this category is indicated by *N*, in this case 4,570 realizations. The half-life in air is important because the release is to the air compartment, and thus a significant portion of the chemical can be found in this phase. With a high Henry's law constant, the chemical is more likely to remain in the air and thus is influenced by the short half-life in air (rather than partitioning to the water).

As an alternative way to display the results, the data from the tree can be translated to a two-dimensional plot depicting the regions of ''persistent'' and ''nonpersistent'' chemicals, as shown in Figure 3. The two axes are the half-life in air and the Henry's law constant. Green regions are classified as ''nonpersistent''; red regions need to go to a second tier of testing.

Table 3. Ranges of chemical properties

Property	Symbol	Distribution type	Lower end	Upper end	Example of chemicals with property at lower end of range	Example of chemicals with property at upper end of range
Henry's law constant (Pa·m^3/mol)	K_H	Log uniform	1×10^{-3}	1×10^{5}	Phenol	Nitrogen gas
Octanol–water partition coefficient	K_{ow}	Log uniform	1	1×10^{9}	Butanol, methylchloride	Di-*n*-octyl-phthalate
Decay rate in air (1/d)	k_a	Log uniform	4×10^{-4}	1×10^{2}	Toxaphene, bromodichloromethane	Benzo(*a*)pyrene
Decay rate in water (1/d)	k_w	Log uniform	1×10^{-5}	1×10^{2}	Hexachloroethane	Pyrene
Decay rate in soil (1/d)	k_s	Log uniform	1×10^{-5}	1×10^{2}	PCB	Anthracene
Vapor pressure (Pa)	*VP*	Log uniform	1×10^{-6}	1×10^{5}	Chrysene, TCDD	Atmospheric pressure
Melting point (K)	*Tm*	Uniform	100	600	Vinyl chloride	Chrysene, beta-HCH, TCDD
Diffusion coefficient in pure air (m^2/d)	D_{air}	Uniform	0.1	1.7	Hexachloroethane	2,4-dinitrotoluene
Diffusion coefficient in pure water (m^2/d)	D_{water}	Uniform	3×10^{-5}	1×10^{-4}	Endrin	Vinyl chloride

