APPLICATION OF A GENETIC ALGORITHM TO CALIBRATE A FOOD WEB BIOACCUMULATION MODEL OF CONTAMINANTS

by

Daniel Ricard

Bachelor of Science (honors), University of British Columbia 1996

RESEARCH PROJECT SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF RESOURCE MANAGEMENT

IN THE

SCHOOL OF RESOURCE AND ENVIRONMENTAL MANAGEMENT

SIMON FRASER UNIVERSITY

July 2000

Abstract

The extent of bioaccumulation of organic contaminants in aquatic ecosystems has traditionally been assessed by two philosophically different methods. This research aimed at developing a hybrid methodology combining the benefits of the empirical and theoretical approaches. A genetic algorithm (GA) was developed to calibrate Gobas' mass-balance, steady-state model of contaminant bioaccumulation to an available set of empirical concentration data in the Lake Ontario ecosystem. The GA was found to be able to perform the complex calibration of a subset of parameters representing the structure of the aquatic food web. GA-derived feeding preferences for four Lake Ontario fish species were found to be in good agreement with independently-derived values. The feeding preferences were also used to establish the food web hierarchy throught an index of trophic position. The trophic position of the four Lake Ontario fish species derived by the GA were also found to be similar to those obtained by other research efforts. Loosening the search constraints was found to improve the ability of the GA to find the optimal solution to the calibration problem. The results suggested that PCBs and other hydrophobic organics can be used as tracers of food web structure. The main limitations of this study included the lack of explicit integration of uncertainty in the empirical contaminant data set used and the reliance on the value of the non-calibrated model parameters and equations.

Acknowledgments

My first and foremost thanks go to my senior supervisor, Dr. Frank Gobas, for continuous support throughout this research project. His insightful comments were very valuable and his thorough understanding of modeling really helped me. I would also like to thank Dr. Lou Hafer whose expertise in optimization and computing were essential to the completion of my research.

Ronnie Drever and Leslie Tse provided very insightful editing comments on the early drafts of my manuscript. Thanks also to Jon Arnot for chairing my oral defense. Many thanks to the REM administrative staff (Bev Hunter, Rhonda Keleher, Julie Bowman, and Mary-Ann Pope) for their presence and help throughout my stay in REM. Laurence Lee was always helpful when I faced computer glitches.

Thanks to all REMers. It was a real opportunity to meet you all. Merci à mes parents pour leur soutien constant lors de mes études.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The bioaccumulation factor (BAF) is widely used to express the degree of chemical accumulation in aquatic organisms. It is defined as the ratio of the concentrations of a compound in biota and water[1]. The BAF is specific to each compound, to each species, and to each ecosystem. In addition, the BAF can vary in a given system as a function of the environmental conditions (e.g. seasonal changes). It is used in various regulations that manage chemical use, in particular, in the development of water and sediment quality criteria, in hazard and exposure assessment, and in evaluating remediation scenarios.

Two philosophically different approaches are used to derive BAFs. The empirical approach derives BAFs from measured concentrations of contaminants in organisms and in water. The applicability of this method is often limited because of the lack of data for different compounds, species, and ecosystems. Also, this method does not provide information about how the BAFs might change under different environmental conditions, chemical emissions, and changes in food web structure.

The other approach is to use theoretical models to derive BAFs. Bioaccumulation models (BM) use knowledge of chemical transport and uptake kinetics to quantitatively predict internal contaminant concentrations in different living organisms of an ecosystem. Models are popular since they can be used for a variety of species, compounds, and ecosystems and do not require site-specific and organism-specific concentration data. Their strength lies in the understanding of the mechanisms controlling the distribution of contaminants in biota. Their weakness lies in the uncertainties in the form of the model equations and in the availability of appropriate data to parameterize the model.

In many practical exposure assessments of contaminants, there is often insufficient data

[1] $BAF = Conc_{biota}/Conc_{water}$

available to satisfactorily apply either an empirical or a theoretical approach to assess the BAFs. It is equally common that a limited data set of empirical concentration data and model input data does exist. It is the aim of this research to investigate the application of genetic algorithms to calibrate a theoretical bioaccumulation model (which lacks crucial input data) using available empirical data (which may also be incomplete). The new, calibrated model combines the attributes of theoretical bioaccumulation models and of available empirical data sets.

The BM developed by Gobas [11] was chosen since it has gained general scientific acceptance and is now used by the US Environmental Protection Agency (EPA) to estimate BAFs of hydrophobic organic compounds under the Great Lakes Water Quality Initiative [5]. The model predicts internal contaminant concentrations through a first-order kinetic mass-balance equation (Equation A.5) that includes the main routes of uptake and elimination to and from the fish tissues. It also captures food web biomagnification by allowing multiple feeding interactions between organisms. The Lake Ontario freshwater ecosystem was investigated in this study.

When facing a situation where certain model parameters are unknown or uncertain, calibration is often applied to deduce model parameters from an available empirical data set. Among the most uncertain BM parameters are those determining the structure of the food web. The diet preferences of each species have to be explicitly specified for the model calculations [11]. However, this information is difficult to determine experimentally. The calibration of the food web structure parameters is analogous to resolving the food web structure using the contaminants as tracers of food web interactions.

In the original BM application to the Lake Ontario ecosystem, the food web structure is determined by a total of ten trophic links (Figure 1.1). The default model calibration for the foodweb structure was derived from the literature [9]. The parameter values were calculated by model calibration to an available data set of hydrophobic organic contaminants [30]. The calibration results were compared to independently-derived values of foodweb structure. Agreement between the two sets of foodweb structure parameter values would indicate that the data set contained the information necessary to resolve the foodweb structure. The new, calibrated model still maintained its predictive ability and, moreover, is consistent with available data.

Organic contaminants such as PCBs are ubiquitous in the environment and their distribution in the different organisms of a food web is highly dependent on trophic interactions [2]. Their distribution will likely be affected by changes in feeding habits or species inva-

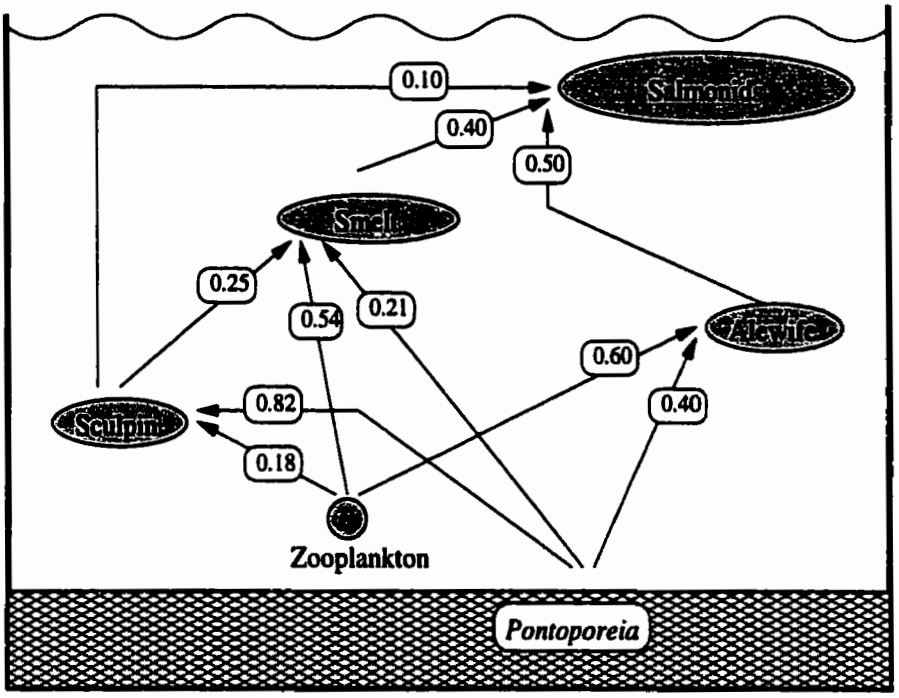

Figure 1.1: Hypothesized structure of the Lake Ontario food web in the late 1980s as derived by Flint [9] and subsequently used by Gobas [11]. The different trophic links and the proportion that a prey item represents in an organism's diet are shown for the different fish species.

sion leading to a different food web structure. Just as stable isotopes [1, 21], PCBs could provide time-integrated measures of feeding habits.

Simple models with few parameters can be calibrated using traditional methods such as minimum least squares or maximum likelihood. However, the BM is complex and the number of parameters to calibrate can be large. Traditional calibration methods are ill-suited for calibration of the BM. In this study, the genetic algorithm (GA) was investigated as a possible calibration method since it was found to be a robust and adaptable optimization algorithm in many comparable situations [14, 37, 20, 10, 31, 29]. The calibration results of the BM food web structure parameters were compared to independently derived parameter values [9, 35].

The objectives of this study were to:

1. investigate the applicability of the GA to develop a site-specific BM based on incomplete data set

2. evaluate the possibility of resolving food web structure information using chemical residue data in the different fish species of the Lake Ontario ecosystem through the calibration of the BM

3. guide future research on the use of advanced numerical methods in bioaccumulation modeling

There are many other examples where the methodology developed here is applicable. It can be applied to other modeling applications and their associated data set. The parameters to be calibrated, usually the most uncertain ones and those that the model is most sensitive to, will vary from application to application. The methodology provides a mathematically sensible way to calibrate a variety of parameters in complex modeling applications.

Chapter 2

Methods

The calibration of the BM first required its implementation in a programming language. A subset of model parameters to be calibrated was identified. A GA was developed to evolve a population of potential solutions based on the fitness of each individual. The fitness of each individual was calculated through an objective function that expresses the level of agreement between the predicted and observed concentration data for 44 different chemicals. The value of the different GA control parameters dictated its performance in solving the optimization problem. It was therefore important to choose them appropriately to calibrate the BM. It was also essential to ensure that the GA had enough time to converge towards a solution approaching the minimum value of the objective function. Finally, search constraints were implemented for the different calibrations. Four separate calibrations using different objective functions and search constraints were performed. The different calibration results were evaluated through an independent perfomance measure of the BM and were compared to independently derived parameter values.

2.1 Implementation of the bioaccumulation model

The BM was programmed in the Quick Basic language using the Microsoft QuickBasic 4.5 compiler [27]. The BM's mass-balance equations can be found in Appendix A.1 and its different parameters can be found in Appendix A.2.

2.2 Selection of bioaccumulation model parameters for calibration

The BM contains many parameters (Appendix A.2) and the methodology presented here could be used to calibrate any of them. Since this project examines the applicability of the GA to calibrate the BM, a subset of uncertain parameters was chosen to represent a fairly complex calibration problem.

While some model parameters have values that are easily obtainable, others are difficult to measure. For example, the average weight and lipid content of a biological species can be relatively easily determined. On the other hand, the diet composition of a fish species is more uncertain because it is more difficult to measure. Existing methodologies to determine food web structure have significant limitations. The trophic interactions used in the original BM [11] are therefore in doubt.

For the purpose of this study, it was decided to explore the application of the GA to calibrate the feeding preferences which represent the trophic structure of the food web. The food web linkages are represented by ten different parameter values (Table 2.1 and Figure 2.1). The food web structure parameters are also an interesting subset because they allow the use of the calibration results to test whether PCBs can be used as tracers of food web structure. Assuming that the BM is adequate in predicting the chemical concentrations of PCBs in the different fish species, the calibration results may be able to elucidate the food web structure that best explains the data observed. Those parameters are not the only uncertain parameters used in the model but their calibration would represent an interesting alternative to existing ecological methods for resolving food web structure.

The food web structure was represented by a vector containing the 10 parameter values:

$$\Theta = \begin{cases} \text{diet composition of sculpin} & P1 & P2 \\ \text{diet composition of smelt} & P3 & P4 & P5 \\ \text{diet composition of alewife} & P6 & P7 \\ \text{diet composition of salmon} & P8 & P9 & P10 \end{cases}$$

$$\Theta = \{P1, P2, P3, P4, P5, P6, P7, P8, P9, P10\}$$

The parameters were defined on a scale from 0 to 1 at a precision of ± 0.01. When viewed as a stream of 10 numbers, the structure of the food web is difficult to interpret. An

Table 2.1: Fish species considered in the BM, their prey items, and the parameters representing their corresponding feeding preferences.

Predator	Prey Item	Parameter
Sculpin	Zooplankton	P1
	Pontoporeia	P2
Smelt	Zooplankton	P3
	Sculpin	P4
	Pontoporeia	P5
Alewife	Zooplankton	P6
	Pontoporeia	P7
Salmon	Sculpin	P8
	Smelt	P9
	Alewife	P10

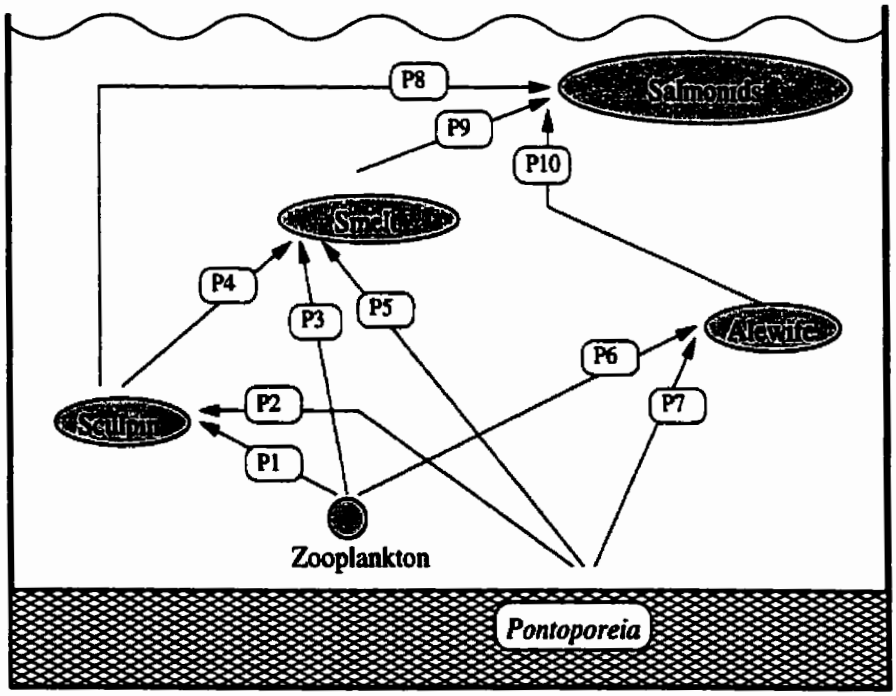

Figure 2.1: The ten parameters in the BM that describe the food web structure.

index of trophic position was used to assess the feeding hierarchy in the ecosystem [22, 35]. Assigning a subjective trophic position to the two lowest prey species in our model of Lake Ontario allowed the calculation of the trophic position of the different fish species using the formula:

$$T_a = [\sum (P_i * T_i)] + 1 \tag{2.1}$$

where,

T_a is the trophic position of fish species a

P_i is the volumetric contribution of prey item i to predator a

T_i is the trophic position of prey item i

In the simplified food web, the zooplankton were assigned a trophic position of 3, the benthic invertebrates a trophic position of 2, and the primary producers a trophic position of 1. This index was used to calculate the trophic position of each fish species. For example, a fish species whose diet consisted of 50% zooplankton and 50% benthic invertebrates had a trophic position of $[(50\% \times 3) + (50\% \times 2)] + 1 = 3.5$. The index allows comparison of the trophic position between different fish species. The availability of independently collected trophic position data provided an opportunity to evaluate the model calibration data [9, 35].

2.3 Selection of forcing variables

In its original application to the Lake Ontario ecosystem, the BM predicted contaminant concentrations in different organisms, including zooplankton and benthic organisms. The predicted concentrations in zooplankton and benthic organisms were then used in the calculations for the predatory fish species. Since the parameters in Θ only affect the predicted values of the concentration in the fish species, the BM can be simplified so that empirical values, instead of predictions, of concentration in zooplankton and in benthic organisms can be used in the calculations for fish species. That is:

$$Conc._{predicted} = f\begin{pmatrix} C_w, C_s, K_{ow} \\ \vec{\rho}, \Theta \end{pmatrix} \tag{2.2}$$

can become:

$$Conc._{predicted} = f\begin{pmatrix} C_w, K_{ow} \\ Conc._{zooplankton} \\ Conc._{Pontoporeia} \\ \vec{p}, \Theta \end{pmatrix} \tag{2.3}$$

where,

$Conc_{predicted}$ is the predicted internal concentration in a given species ($\frac{ng}{g}$ wet weight)

$f(...)$ is the BM's mathematical equations (Appendix A.1)

C_w is the concentration in water ($\frac{ug}{L}$)

C_s is the concentration in sediments ($\frac{ng}{g}$ dry weight)

K_{ow} is the compound's octanol-water partition coefficient (unitless)

$Conc._{zooplankton}$ is the observed concentration in zooplankton ($\frac{ng}{g}$ wet weight)

$Conc._{Pontoporeia}$ is the observed concentration in *Pontoporeia* ($\frac{ng}{g}$ wet weight)

$\vec{p}$ are the BM parameters (Appendix A.2)

We will refer to the BM form in equation 2.2 as $BM_{predicted}$ and that in equation 2.3 as $BM_{empirical}$. This new model form eliminates potential errors stemming from predictions of $Conc._{zooplankton}$ and $Conc._{Pontoporeia}$. The BM's predictions for zooplankton and benthic species have been refined since its original application to Lake Ontario [28].

2.4 Implementation of the genetic algorithm

To have complete flexibility over the calibration of the BM, a GA-based optimizer was developed and implemented in the Quick Basic language using the Microsoft QuickBasic 4.5 compiler [27]. While some off-the-shelf GA software applications are available, they lack the expandability and adaptability required for this project.

The GA designed included a mutation operator, a 2-point crossover recombination operator [33], and a reproduction operator. The reproduction operator used the elitist strategy and standard fitness [4] (see Appendix B). Potential solutions were encoded on adjacent

genes stored on binary haploid chromosome. Parameter values were encoded in binary-reflected Gray code [17, 18]. In each generation, raw fitness values for each individual in the population were first calculated by an objective function. Individuals were then ranked and their standard fitness was calculated.