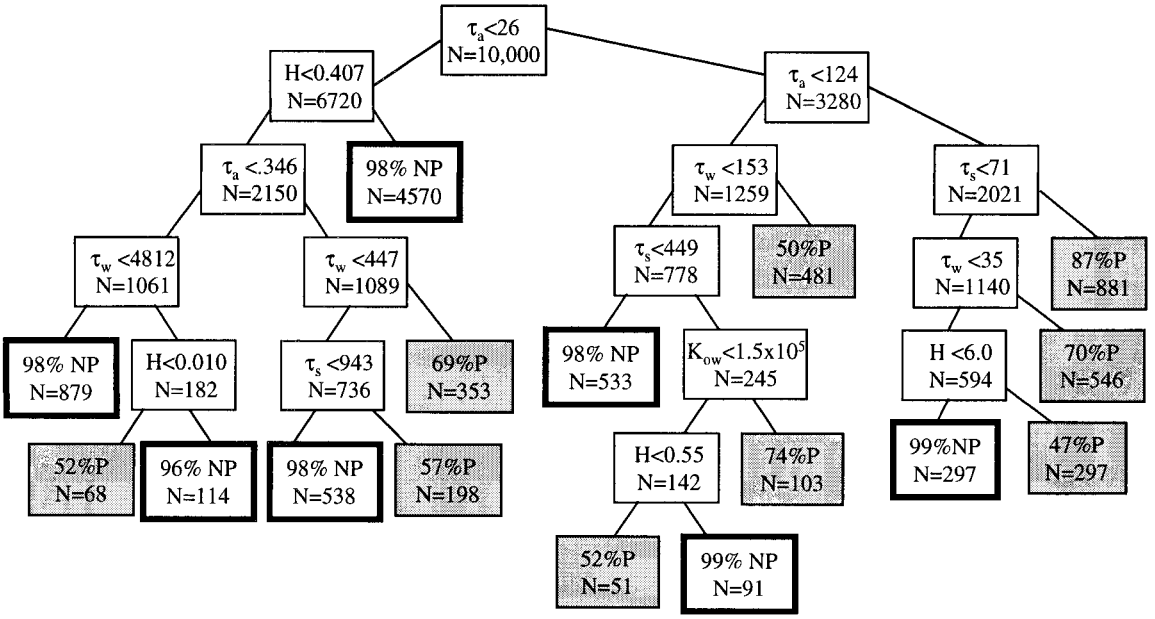


Fig. 2. The classification and regression tree (CART) for emissions to air, persistence defined as six months; all chemical and landscape properties used as predictor variables. The tree is read by asking a question regarding the inequality listed. If the answer is yes, the user follows the left branch, while if the answer is no, the user follows the right branch. The user follows this procedure until reaching a terminal node, which indicates if the chemical can be classified as ''nonpersistent'' or if the chemical needs to be evaluated in a second tier evaluation. N defines the number of chemical realizations in that node. The terminal nodes indicate the percent of the realizations that are either ''persistent,'' P, or ''nonpersistent,'' NP.

Yellow regions are ''nonpersistent'' only if they meet the additional qualifications listed in that region. If not, they are classified as ''persistent'' and need to be evaluated in the second tier of testing.

An issue of concern when interpreting CART results is that the misclassified chemicals might be near these borders; however, a systematic evaluation demonstrated that this is not necessarily the case. For example, consider the yellow region in the lower left-hand corner. While some of the false negative values have a half-life in water close to this limit, others have a high octanol–water partition coefficient and a long half-life in soil. In either case, the false negatives often have an overall half-life just over the cutoff value; thus, they will not persist much longer than the nonpersistent chemicals.

Let us consider a chemical whose reaction half-life in air is greater than 26 d. Following the right branch, the next question the user would need to ask (see Fig. 2) is whether the half-life in air is more or less than 126 d. If it is less, we follow the left branch and ask about the half-life in water. Perhaps the user does not know the half-life in water but does know the octanol–water partition coefficient and Henry's law constant. We have a situation where the user has an incomplete data set. This leads us to our second case. A new tree using only the half-life in air, the octanol–water partition coefficient, and the Henry's law constant is generated as shown in Figure 4. The results of this case are also plotted as regions of ''persistence'' and ''nonpersistence'' in Figure 5. Now the user can classify the chemical based on only the known properties. Note that in the ''nonpersistent'' region on the right side of the tree, only 95% of the chemicals are nonpersistent. This is in contrast to the tree in Figure 2, where the nonpersistent regions on the right-hand side of the tree are either 98 or 99% nonpersistent. By using fewer data (i.e., not including the half-life in soil and water), we decrease the reliability of the classifications. Also note that the red regions are larger when the complete data set is not used.

Another example of limited data is considered in the third case, using only the half-life in air, water, and Henry's law constant as predictor variables. The regional plot of persistence is displayed in Figure 6 (the classification tree is not shown for brevity). There is again a significant range of chemical property value sets that can be classified as nonpersistent based on just these three parameters. If we look at the lower centered region in Figure 3, we notice we must know both the half-life in soil and air. In Figure 6, there are now two regions. The lower one has a much more restrictive value on the half-life in water. This is the effect of not being able to use the half-life in soil in developing the classification tree.

These examples illustrate the value of complete data and two disadvantages of incomplete data. In both examples, the total number of nonpersistent chemicals requiring second tier analysis has increased. The red regions are much larger and some of the qualifications in the yellow regions are more restrictive. Thus, the trees generated with a reduced set of data are not as efficient at screening out nonpersistent chemicals. An additional disadvantage was seen in the first tree, where we were only 95% confident, as opposed to 98 or 99% confident when we used the half-life in soil and water. A reduction in the confidence of the classification may occur when fewer data are used.

The CART method can be used to help direct data collection. For example, consider a chemical that has a half-life in air of 10 d, a Henry's law constant of 0.005, and a half-life in water of 100 d. Using Figure 6, the user would discover that it is likely to be classified as persistent because the half-life in water is 100 d, exceeding the cutoff value, and will require a tier 2 analysis. When determining what parameter to determine next, the user notices that in Figure 3, the listed