Two different objective functions were used to measure the goodness-of-fit of the BM to the data. Both were sums of squared residuals, the first used untransformed values of predictions and observations while the second used their natural logarithm. The objective functions ($S1$ and $S2$) are:

$$S1 = \left[\sum_{j=1}^{n} w_j \left[\sum_{i=1}^{m} \left(Conc._{pred._{i,j}} - Conc._{obs._{i,j}}\right)^2\right]\right] \tag{2.4}$$

$$S2 = \left[\sum_{j=1}^{n} w_j \left[\sum_{i=1}^{m} \left(log_e\left(Conc._{pred._{i,j}}\right) - log_e\left(Conc._{obs._{i,j}}\right)\right)^2\right]\right] \tag{2.5}$$

where,

$Conc._{pred._{i,j}}$ is the predicted concentration of compound j in fish i

w_j is the weight given to compound j

m is the number of fish species

n is the number of compounds

All compounds were given an equal weight (i.e. $w = 1/n$).

Standard fitness was calculated by a half normal distribution equation that was normalized to 1 [4]:

$$fitness(\text{rank i individual}) = \frac{\exp\left[-\frac{(\text{ranked position i})^2}{2\sigma^2}\right]}{\sum_{j=1}^{\#\text{individuals}} \exp\left[-\frac{(\text{ranked position j})^2}{2\sigma^2}\right]} \tag{2.6}$$

The standard deviation in the normal distribution (σ), dictates the evolutionary pressure. A high value of σ yields a flat distribution and the pre-assigned fitness values do not differ much from individual to individual. The best individual still has a larger share of the total population fitness but its proportional advantage is reduced as σ increases. Conversely, a small σ value gives a strong bias towards the best individual in a given generation.

Chemical residue data in different organisms of the Lake Ontario food web were used for the BM calibration [30]. The subset of compounds used had data points for each fish species, for zooplankton, for *Pontoporeia*, and had $\log K_{ow} > 5.0$. The $\log(K_{ow})$ of the different compounds were obtained from [12] and [16]. The $\log(K_{ow})$ of γ-chlordane was calculated according to methods described in [25] using a free online version of ESC's Estimation Software [8]. In total, 44 compounds were used for the calibration of $BM_{empirical}$ (please refer to Appendix E for the data).

The calibrated parameters ($\Theta = \{P1, P2, P3, P4, P5, P6, P7, P8, P9, P10\}$) were searched under two different sets of constraints. Under the first constraints scenario, each parameter was searched on a $[0, 1]$ interval at a precision of ± 0.01. The sum of diet preferences for each fish species equaled 1.0. The default calibration ($\theta_{Gobas93}$) fulfilled this constraint since the sum of diet preferences for each fish species was equal to 1.0.

It was already established that the parameter values are uncertain and that not all trophic links in the ecosystem were accounted for. To address this uncertainty, a second constraints scenario was evaluated where each parameter was searched on a $[0,2]$ interval, still at a precision of ± 0.01. In this case, the sum of diet preferences was allowed to vary between 0 and 2. Allowing the diet of a fish to break the strict constraints was similar to varrying the level of contaminants input from the organism's food. The constraints were then:

1. Strict constraints

   $0 \leq P_i \leq 1$

   $P1 + P2 + P3 = 1$

   $P4 + P5 + P6 = 1$

   $P7 + P8 = 1$

   $P9 + P10 = 1$

2. Under loose constraints

   $0 \leq P_i \leq 2$

   $0 \leq P1 + P2 + P3 \leq 2$

   $0 \leq P4 + P5 + P6 \leq 2$

   $0 \leq P7 + P8 \leq 2$

   $0 \leq P9 + P10 \leq 2$

If a potential solution was found to break the constraints, it was penalized to reduce its chances of surviving the next generation. To implement this, a penalty function was added to the objective function ($S = S + penalty$) [14, 26]. The penalty needed to be large enough to ensure the nonoptimality of solutions breaking the constraints [29]. Penalty weights (ω) of 10000 for calibration using $S1$ and of 100 when using $S2$ were used in a penalty function of the form:

$$penalty = \omega * \sum_{j=1}^{n} \left| 1 - \left(\sum_{i=1}^{m} P_i \right) \right| \tag{2.7}$$

where,

ω is the penalty weight

P_i is the contribution of prey item i to predator j

m is the number of prey species

n is the number of fish species

The loose constraints used the same penalty function but were implemented as:

$$penalty = \begin{cases} 0 & if \quad \sum_{j=1}^{\#prey} P_i \leqslant 2.0 \\ penalty & if \quad \sum_{j=1}^{\#prey} P_i > 2.0 \end{cases} \tag{2.8}$$

The population was randomly initialized at the beginning of each run, which meant that solutions were very likely breaking the constraints in the earlier part of the evolution. The penalty function ensured that solutions which did not fulfill the constraints were at an evolutionary disadvantage and that their chance of surviving to the next generation were accordingly reduced.

The trophic position of the different fish species (Equation 2.1) was affected by the change in constraints. For example, a sum of diet preferences larger than 1 artificially elevates an organism's trophic position. When conducting calculations under loose constraints, it was necessary to adapt Equation 2.1 so that trophic position accounted for the possibility of a sum of diet preferences different than 1.0. This correction returned a trophic position that took into account the relative contribution of the different prey items instead of their absolute value. The corrected realized trophic position was calculated as:

$$corrected T_a = \left[\sum (P_i * corrected T_i) / \sum P_i\right] + 1 \qquad (2.9)$$

where,

$corrected T_a$ is the corrected trophic position of fish species a

$corrected T_i$ is the corrected trophic position of prey item i

Just as the contaminant predictions in the BM, the trophic position calculations have to be performed in an order that respects the hierarchy in the food web. Trophic positions are first calculated for alewife and sculpin since they feed on the two lowest prey items, namely benthic organisms ($T_{pontoporeia} = 2$) and zooplankton ($T_{zooplankton} = 3$). The trophic position of smelt can then be calculated using $corrected T_{sculpin}$, along with $T_{pontoporeia}$ and $T_{zooplankton}$. Finally, the trophic position of the salmonids can be established using $corrected T_{sculpin}$, $corrected T_{alewife}$, and $corrected T_{smelt}$.

2.5 Debugging and testing of computer code

The code of the BM was verified by comparing its output to results published in [11].

To ensure the GA was correctly implemented, it was first used to calibrate different linear and non-linear functions of increasing complexity. The calibration of such models can be performed by normal regression analysis, hence providing a benchmark for testing the results obtained by the GA. Discrepancies between GA calibrations and the benchmark calibrations provided information necessary for the debugging of the computer code. At that point, the control parameters used varied in values, a situation that led to the meta-level optimization (Section 2.6).

Calibration of the parameters α and β in a linear model of the form $\gamma = \alpha x + \beta$ was the first GA calibration undertaken. Fifty data sets were generated using different values of α and β. Each data set consisted of 50 data points and were used to test the GA. After the GA showed its ability to calibrate a linear model using the "synthetic" data, a random normal deviate was added to the data and the GA was used to calibrate the same model using the "noisy" data.

The calibration of non-linear models was also conducted. The model forms used to calibrate are taken from [32]. Synthetic data was generated and used to test the GA's ability to

correctly calibrate the different non-linear models. The three non-linear models calibrated are:

1. $f(x) = 1 - (x^{-\alpha})$
2. $f(x) = (\alpha x) / (\beta + x)$
3. $f(x) = \alpha / [1 + \exp(\beta - \gamma x)]$

The final test of the GA before applying it to the BM was a calibration of a non-linear model using empirical data. The calibrated model is a Michaelis-Menten type equation:

$$y = [(\alpha x) / ((\alpha/\beta) + x)]$$

The data and analyses of Drever [7] were used as a benchmark for calibration of the α and β parameters. Drever used the nonlinear regression package from SPSS [34].

2.6 Meta-level optimization of the genetic algorithm control parameters

The GA has four control parameters; the number of individuals in the population, σ in the standard fitness calculation, the probability of recombination and the probability of mutation. The value of the GA control parameters affect its ability to solve a given optimization problem. For example, if the probability of mutation is too high, the GA will behave just as a random search and will not explore the promising regions of the search space. The number of individuals in the population controls the amount of search space that gets sampled over a run. The other three control parameters, dictate the GA's level of exploration/exploitation. For this reason, it was important to test different GA configurations for BM calibration. A population size of 50 was deemed sufficient. The population size was not explicitly evaluated. It was necessary to include σ as part of the meta-optimization because its appropriate value is dependent on the optimization problem [4]. This optimization of the GA control parameters is called meta-level optimization [15].

The meta-level optimization of the GA control parameters was performed using a synthetic concentration data set. This data was generated by running the BM with a set of default parameters (Table A.1, Appendix A.2) for the 44 compounds used (Appendix E).

Since the data used were synthetic, the answer sought by the GA was known in advance. The idea was to calibrate the model using different GA configurations and to evaluate which configuration performed better.

The appropriate probability of mutation for calibrating the BM was found by evaluating 6 different values (0.001, 0.005, 0.01, 0.05, 0.1, 0.25). Six different values of the probability of recombination were also evaluated (0.0, 0.2, 0.4, 0.6, 0.8, 1.0). Finally, σ (evolutionary pressure) in the standard fitness operator was evaluated for 5 different values (3, 5, 10, 25, 50).

The meta-level optimization used the sum-of-squared residuals (equation 2.4) as the objective function. Each of the 180 combinations of configuration parameters was evaluated through 10 runs of 200 generations. Each optimization routine used the same seed for initializing the random number generator to insure that the differences in performance were attributable to the configuration of the GA and not to stochasticity. The mean value of S_{min} after 10 runs was used as an indicator of the ability of the GA configuration to calibrate the BM.

2.7 Number of generations

If the GA can solve the optimization routine, the value of the objective function is expected to asymptotically reach a minimum value as the number of generations becomes larger. To estimate the number of generations neccessary to approach the minimum value of the objective function (hereafter referred to as the "stabilizing number of generations") required another meta-level analysis.

The GA used the data in Appendix E to calibrate the BM through 10 runs of an increasing number of generations (25, 50, 100, 200, 300, 500, 1000, 2000, and 5000 generations). The GA configuration found by the meta-level optimization of the GA control parameters (section 2.6) and strict constraints were used for those runs. The minimum value of the objective functions ($S1_{min}$ and $S2_{min}$) were not known in advance. The GA was therefore allowed to run for a large number of generations to establish if it converged to a solution approaching $S1_{min}$ and $S2_{min}$. The number of generations required for the GA to asymptotically approach $S1_{min}$ and $S2_{min}$ was subsequently used for the different calibrations.

2.8 The calibration routines

The problem statement for the calibration of the BM's food web structure parameters may be expressed as: find Θ such that $S(\Theta)$ is minimized.

Forty GA runs of the stabilizing number of generations (section 2.7) were performed for four calibration routines:

1. Calibration of $BM_{empirical}$ by minimization of $S1$ (equation 2.4) under strict constraints
2. Calibration of $BM_{empirical}$ by minimization of $S1$ (equation 2.4) under loose constraints
3. Calibration of $BM_{empirical}$ by minimization of $S2$ (equation 2.5) under strict constraints
4. Calibration of $BM_{empirical}$ by minimization of $S2$ (equation 2.5) under loose constraints

The solution resulting in the lowest value of the objective function after 40 runs was considered as the solution to the calibration problem. The combined set of 40 GA solutions were statistically investigated to evaluate the robustness of the algorithm at properly calibrating the BM.

2.9 The performance measure

To evaluate the effects of calibration on the BM, an independent metric of the model's predictive capability was required. The mean and standard deviation of the natural logarithm of the ratio of predicted and observed concentrations was chosen as the performance measure of the model's predicitive ability:

$$\overline{\log_e MB} = \left[\sum_{j=1}^{n} \left(\sum_{i=1}^{m} \log_e \left(\frac{Conc._{pred.i,j}}{Conc._{obs.i,j}} \right) \right) / m \right] / n \tag{2.10}$$

$$\sigma_{\log_e MB} = \sqrt{\left[\sum_{j=1}^{n} \left(\sum_{i=1}^{m} \left(\log_e MB - \overline{\log_e MB} \right)^2 \right) \right] / (p-1)} \tag{2.11}$$

where,

m is the number of fish species

n is the number of compounds

p is the number of data points (i.e. $p = m * n$)

This metric was useful because it provided an all-encompassing measure of the model bias. The systematic over or under-prediction of empirical concentration data can stem from many sources, including measurement errors in contaminant concentrations, natural variability within each fish species, value of model parameters, and form of model equations. A negative $\overline{\log_e MB}$ meant that, on average, our model systematically underpredicted the concentrations in fish. A positive value pointed to systematic overprediction, and a value of 0 meant that, on average, our model predictions were consistent with observed data. The results from equation 2.10 are referred to as "mean model bias". The standard deviation of the model bias (equation 2.11) indicated the variability around the mean model bias. The fish-specific value of these metrics were also used to evaluate the fit of the BM for each fish species:

$$\overline{\log_e MB_i} = \left[\sum_{j=1}^{n} \log_e \left(\frac{Conc._{pred.i,j}}{Conc._{obs.i,j}} \right) \right] / n \tag{2.12}$$

$$\sigma_{\log_e MB_i} = \sqrt{\left[\sum_{j=1}^{n} \left(\log_e MB - \overline{\log_e MB_i} \right)^2 \right] / (n-1)} \tag{2.13}$$

Chapter 3

Results and discussion

The results from the debugging and meta-level optimization of the GA control parameters are presented and discussed first. The calibration results of the food web parameters are presented and discussed in the later part of this chapter.

3.1 Debugging and testing

Calibration results of the linear models by the GA were similar to those obtained by regression analysis for both the "synthetic" and "noisy" data sets. The GA also properly calibrated the different non-linear functions with either one, two, or three parameter using both synthetic and noisy data. GA results for the calibration of the Michaelis-Menten model using data from [7] were similar to those obtained by SPSS.

These results were expected from previous research efforts as the GA was successfully used to calibrate a variety of linear and non-linear models [10, 14, 29, 31, 37]. In certain cases, the GA solution yielded a better fit to the available data than the regression model. The analysis of Drever [7] used the nonlinear regression package from the SPSS statistical software [34]. The non-linear least square estimates of parameter value is implemented by the Marquardt algorithm [23]. The results returned by such an algorithm are only approximate since the problem does not have an analytical solution and the calibration can only be performed by search [3]. Our results suggest that the GA calibration of the linear and non-linear models was comparable to that obtained through traditional methods. The results proved that the GA computer code was operating as intended.

Table 3.1: GA configuration derived from meta-level optimization of GA control parameters. The three control parameters considered were the probability of recombination, the probability of mutation, and σ.

GA control parameter	Optimal value
p(recombination)	0.4
p(mutation)	0.01
σ	5

3.2 GA control parameters

The meta-level optimization results can be found in Tables C.1, C.2, C.3, C.4, and C.5. The results are further illustrated in Figures C.1, C.2, C.3, C.4, and C.5. The meta-level results were interpreted in a qualitative manner. For the different σs, the lowest mean values of $S1$ consistently occurred at intermediate levels of p(mutation) and p(recombination). The GA configuration that was deemed best at solving the calibration problems and that was subsequently used for the BM calibrations can be found in Table 3.1.

None of the runs performed during the meta-level optimization of the GA control parameters successfully found the known optimal value of Θ. This is probably due to the fact that all meta-level runs evolved only 200 generations. As will be noted in section 3.3, the GA requires a larger number of generations to converge to the optimal solution. In addition, the objective function used was S1 (equation 2.4) which made the calibration biased towards fitting the large concentration data points. However, the results of the meta-level optimization still stand since they allowed us to evaluate how the different GA configurations compared when solving the optimization problem. The GA configuration that was found to best calibrate the BM exhibited intermediate levels of exploitation and exploration. A probability of mutation of 0.01 appeared to be low enough not to make the GA act in a near-random manner. At this level, the operator is expected to perform, on average, one mutation every other generation. The level of disturbance in evolving the GA population was kept at a level that allowed new areas of the search space to be sampled while allowing the algorithm to converge towards promising regions of the search space by allowing recombinations to occur between reproducing individuals at a probability of 0.4. The value of σ found to optimize the GA's search ability was also at an intermediate value of its documented range of application [4].

Unlike the work done by Grefenstette [15], the meta-optimization performed here did

not evaluate other GA parameters such as population size, selection strategy, and generation gap. The value of those general GA control parameters were based on the results in [15] and from empirical results in many GA applications [33, 26, 14, 29] which showed that a population size of 50, the elitist strategy, and generation gap of 1.0 were appropriate. Grefenstette's study aimed at optimizing the performance of the GA through the systematic evaluation of GA configurations. The performance of the GA is a measure of how fast it finds a solution to an optimization problem. The GA application presented in this study did not aim at making the algorithm fast but was more concerned with making it robust. The meta-level optimization was aimed at finding the GA configuration that strikes the right balance between exploration of the search space and exploitation of promising solutions. These two aspects of the search are usually dictated by the mutation and recombination operators and by σ. The goal of the meta-level optimization was not to find the GA configuration that finds the solution fastest but rather to isolate the configuration whose exploitation/exploration balance was best-suited to calibrate the food web structure paramaters of the BM.

3.3 Number of generations

The value of the objective functions was found to decrease as the number of runs increased, as seen in Figures 3.1 and 3.2. Both objective functions asymptotically approached a minimum value as the number of generations increased. The best solution out of 10 runs doesn't improve much after 1000 generations but the spread between the best and worst solutions narrowed as the evolution time increased, especially for $S2$.

When using $S1$ as the objective function, the GA's best answer out of 10 runs yielded a better fit than $\Theta_{Gobas93}$ after only 25 generations. After 200 generations and more, the worst solutions found by the GA yielded a better fit than the default calibration (Table C.6). The results were similar when using $S2$. The best answer found by the GA needed 100 generations to yield a better fit than the default calibration. However, the GA's worst solutions yielded a better fit than the default calibration only after 5000 generations (Table C.7).