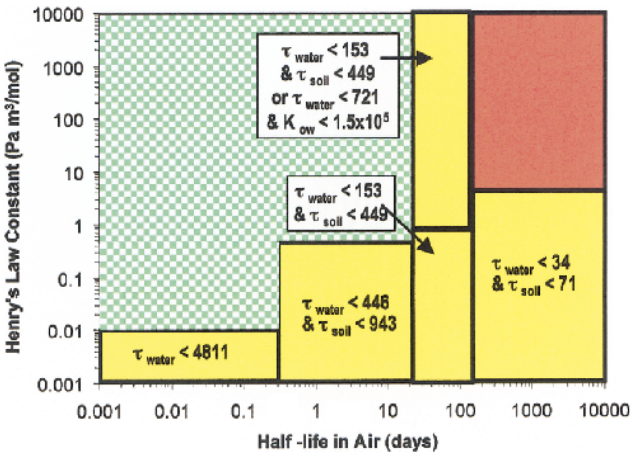


Fig. 3. Results of classification and regression tree (CART) analysis for case 1, emissions to air, persistence defined as six months. All chemical and landscape properties used as predictor variables. Green = ''nonpersistent''; red = ''persistent''; yellow = ''nonpersistent'' if the chemical meets the additional inequalities specified in the region or caption; otherwise, second tier testing is needed. Each yellow region (as defined by the black border) is separate and must meet only the specifications in that region.

parameter set falls in a yellow region with the qualifications of a half-life in water less than 447 d and a half-life in soil less than 943 d. The user should thus strive to obtain reliable data on half-life in soil to classify the chemical.

For the fourth case study, a CART tree was generated that used five years as the cutoff value for persistence and translated into the regional plot in Figure 7 (the tree is not shown for brevity). The first question the user must ask is whether the half-life in air is less than 4 d. The mean value for the transport processes from air to soil or water for the 10,000 simulations is approximately 10 d. For a substantial portion of the mass to be decayed in air, unaffected by the decay rates in soil and water, the decay rate in air must be faster than the rate of the competing transport processes from the air to the other media. In case 4, the additional qualifications on the half-life values for soil and water are approximately one order of magnitude greater than those from case 1, corresponding to the order of magnitude increase in the cutoff value. To understand the differences in property values resulting in a moderately persistent (greater than six months) and highly persistent (greater than five years) chemical, one can compare Figures 3 and 7.

A final tree was generated for emissions to water (not shown for brevity). The results are plotted in Figure 8, this time with axes showing half-life in water and half-life in air. Large regions can be classified as ''nonpersistent'' based on value ranges of these two properties. The regions of nonpersistence differ from those for emissions to air and demonstrate the importance of classifying the mode of entry of the chemical into the environmental system.

We compared the results from CART to another screening tool under consideration by the United Nations Environmental Program (UNEP) [18]. In this method, a chemical is considered persistent if either the half-life in water is greater than two months or the half-life in soil or sediment is greater than six months [18]. The chemical has the potential for long-range transport if the half-life in air is greater than 2 days. We compared the CART methodology to the UNEP criteria for a chemical being both persistent and having the potential for long-