The constraints were implemented through a penalty function which meant that solutions breaking the constraints were weeded out of the evolving population. The beginning of the evolution had many solutions breaking the constraints since the population was randomly initialized. As the run progressed, the GA solutions tended to increasingly fulfill

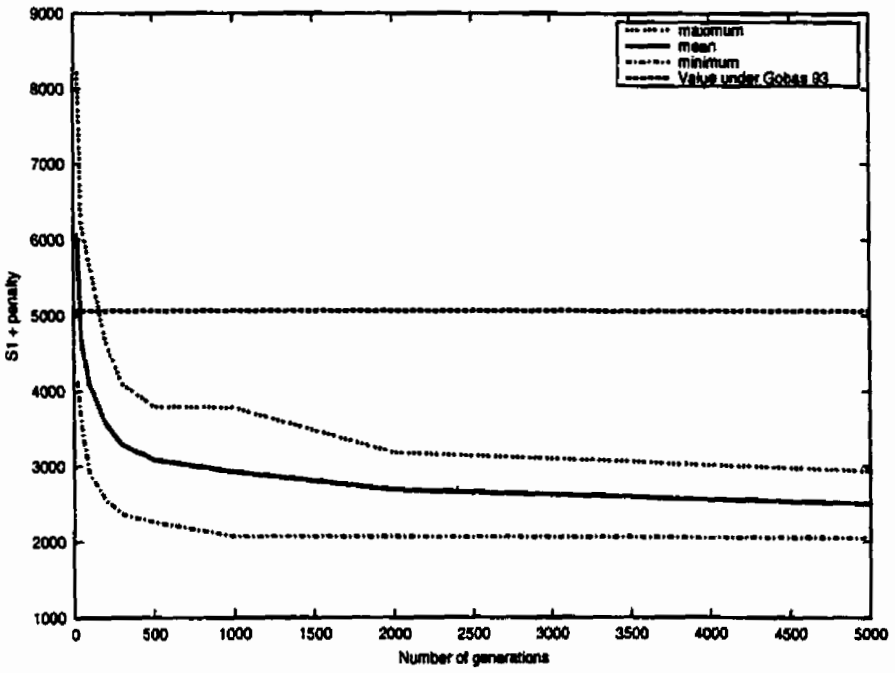

Figure 3.1: Minimum, mean, and maximum value of ($S1 + penalty$) out of 10 runs versus number of generations per run. The value of $S1$ under $\Theta_{Gobas93}$ is also shown. The data used to generate this Figure can be found in Table C.6.

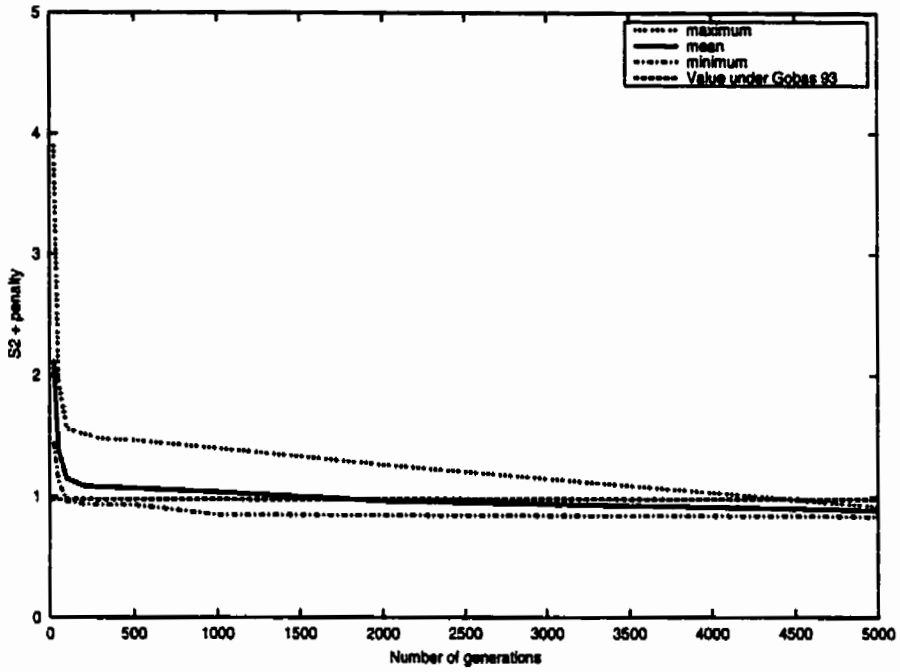

Figure 3.2: Minimum, mean, and maximum value of ($S2 + penalty$) out of 10 runs versus number of generations per run. The value of $S2$ under $\Theta_{Gobas93}$ is also shown. The data used to generate this Figure can be found in Table C.7.

the constraints. The initial decrease in the value of the objective functions was due to two factors. First, the value of $S1 + penalty$ decreased as the number or generations increased because an increasing number of solutions fulfilled the constraints. Second, the actual calibration of the food web parameters took place which further reduced the value of the objective functions.

1000 generations was deemed an appropriate criterion to stop the GA evolution during the calibrations. It is, however, important to run the GA more than once because a given evolution of 1000 generations still has the potential of yielding a solution that yields a worse fit than the default calibration. For this reason, only the 10 best solutions out of the 40 runs were used as the results of the different calibrations.

3.4 Default calibration

The predictions of $Conc._{zooplankton}$ and $Conc._{Pontoporeia}$ in $BM_{predicted}$ were calculated using a chemical partitioning model between the organism and the water for zooplankton and between the organism's lipids, the organic carbon fraction of the sediments and the interstitial water [11]. These models were later refined and the equilibrium partitioning approach was replaced by uptake and elimination kinetics models [28].

The default value of Θ was obtained from [9]:

$$\Theta_{Gobas93} = \{0.18, 0.82, 0.54, 0.25, 0.21, 0.6, 0.4, 0.1, 0.4, 0.5\}$$

The foodweb structure under $\Theta_{Gobas93}$ and the trophic position of the different fish species is illustrated in Figure 3.3. Values of the two objective functions and performance measure of $\Theta_{Gobas93}$ can be found in Table 3.2. The model fit to available data are presented in Figures D.2 and D.3.

Using $Conc._{zooplankton}$ and $Conc._{Pontoporeia}$ as forcing variables yielded a model that bypassed the uncertainties associated with the BM predictions in benthic organisms and zooplankton. This simple transformation of the model from $BM_{predicted}$ to $BM_{empirical}$ significantly improved its fit to the data. This is reflected by the 58% reduction in the value of $S1$, the 39% reduction in $S2$, and the fact that the percentage of predictions falling within a factor of 2 of the data went from 72.7% to 83.5% (Table 3.2). The overall model bias is similar for the two model forms but the standard deviation of the $\log_e(MB)$ was 0.6328 under $BM_{predicted}$ while it was 0.4939 under $BM_{empirical}$. This was further illustrated by the larger scatter of points in Figure D.2 when compared to Figure D.3. The large reduction

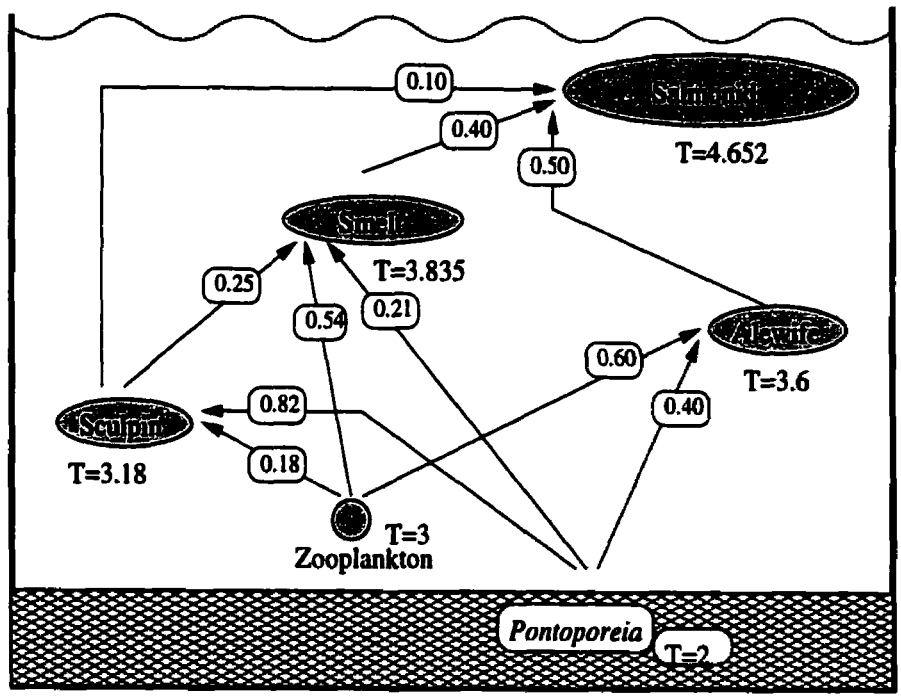

Figure 3.3: Feeding preferences and trophic position of the different Lake Ontario fish species as derived by Flint [9] and subsequently used by Gobas [11].

Table 3.2: Value of $S1$ (equation 2.4), $S2$ (equation 2.5), mean model bias (equation 2.10), and fish-specific mean model bias (equation 2.12) when using $\Theta_{Gobas93}$.