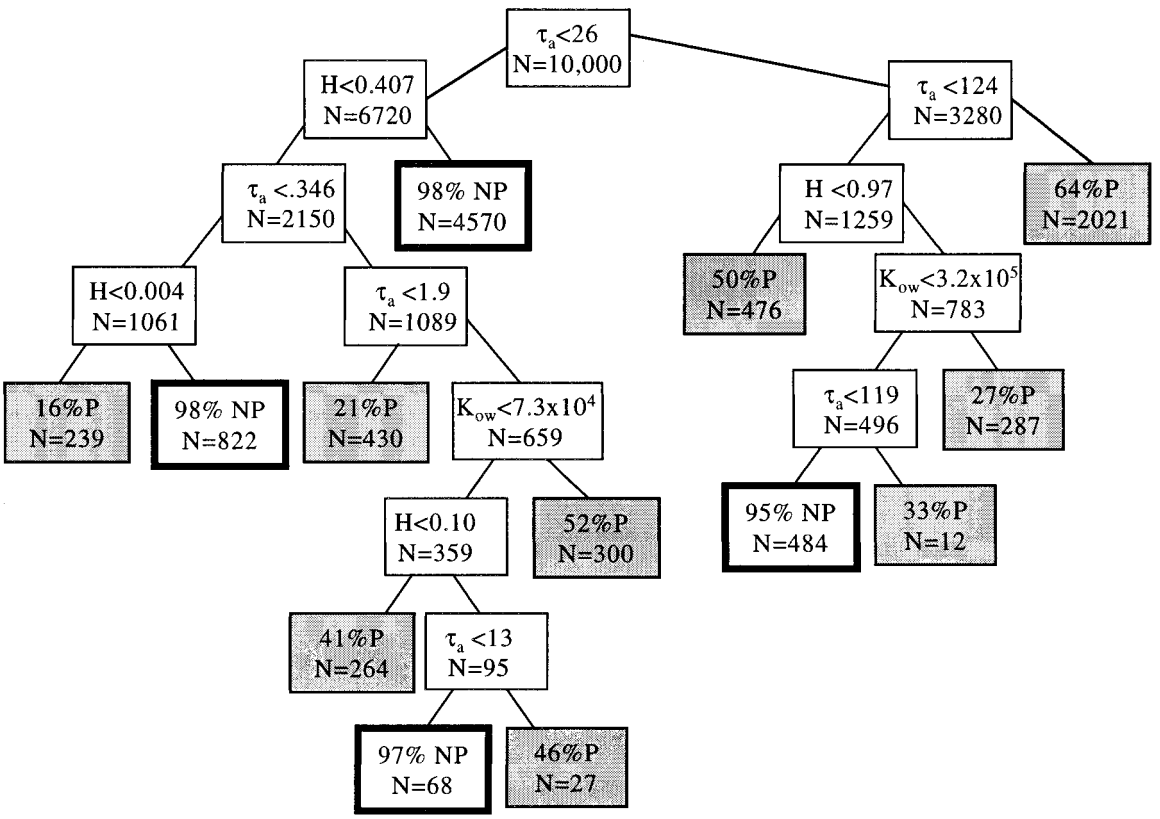


Fig. 4. Classification and regression tree (CART) for case 2, emissions to air, persistence defined as six months. Only half-life in air, octanol–water partition coefficient, and Henry's constant used as predictor variables. The tree should be read as in Figure 2.