Metric	Value under $BM_{predicted}$	Value under $BM_{empirical}$
$S1$	12196.77	5063.38
$S2$	1.5975	0.9745
$\overline{\log_e(MB)}$	0.0342	0.0328
$\sigma_{\log_e(MB)}$	0.6328	0.4939
Sculpin $\overline{\log_e(MB)}$	0.2193	0.0281
Alewife $\overline{\log_e(MB)}$	-0.2008	-0.0735
Smelt $\overline{\log_e(MB)}$	0.0640	0.0819
Salmon $\overline{\log_e(MB)}$	0.0543	0.0948
predictions within a factor of 2	72.7%	83.5%
predictions within a factor of 10	100%	100%

Table 3.3: GA configuration used for the different BM calibrations

GA control parameter	Value
No. runs	40
No. generations per run	1000
No. individuals	50
p(recombination)	0.4
p(mutation)	0.01
σ	5
BM used	$BM_{empirical}$

in the value of $S1$ observed when switching from $BM_{predicted}$ to $BM_{empirical}$ is due to the better fit of high concentration data points. Since $BM_{empirical}$ yielded a much better fit than $BM_{predicted}$ and eliminated important uncertainties in the BM predictions for zooplankton and benthic organisms, $BM_{empirical}$ was used for all calibrations.

3.5 GA calibrations of food web parameters

The GA control parameters used for the four calibration routines were those in Table 3.3. The food web parameters results are presented as the best solution found by the GA after 40 runs of 1000 generations. The results of the 10 best GA runs, along with those of all 40 GA runs, were statistically evaluated to determine the efficacy of the GA to find an appropriate answer to the calibration problem. The solution found by the GA at each run was likely to be a good solution to the problem at hand but not necessarily the best solution. Since most solutions fell within a promising region of the search space, they can be amalgamated to provide an overview of the efficacy of the GA at calibrating the BM.

The GA's best solution to the calibration was used to produce the graphs of model fit to available data (Figures D.5, D.7, D.9, and D.11) and to calculate the different measures of fit (Table 3.4 and 3.6). The food web structure figures (Figures D.4, D.6, D.8, and D.10) include both the best GA answer and the mean and standard deviation of the GA's 10 best answers to the calibrations. The GA's 10 best answers were also used to determine mean and standard deviation of the trophic positions which are reported in Tables 3.5 and 3.7.

The best answer found by the GA reduced the value of the objective function from its default value (i.e. under $\Theta_{Gobas93}$) for all calibrations. The loose constraints also improved

the fit compared to the strict constraints.

When calibrating the food web structure of the bioaccumulation model through a 1000 generation-long GA run, the objective function was evaluated 5×10^5 times (1000 generations $\times$ 50 individuals = 50000 function evaluations = 5×10^4). Under the food web structure parameters calibrations, the search space facing the GA had 2.5503×10^{11} possibilities under the strict constraints scenario. Loosening the constraints increased the search space to 7.4002×10^{20} possibilities. Each GA run sampled a really small subset of this search space. When calibrating under strict constraints, the GA sampled $1.9606 \times 10^{-4}\%$ of the search space ($(5 \times 10^4 / 2.5503 \times 10^{11}) \times 100 = 1.9606 \times 10^{-4}\%$). This was the maximum proportion searched by the GA since there can exist multiple copies of the same solution in an evolving population, hence reducing the proportion of the search space sampled. Under loose constraints, the GA only sampled $6.7566 \times 10^{-15}\%$ of the search space. The fact that the GA can consistently find solutions that reduce the value of the objective function in such a large search space demonstrates its efficiency to solve the very complex calibration problem.

The GA is known to have somewhat poor convergence behavior when the solution gets close to the optimum [36]. The GA converges on the promising regions of the search space but fails to search those promising regions in greater details. This shortcoming has lead some researchers to combine the GA with traditional optimization techniques [36]. The GA used in this study included the elitist strategy, ensuring that the fitness of the best individual never regressed. The standard fitness operator aimed at making the GA more robust by applying the correct evolutionary pressure on the evolving population of solutions. Moreover, the GA was allowed to evolve for a number of generations sufficiently long to find a near-optimal solution in the promising region of the search space. Finally, the GA solution was derived from only the 10 best GA answers after 40 runs of 1000 generations.

The two objective functions used in this study yielded discernably different calibration results. Since the empirical concentration data span a wide range of values (2.8 to 860 $\frac{ng}{g}$, see Table E.1), calibrating the BM to S1 (equation 2.4) tended to concentrate on fitting the highest concentration data. This is illustrated by the highest concentration data points in Figures D.5 and D.7 being located on the 1:1 line. The GA succeded at reducing the value of $S1$ from its value under $\Theta_{Gobas93}$ but the calibrated parameters values yielded a poorer overall model fit to the available empirical data. This is because the GA solutions did not yield a good fit for the concentration data of lesser values. The BM predictions tended to underestimate the chemical concentrations in the different fish species as reflected in

Table 3.4: Value of $S1$ (equation 2.4), $S2$ (equation 2.5), mean model bias (equation 2.10), and fish-specific mean model bias (equation 2.12) when using GA calibration results obtained by minimizing $S1$. The results are given for the GA's best answer. The values under $\Theta_{Gobas93}$ are also shown for comparison.

Metric	$\Theta_{Gobas93}$	$\Theta_{S1_{min}}, strict$	$\Theta_{S1_{min}}, loose$
$S1$	5063.38	2073.38	1611.64
$S2$	0.9745	1.4325	0.953244
$\overline{\log_e(MB)}$	0.0328	-0.3084	-0.1582
$\sigma_{\log_e(MB)}$	0.4939	0.5143	0.4631
Sculpin $\overline{\log_e(MB)}$	0.0281	-0.5815	-0.0257
Alewife $\overline{\log_e(MB)}$	-0.0735	-0.1613	-0.3665
Smelt $\overline{\log_e(MB)}$	0.0819	-0.3229	-0.1457
Salmon $\overline{\log_e(MB)}$	0.0948	-0.1681	-0.0949
% within a factor of 2	83.5%	81.3 %	85.2 %
predictions within a factor of 10	100%	99.4%	100%

a negative value of $\overline{\log_e MB}$ (Table 3.4) for both GA answers. When using $\Theta_{S1_{min}}, strict$, the BM's overall bias underestimated the empirical concentration data by a factor of 0.735 ($e^{-0.3084} = 0.735$). $\Theta_{S1_{min}}, loose$ underestimated the data by a factor of 0.854 ($e^{-0.1582} = 0.854$). Considering that the BM with the default calibration only overestimates the data by a factor of 1.033, the calibrations by minimization of $S1$ do not improve the fit of the BM to the concentration data. The value of $S1$ is greatly reduced by the GA answers but at the expense of a more biased model. The GA calibration results also return trophic positions that were beyond their independently-derived values for most of the different fish species (Table 3.5). It was therefore important to log transform our data and use the results from calibrations 3 and 4 (minimization of $S2$) to explore the use of PCBs as tracers of food web processes.

The calibrations of the BM through the minimization of $S2$ were more interesting. The answers found by the GA ($\Theta_{S2_{min}}$) improved the model fit to the available data as indicated by the decrease in the value of both $S1$ and $S2$ compared to the default calibration. The food web structure found by the GA also maintained a fairly unbiased model for the strict and loose constraints cases. The BM underpredicted the concentration data by a factor of 0.912 ($e^{-0.0919} = 0.912$) with $\Theta_{S2_{min}}, strict$ and by a factor of 0.996 ($e^{-0.0042} = 0.996$) with

Table 3.5: Trophic position (Equation 2.9) of the different fish species of Lake Ontario under $\Theta_{Gobas93}$, as obtained by Vander Zanden [35], and derived from GA calibrations using $S1$. Standard deviations of the mean trophic positions are in parentheses.

Fish Species	$\Theta_{Gobas93}$	Value from [35]	$\Theta_{S1_{min,strict}}$	$\Theta_{S1_{min,loose}}$
Sculpin	3.18	3.70 (0.28)	3.82 (0.14)	3.41 (0.032)
Alewife	3.6	3.11 (0.11)	3.7 (0.27)	3.82 (0.060)
Smelt	3.835	3.66 (0.29)	4.21 (0.12)	3.68 (0.077)
Salmon	4.652	4.38 (0.38)	5.00 (0.060)	4.70 (0.019)

$\Theta_{S2_{min}}$, *loose* (Table 3.6). The variability around the mean model bias was also reduced. The food web structures that minimize $S2$ under strict and loose constraints yielded average trophic positions that were consistent with those derived by other research initiatives [9, 35] for all fish species except for smelt under the loose constraints (Table 3.7).

Loosening the search constraints allowed the expansion of the search space (2.5503×10^{11} to 7.4002×10^{20} possibilities). The new search space was much larger but the GA was still able to find appropriate solutions. An interesting result was that the GA solutions stayed well within the constraints. Loosening the constraints was similar to allowing the value of the food web structure parameters to resolve errors arising elsewhere in the model equations, model parameters, and empirical data. The solutions found did not reach the upper or lower limits of the constraints, suggesting that the total error of the model may be relatively small and that it captured the main mechanisms of chemical uptake, elimination, and biomagnification. In essence, the loose constraints allowed the GA to maximize the fit of the model even if it meant that the parameter values are ecologically impossible.