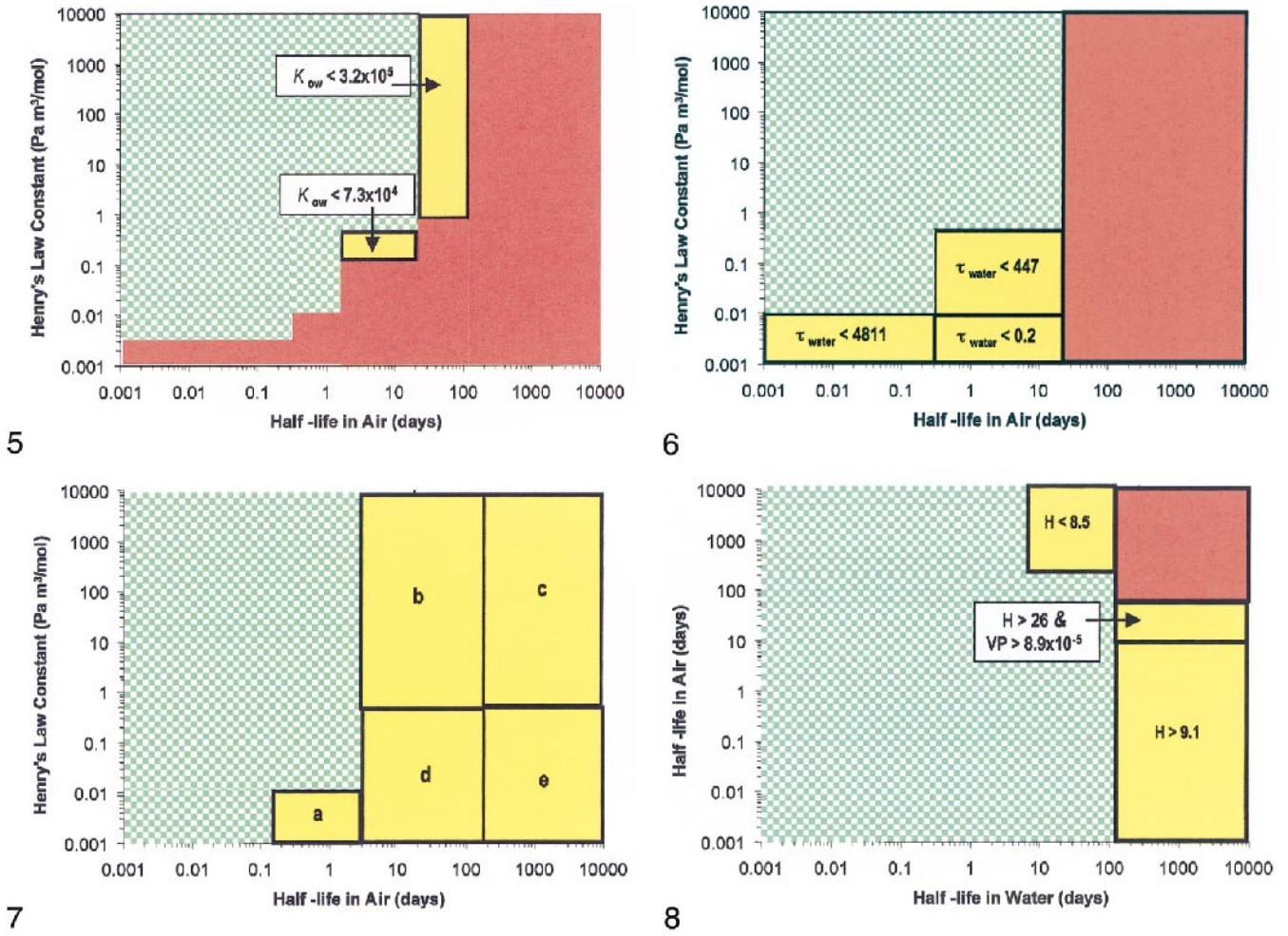


Fig. 5. Results of classification and regression tree (CART) analysis for case 2, emissions to air, persistence defined as six months. Only half-life in air, octanol–water partition coefficient, and Henry's constant used as predictor variables. The figure should be read as in Figure 3.
Fig. 6. Results of classification and regression tree (CART) analysis for case 3, emissions to air, persistence defined as six months. Only half-life in air, half-life in water, and Henry's constant used as predictor variables. The figure should be read as in Figure 3.
Fig. 7. Results of classification and regression tree (CART) analysis for case 4, emissions to air, persistence defined as five years. All chemical and landscape properties used as predictor variables. Inequalities for (**a**) $\tau_{water} < 2{,}016$, (**b**) $H > 0.54$ or $\tau_{water} < 1{,}856$ and $\tau_{soil} < 3{,}404$, (**c**) $\tau_{soil} < 1.1$ or $\tau_{water} < 1{,}856$ and $\tau_{soil} < 3{,}404$, (**d**) $H > 0.54$ or $\tau_{water} < 1{,}856$ and $\tau_{soil} < 3{,}404$ or $\tau_{water} < 406$ and $K_{ow} < 14{,}270$, (**e**) $\tau_{water} < 406$ and $K_{ow} < 14{,}270$ or $\tau_{water} < 1{,}856$ and $\tau_{soil} < 3{,}404$ or $\tau_{soil} < 1.1$. The figure should be read as in Figure 3.
Fig. 8. Results of classification and regression tree (CART) analysis for case 5, emissions to water, persistence defined as six months. All chemical and landscape properties used as predictor variables. The figure should be read as in Figure 3.

range transport. For a separate set of 10,000 chemical realizations, we calculated both the percent of false negatives and of false positives for both the CART screening method and the UNEP screening method, with the results shown in Figure 9. It is important to note that the comparisons are based on model results. The percent of both false negatives and false positives were reduced using the CART method. Additionally, we determined the percent false negative and false positive using the criteria that a chemical was either persistent or had the potential for long-range transport. In this case, less than 1% were false negatives, but well over 50% were false positives, demonstrating that these are conservative measures. The real benefit of the CART methodology is that it considers the combined effect of partitioning and half-lives in multiple environmental media.

DISCUSSION

We have proposed the use of a screening level classification system (CART) for determining if a chemical pollutant is persistent or nonpersistent. Classification of persistence is determined by comparing the overall residence time of a compound in a multimedia evaluation environment to a reference time value. The multimedia model is used with 10,000 simulations to capture a broad range of chemical-property data sets, allowing us to identify ranges of property values that result in a nonpersistent residence time in the multimedia model. We view this process as the first level of a tiered approach, sorting out those chemicals that can easily be classified as nonpersistent, thus eliminating the need for further analysis. For chemical property data sets that classify as ''persistent,'' a more detailed and complex evaluation is required at the second tier. At the second tier, persistence may be calculated with the same multimedia model used to generate the classification tree. However, it may still not be possible to classify the chemical because parameter uncertainties may prevent a reliable classification.

The ability to sort out large numbers of chemicals at the first tier provides a distinct regulatory advantage by eliminat-

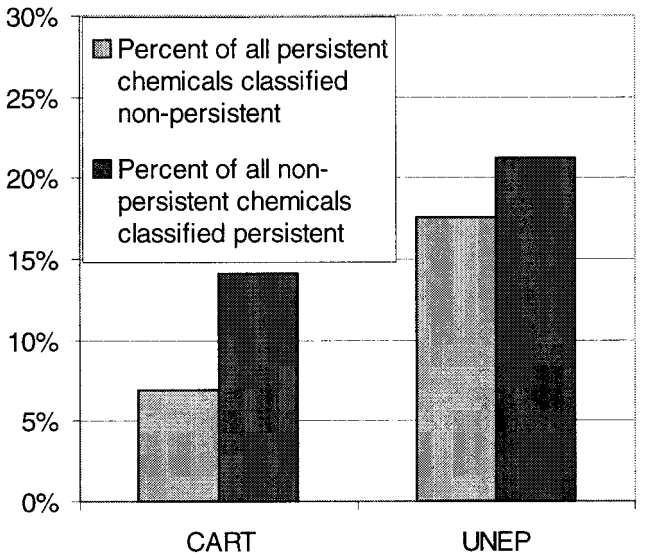


Fig. 9. The percent of all nonpersistent chemicals that were classified as persistent and the percent of all persistent chemicals that were classified as nonpersistent for both the classification and regression tree (CART) and United Nations Environmental Program (UNEP) screening methods.

ing the need to assess persistence by running a complex model for each of the chemicals under consideration. We compared the method to another classification system under consideration by the UNEP and found that, for a second set of chemical realizations, CART resulted in a similiar number of false negatives and fewer false positives. This builds confidence in the method as an appropriate method for screening chemicals for persistence.

When using a screening-level multimedia model, which has model uncertainties, we acknowledge that there is a defined reliability limit or uncertainty level within which the model can identify the overall residence time associated with a chemical. This reliability limit is sufficient for preliminary analysis. If a model such as CART were to be used for policy decisions, this reliability limit would be further evaluated. Additionally, the appropriate cutoff time for classification as persistent or nonpersistent would need to be evaluated (we used six months and five years in the case study). Finally, regulators would need to select an acceptable percent of allowable false negatives (we used 5% in the case study). As the percent of false negatives is decreased, there is a possibility that many more specifications must be made on the regions to determine if a chemical is nonpersistent. This would include converting areas of the green regions to yellow regions or more qualifications in the yellow regions, both requiring more data. The result would be that more chemicals would have to be evaluated at the second tier. These two factors must be balanced when determining the desired confidence limit.

It should be noted that there is variation in the trees generated from different random data sets generated from the same parameter distributions. The form of the tree is generally similar. However, there can be variations in the exact cutoff values. For this reason, cutoff values should only be taken as order-of-magnitude values. Variations in the additional inequalities that must be satisfied can also occur (i.e., in one simulation, the additional qualifications may specify the half-life in water and soil while another simulation results in a more restrictive qualification for just the half-life in soil). These variations tend to occur when less data is used in making a classification but the various qualifications tend to be consistent with what one would expect based on multimedia partitioning.

For the cases with emissions to air, the most important chemical parameter value needed to make an initial determination about the likely persistence is the decay rate in the air. The Henry's law constant and the decay rates in water and soil are also important parameters for classification of a chemical. The variables were treated as probabilistically independent in the cases presented in this paper; however, a simulation was completed with the half-life in water and soil correlated (using a correlation coefficient of 0.5). There were fewer false negatives because there were fewer cases with a high value for one decay rate and a low value for the other.

For both the air-emission scenario and the water-emission scenario, the classification trees developed did not depend on the values of any of the landscape parameters used in the model. We infer from this that the landscape parameters are not important for determining the persistence over the ranges of landscape and chemical property values selected in these case studies. The proportion of water relative to soil and the temperature dependency of the chemical parameters were not explicitly tested.