The variability around the mean values of the food web structure parameters measured the ability of the GA at finding the optimal solution to the calibration problem. The large variability in the results obtained from calibration 3 (Figure D.8) indicated that the strict constraints were an obstacle to the calibration problem. Finding a solution that fulfilled the strict constraints seemed to represent a calibration problem whose results had a higher level of variability. Different solutions respecting the strict constraints resulted in a good fit to the empirical data but failed at distinguishing themselves from other good solutions. The variability around the mean values of parameters in $\Theta_{S2_{min,strict}}$ pointed to minimum and maximum values that almost spanned the interval on which the parameters were defined. For example, the value of $P1$ in the GA solutions had a mean value of 0.49 and a standard deviation of 0.19. This means that 95% of the $P1$ values will occurred in the interval

Table 3.6: Value of $S2$ (equation 2.5), $S1$ (equation 2.4), mean model bias (equation 2.10), and fish-specific mean model bias (equation 2.12) when using GA calibration results obtained by minimizing $S2$. The values under $\Theta_{Gobas93}$ are also shown for comparison.

Metric	$\Theta_{Gobas93}$	$\Theta_{S2_{min}, strict}$	$\Theta_{S2_{min}, loose}$
$S2$	0.9745	0.8375	0.7071
$S1$	5063.38	3497.464	2283.184
$\overline{\log_e(MB)}$	0.0328	-0.0919	-0.0042
$\sigma_{\log_e(MB)}$	0.4939	0.4495	0.4216
Sculpin $\overline{\log_e(MB)}$	0.0281	-0.1647	-0.0036
Alewife $\overline{\log_e(MB)}$	-0.0735	-0.0818	-0.0049
Smelt $\overline{\log_e(MB)}$	0.0819	-0.1109	-0.0080
Salmon $\overline{\log_e(MB)}$	0.0948	0.0102	-0.0003
predictions within a factor of 2	83.5%	85.8 %	90.3 %
predictions within a factor of 10	100%	100%	100%

Table 3.7: Trophic position (Equation 2.9) of the different fish species of Lake Ontario under $\Theta_{Gobas93}$, as obtained by Vander Zanden [35], and derived from GA calibrations using $S2$. Standard deviations of the mean trophic positions are in parentheses.

Fish Species	$\Theta_{Gobas93}$	Value from [35]	$\Theta_{S2_{min,strict}}$	$\Theta_{S2_{min,loose}}$
Sculpin	3.18	3.70 (0.28)	3.49 (0.14)	3.42 (0.0017)
Alewife	3.6	3.11 (0.11)	3.55 (0.14)	3.54 (0.0019)
Smelt	3.835	3.66 (0.29)	3.76 (0.15)	3.41 (0.0026)
Salmon	4.652	4.38 (0.38)	4.61 (0.10)	4.59 (0.0041)

[0.12,0.86] i.e. $[\mu \pm (1.96 * \sigma)] = [0.49 \pm (1.96 * 0.19)] = [0.12, 0.86]$. Recall that the parameters were searched on a $[0, 1]$ interval. Loosening the constraints significantly reduced the variability around the mean value of the calibrated food web structure paramaters found by the GA. The GA's convergence to a solution was much better, as indicated by the lower variability around the calibrated feeding preferences (Figure D.10).

The loose constraints allowed the GA to find a solution that adjusted the level of contaminant input from the different prey items. The GA solutions needed to be interpreted differently than those from the strict constraints scenario. The fact that a given fish species can have a preference for a prey item that is greater than one might not make sense at first. Once corrected for the sum of diet preferences, the trophic position calculated through equation 2.9 actually used the relative proportion of each prey item to the organism's diet. The GA answer was still usable to determine the trophic position of the different fish species and also encoded the appropriate level of contaminant uptake from the different prey items.

3.6 Resolving food web structure

The calibration results consisted of the food web structures that gave the best fit to the available contaminant concentration data set. Since there were independently derived estimates of the diet preferences of the Lake Ontario fish species, the results could be used to determine if GA-derived food webs compared to previously published results. Using the methodology developped here, a subset of organic contaminants could be used with the BM to backcalculate the food web structure that best explained the observations in chemical concentrations. If the BM equations captured the main mechanisms of bioaccumulation, and if the value of the other BM parameters were well known, calibration results of Θ could provide an alternative to the existing ecological methods for determining food web structure.

The trophic position of the four Lake Ontario fish species as determined by Flint [9] and Vander Zanden [35] were different from each other. The food web hierarchy for Lake Ontario as suggested by Flint [9] and used by Gobas [11] ($\Theta_{Gobas93}$) puts salmonids at the top, followed by smelt and alewife at intermediate levels, and sculpin at the bottom. Vander Zanden also puts salmonids at the top of the food web but sculpin is at an intermediate level while alewife is at the bottom. The trophic positions for Flint's calibration fell within the 95% confidence region of those from Vander Zanden for sculpin, smelt, and salmon (i.e. $T_{sculpin_{Flint}} = 3.18 \in [3.70 \pm (1.96 * 0.28)]$, $T_{smelt_{Flint}} = 3.835 \in [3.66 \pm (1.96 * 0.29)]$ and

$T_{salmon_{Flint}} = 4.652 \in [4.38 \pm (1.96 * 0.38)]$). The trophic position of alewife, on the other hand, was dissimilar for the two benchmarks (i.e. $T_{alewife_{Flint}} = 3.6 \notin [3.11 \pm (1.96 * 0.11)]$). Changes in the structure of the Lake Ontario food web have been documented [13] and could explain the discrepancies between the two food web structures used to evaluate the GA solutions.

$\Theta_{Gobas93}$ represents a food web where sculpin is a low-level predator at $T_{sculpin} = 3.18$. However, Vander Zanden found the same species to occupy a higher position at $T_{sculpin} = 3.70$ [35]. The food web structures determined by $\Theta_{S2_{min,strict}}$ and $\Theta_{S2_{min,loose}}$ yielded trophic positions for sculpin ($T_{sculpin} = 3.49$ and $T_{sculpin} = 3.42$, respectively) that fell within the values from the benchmarks. In the simplified food web, sculpins had two prey items at trophic position $T_{zooplankton} = 3$ and $T_{Pontoporeia} = 2$ which allowed them a potential trophic position ranging from 3.00 to 4.00. The low variability around the mean trophic position of sculpin as determined from the GA's 10 best answers under loose constraints was due to the clear convergence of the GA towards solutions approaching $S2_{min}$. $\Theta_{S2_{min,strict}}$ systematically underpredicted sculpin concentrations by a factor of 0.848, as indicated by $\overline{\log_e(MB)} = -0.1647$. $\Theta_{S2_{min,loose}}$ gave an almost unbiased model for sculpin ($\overline{\log_e(MB)} = -0.0036$). The GA solutions were in agreement with independently-derived values for the diet composition of Lake Ontario sculpin.

Alewife in $\Theta_{Gobas93}$ is an intermediate-level predator at $T_{alewife} = 3.6$ while [35] found the same species to occupy a lower position at $T_{alewife} = 3.11$. The food web structures determined by $\Theta_{S2_{min,strict}}$ and $\Theta_{S2_{min,loose}}$ yielded trophic positions for alewife ($T_{alewife} = 3.55$ and $T_{alewife} = 3.54$, respectively) that were between the two benchmark values, and that are practically similar to those suggested by Flint [9] and used in $\Theta_{Gobas93}$. In the simplified food web, alewife also had a potential trophic position ranging from 3.00 to 4.00. Here again, the very low variability around the mean trophic position of alewife as determined from the GA's 10 best answers under loose constraints indicated that the GA has a strong tendency to converge to this solution. $\Theta_{S2_{min,strict}}$ systematically underpredicted alewife concentrations, as indicated by $\overline{\log_e(MB)} = -0.0818$. $\Theta_{S2_{min,loose}}$ gave an almost unbiased model for alewife ($\overline{\log_e(MB)} = -0.0049$). The GA solutions were in agreement with independently-derived values for the diet composition of Lake Ontario alewife.

$\Theta_{Gobas93}$ represents a food web where smelt is a high-level predator at $T_{smelt} = 3.835$ while [35] found the same species to occupy a lower position at $T_{smelt} = 3.66$. The food web structure determined by $\Theta_{S2_{min,strict}}$ yielded a trophic position for smelt that was between the benchmark values ($T_{smelt} = 3.76$). The food web determined by $\Theta_{S2_{min,loose}}$ yielded a trophic

position for smelt that was lower than both benchmark values ($T_{smelt} = 3.41$). In the simplified food web, alewife had a potential trophic position ranging from 3.00 to 5.00. Smelt was different from sculpin and alewife because it had the potential to be piscivorous. The simplified food web allowed smelt to eat sculpin, in a proportion determined by $P4$ (Table 2.1 and Figure 2.1). In $\Theta_{S2_{min,loose}}$, $P4 = 0.00$ which reduced the trophic position of smelt. The GA solution suggested that the diet of smelt was entirely composed of zooplankton and *Pontoporeia*. However, Vander Zanden reported $P4 = 0.30$ and identified smelt as an important source of contaminants to the salmonids because of its elevated trophic position [35]. Flint reported $P4 = 0.25$ [9]. $\Theta_{S2_{min,loose}}$ suggests that contaminant input into smelt from sculpin was not necessary to match the observed concentration data. A potential explanantion for this discrepancy is that the BM consistently overpredicts the contaminant concentrations in smelt, as indicated by a $\overline{\log_e(MB)}$ of 0.0819 in Table 3.2. In order for the GA to improve the fit of contaminant concentrations in smelt, it had to lower the intake of contaminants from the most contaminated prey item, namely the sculpin. The low variability around the mean trophic position of smelt as determined from the GA's 10 best answers under loose constraints was due to the convergence of the GA towards solutions approaching $S2_{min}$.