In addition, the CART method was applied to make persistent or nonpersistent classifications even without a full set of chemical properties. Hence, there is no restriction for the value of the unavailable property. Although the regions for which we have a robust nonpersistent classification are reduced and the percent of false negative classifications is increased in some of the remaining regions, it is still possible to make robust nonpersistent classifications for a large number of chemical property sets. There are clear limitations to using a less than complete data set. Nevertheless, the ability to use less than complete property data sets is advantageous to the decision-making process. It enables a decision maker to classify chemicals with incomplete chemical data sets. This methodology can also help direct data gathering efforts by allowing one to assess the value of new information by comparing the classification trees with and without complete data sets.

Many of the properties for which measurements have been published report values that differ significantly among researchers. Thus, uncertainty about these parameters can be large. If a parameter value is uncertain yet the entire range of values for that parameter result in the same classification, the chemical can still be classified regardless of the uncertainties. If the range of values for that parameter results in different classifications, then either more precise parameter values must be obtained or second-tier testing would be required for that chemical. In the second tier of testing, a Monte Carlo simulation can be completed to determine if the chemical can be classified as ''persistent'' or ''nonpersistent'' given the parameter uncertainty.

Perhaps the most important advantage of CART is the ability to reduce the results of a model that includes multimedia interactions to a classification diagram that does not require an understanding of multimedia models for interpretation. In many situations, such as the case study presented here, the multidimensional response surface of a model can be reduced to a two-dimensional plot with the two most important predictor parameters on the axes. The importance of other parameters shows up in the subregions of this plot. The CART method provides a graphical way of mapping inputs to outputs based on parameter importance in complex systems such as

multimedia transport models. This type of mapping provides an important link between the need among decision makers for clearly illustrated causative relationships and the need among scientists for scientifically defensible models that have the flexibility to account for a complex web of transport processes.

Acknowledgement—This work was supported in part by the U.S. Environmental Protection Agency and was carried out at Lawrence Berkeley National Laboratory through the U.S. Department of Energy under Contract DE-AC03-765F00098. U.S. Environmental Protection Agency funding was provided by the National Exposure Research Laboratory through Interagency Agreement DW-988-38190-01-0. The U.S. Environmental Protection Agency STAR Fellowship program also provided funding for D.H. Bennett.

REFERENCES

1. Scheringer M. 1997. Characterization of the environmental distribution behavior of organic chemicals by means of persistence and spatial range. *Environ Sci Technol* 31:2891–2897.
2. Bennett DH, Kastenberg WE, McKone TE. 1998. General formulation of characteristic time for persistent chemicals in a multimedia environment. *Environ Sci Technol* 33:503–509.
3. Webster E, Mackay D, Wania F. 1998. Evaluating environmental persistence. *Environ Toxicol Chem* 17:2148–2158.
4. Mullerherold U, Caderas D, Funck P. 1997. Validity of global lifetime estimates by a simple general limiting law for the decay of organic compounds with long-range pollution potential. *Environ Sci Technol* 31:3511–3515.
5. Breiman L, Freidman J, Olshen R, Stone C. 1984. *Classification and Regression Trees.* Wadsworth Publishing, Monterey, CA, USA.
6. Klein A. 1985. OECD Fate and mobility test methods. In Hutzinger O, ed, *Handbook of Environmental Chemistry.* Springer Verlag, Berlin, Germany, pp 1–28.
7. Mackay D, Paterson S. 1991. Evaluating the multimedia fate of organic chemicals—A level-III fugacity model. *Environ Sci Technol* 25:427–436.
8. Mackay D. 1991. *Multimedia Environmental Models, The Fugacity Approach.* Lewis, Chelsea, MI, USA.
9. Eisenberg JNS, Bennett DH, McKone TE. 1998. Chemical dynamics of persistent organic pollutants: A sensitivity analysis relating soil concentration levels to atmospheric emissions. *Environ Sci Technol* 32:115–123.
10. Eisenberg JNS, McKone TE. 1998. Decision tree method for the classification of chemical pollutants: Incorporation of across-chemical variability and within-chemical uncertainty. *Environ Sci Technol* 32:3396–3404.
11. Eisenberg JNS, Seto EYW, Olivieri AW, Spear RC. 1996. Quantifying water pathogen risk in an epidemiological framework. *Risk Anal* 16:549–563.
12. Pilote L, Miller DP, Califf RM, Rao JS, Weaver WD, Topol EJ. 1996. Determinants of the use of coronary angiography and revascularization after thrombolysis for acute myocardial infarction. *N Engl J Med* 335:1198–1205.
13. Tronstad R. 1995. Importance of melon type, size, grade, container, and season in determining melon prices. *J Agric Resour Econ* 20:32–48.
14. Spear RC, Grieb TM, Shang N. 1994. Parameter uncertainty and interaction in complex environmental models. *Water Resour Res* 30:3159–3169.
15. Spear RC, Bois FY, Woodruff T, Auslander D, Parker J, Selvin S. 1991. Modeling benzene pharmacokinetics across 3 sets of animal data—Parametric sensitivity and risk implications. *Risk Anal* 11:641–654.
16. Decisioneering Corporation. 1996. *Crystal Ball*, Ver 4.0. Boulder, CO, USA.
17. Steinberg D, Colla P. 1997. *CART—Classification and Regression Trees: Supplementary Manual for Windows.* Salford Systems, San Diego, CA, USA.
18. United Nations Environmental Program. 1998. UNEP Executive Body Decision 1998/2. Nairobi, Kenya.
19. Wania F, Mackay D. 1995. A global distribution model for persistent organic chemicals. *Sci Total Environ* 161:211–232.
20. Bidleman TF. 1988. Atmospheric processes. *Environ Sci Technol* 22:361–367.
21. Thibodeaux LJ. 1996. *Environmental Chemodynamics: Movement of Chemicals in Air, Water, and Soil.* John Wiley & Sons, New York, NY, USA.
22. McKone TE. 1996. Alternative modeling approaches for contaminant fate in soils: Uncertainty, variability, and reliability. *Reliability Engineering and System Safety* 54:165–181.

APPENDIX

Steady-state unit world

The following equations represent the amount of mass in each compartment in the steady-state system with air, water, and soil, including source terms to all compartments.

The following notation is used in the mass balance:

$$T_{a,total} = T_{aw} + T_{as} + k_a(1 - \Phi) \quad \text{(A1)}$$

$$T_{w,total} = T_{wa} + k_w \quad \text{(A2)}$$

$$T_{s,total} = T_{sw} + T_{sa} + k_s \quad \text{(A3)}$$

where $\Phi = Z_{ap}/Z_a$.

The equations for the mass balance in each compartment are

$$M_a T_{a,total} = M_w T_{wa} + M_s T_{sa} + S_a \quad \text{(A4)}$$

$$M_w T_{w,total} = M_a T_{aw} + M_s T_{sw} + S_w \quad \text{(A5)}$$

$$M_s T_{s,total} = M_a T_{as} + S_s \quad \text{(A6)}$$

where M_i is the mass in compartment i (kg), S_i is the source to compartment i (kg/d), and s, a, and w are the subscripts for soil, air, and water, respectively.

The mass in each compartment at steady state in terms of source terms is

$$M_a = \frac{\left(S_a + \frac{T_{wa}S_w}{T_{w,total}} + \left(\frac{T_{sa}}{T_{s,total}} + \frac{T_{wa}}{T_{w,total}}\frac{T_{sw}}{T_{s,total}}\right)S_s\right)}{\left(T_{a,total} - \frac{T_{wa}T_{aw}}{T_{w,total}} - \frac{T_{sa}T_{as}}{T_{s,total}} - \frac{T_{wa}T_{sw}}{T_{w,total}}\frac{T_{as}}{T_{s,total}}\right)} \quad \text{(A7)}$$

$$M_w = \frac{T_{aw}}{T_{w,total}}M_a + \frac{S_w}{T_{w,total}} + \frac{T_{sw}}{T_{w,total}}\frac{T_{as}}{T_{s,total}}M_a + \frac{T_{sw}}{T_{w,total}}\frac{S_s}{T_{s,total}} \quad \text{(A8)}$$

$$M_s = \frac{T_{as}}{T_{s,total}}M_a + \frac{S_s}{T_{s,total}} \quad \text{(A9)}$$

The characteristic time is defined for the unit world environment as

$$\tau = \frac{M_a k_a(1 - \Phi) + M_w k_w + M_s k_s}{M_a + M_w + M_s}$$