$\Theta_{Gobas93}$ represents a food web where salmonids are the top predator at $T_{salmon} = 4.652$ while [35] found the same species to occupy a lower position at $T_{salmon} = 4.38$. The food web structures determined by $\Theta_{S2_{min,strict}}$ and $\Theta_{S2_{min,loose}}$ yielded trophic positions for salmon ($T_{salmon} = 4.61$ and $T_{salmon} = 4.59$, respectively) that were between the benchmark values. In the simplified food web, salmon had a potential trophic position ranging from 4.00 to 6.00. The GA solutions were in agreement with independently-derived values for the diet composition of Lake Ontario salmonids.

Considering the many sources of errors already present in [9, 35], the GA results seem to represent a sound alternative to the existing parameter values. The new calibrated model using $\Theta_{S1_{min}}, loose$ yielded an overall unbiased model for all fish species and resolved the trophic position of all fish species except smelt.

Chapter 4

Recommendations for further research

This study demonstrated the ability of the GA to calibrate a set of input parameters of a bioaccumulation model. The calibration scheme investigated appeared to be able to resolve the food web structure of the Lake Ontario ecosystem from contaminant concentration data.

The GA can be used to calibrate BM parameters whose values are uncertain. The methodology presented here is a mathematically sound way to find parameter values and represents a clear advantage over finding parameter values on a trial-and-error basis. A major problem with the development of site-specific BM lies in the difficulty of obtaining certain parameter values. The GA can be used with a limited concentration data set to estimate these parameters. Another application of the GA calibration is to resolve food web structure by using contaminants as tracers of trophic interactions.

Another interesting project would be to compare calibration results of the model using two different data sets taken before and after recorded changes in the trophic structure of an ecosystem (e.g. perform calibrations using data from [24] and compare to the results of this project). This would provide an opportunity to test whether model calibration to observed concentration data captures changes in food web structure.

The calibrations performed in this project used a wide variety of PCB congeners, assigning an equal weight to each in the optimization. Certain PCB congeners are known to be very difficult to metabolize by fish [24], and they could be used as a more robust subset of contaminants to extract food web structure from BM calibration. Alternatively, a subset of PCB congeners representing the majority of the contaminant body burden in fish could be used. Oliver and Niimi identified 10 PCB congeners that constitute over half the PCBs in fish [30].

The methodology developed here can also be applied to incomplete concentration data

sets. An incomplete data set is one that does not have data points for each compound and for each fish species. Since the objective function implemented a multi-objective optimization, the weight of the missing species-compound data points could be set to 0 and calibration could be performed as usual. This method is interesting because incomplete data sets are very common. The data set available to a manager often consists of disparate data for different compounds in a variety of species.

An important limitation to the method presented is that it fails to use the variability in the measured data when performing the calibration. The evolution of the GA solutions did not take this variability into consideration when deciding if a solution was good or not. Instead, the average values of the concentration data was used as a single point estimate for evaluating the fit of the BM. This shortcoming is important and considerations should be taken when deciding which compounds to use for calibration. Variability in the measured data maybe used in conjunction with the results of a Monte Carlo analysis.

The effect of the variability in the value of the forcing variables and model parameters ($C_w, C_s, K_{ow}, Conc._{zooplankton}, Conc._{Pontoporeia}, \vec{\rho}$, and Θ) on the model outcome can be determined by a Monte Carlo analysis. This is performed by running the model for a large number of times and sampling the value of the different forcing variables and model parameters from their known distribution instead of using their mean value as a point estimate. The results from such an analysis show the level of variability in the model predictions resulting from explicitly considering the uncertainty into the values of the forcing variables and model parameters. The uncertainty estimates of the model's predictions can be compared to that of the data set for another indicator of the model's predictive ability. It was not possible to perform such an analysis using the GA because $S2$ was always calculated with the mean value of the reported concentration data. It would be interesting to see if merging the Monte Carlo analysis and GA calibration would result in tangible results. Coupling of both methods would be feasible by using Monte Carlo-like values for running the BM. The GA would evolve a solution using a BM whose parameter values change at each generation based on their distribution. The elitist strategy should not be used under this framework because the BM will be different each time it is evaluated and will have many different levels of fitness. It is expected that the value of $S2$ would eventually stabilize and the subsequent variability around $S2_{min}$ would be attributable to the variability in the forcing variables.

The GA was shown to be a mathematically-sound method for calibrating BM parameters. This research also evaluated the possibility of obtaining food web structure information about an ecosytem by using contaminant concentration data as tracers of trophic

linkages. Such knowledge of an ecosystem is important since it will be a major determinant in the extent of contaminant biomagnification [6, 2].

Appendix A

Bioaccumulation model

A.1 Mass-balance equations

The total mass of chemical present in an organism is:

$$X_{fish} = V_{fish} * C_{fish} \tag{A.1}$$

where,

X_{fish} is the mass of chemical present in the fish (ng)

V_{fish} is the volume of the fish (g)

C_{fish} is the chemical concentration in the fish ($\frac{ng}{g}$)

The net flux of chemical through a fish can be described by the following mass-balance equation:

$$\begin{aligned} d(V_{fish} \cdot C_{fish})/dt = \\ & (k_1 * V_{fish} * C_{water}) + (k_D * V_{fish} * C_{diet}) \\ & - (k_2 * V_{fish} * C_{fish}) - (k_E * V_{fish} * C_{fish}) \\ & - (k_G * V_{fish} * C_{fish}) - (k_M * V_{fish} * C_{fish}) \end{aligned} \tag{A.2}$$

where,

C_{water} is the bioavailable fraction of the total water concentration ($\frac{\mu g}{L}$, Equation A.3)

k_1 is the uptake rate constant from water via the gills (day^{-1})

k_2 is the elimination rate constant to water via the gills (day^{-1})

k_D is the uptake rate constant from food (day^{-1})

k_E is the elimination rate constant by faecal egestion (day^{-1})

k_G is the growth dilution rate constant (day^{-1})

k_M is metabolic transformation rate constant (day^{-1})

The bioavailable fraction of the total water concentration is determined by:

$$C_{water} = BSF * C_{water_{total}} \tag{A.3}$$

$$BSF = 1/[1 + ((K_{ow} * OC_{water})/\rho_{OC})] \tag{A.4}$$

By using the steady-state assumption ($d(X_{fish})/dt = 0$), mass-balance equation A.2 can be simplified to:

$$(k_1 * C_{water}) + (k_D * C_{food}) = C_{fish} * (k_2 + k_E + k_G + k_M)$$

Solving for C_{fish} yields:

$$C_{fish} = [(k_1 * C_{water}) + (k_D * C_{food})] / (k_2 + k_E + k_G + k_M) \tag{A.5}$$

Food web transfer is included in the BM by allowing each fish species to have more than one food item, and by specifying the proportion of each food item in the organism's diet:

$$C_{food} = \sum_{i=1}^{m} (C_{food\,item\,i} * P_i)$$

where,

m is the number of prey species

The "k"s in the mass-balance equation are the first-order rate constants that describe the uptake and elimination routes of contaminants to the fish tissue. They are calculated through existing submodels of chemical transport and from allometric relationships.

$$k_1 = \left[\left(\left(V_{fish}/Q_w\right) + \left(V_{fish}/Q_l\right)/K_{ow}\right)\right]^{-1}$$

$$k_2 = \left[\left(\left(V_l/Q_w\right) * K_{ow}\right) + \left(V_l/Q_l\right)\right]^{-1}$$

$$V_l = V_{fish} * l_{fish}$$

$$Q_w = 88.3 * V_f^{0.6(\pm 0.2)}$$

$$Q_l \approx Q_w/100$$

$$k_D = E_D * \left(F_D/V_f\right)$$

$$k_E = 0.25 * k_D$$

$$F_D = 0.022 * V_f^{0.85} * \exp(0.06 * T)$$

$$E_D = \left[\left(5.3 \times 10^{-8} * K_{ow}\right) + 2.3\right]^{-1}$$

$$k_G = 0.000502 * V_f^{-0.2}$$

$$k_M = 0$$

A.2 Value of non-calibrated parameters

Table A.1 presents the different non-calibrated parameters of the BM and their value in [11].

Table A.1: Value of the different BM parameters (from [11])

Parameter symbol	Parameter name	Units	Value
T	Mean water temperature	°C	8
OC_{water}	Organic content of the water	$\frac{kg}{L}$	2.5×10^{-7}
ρ_{lipids}	Density of lipids	$\frac{kg}{L}$	0.9
ρ_{OC}	Density of organic carbon	$\frac{kg}{L}$	0.9
$V_{sculpin}$	Sculpin weight	g	5.4
$l_{sculpin}$	Sculpin lipid content	%	8.0
$V_{alewife}$	Alewife weight	g	32
$l_{alewife}$	Alewife lipid content	%	7.0
V_{smelt}	Smelt weight	g	16
l_{smelt}	Smelt lipid content	%	4.0
$V_{salmonids}$	Salmonids weight	g	2410
$l_{salmonids}$	Salmonids lipid content	%	16

Appendix B

Genetic Algorithm

GAs are machine-learning techniques first introduced by Holland [19]. The algorithm uses the principles of genetic variation and natural selection to solve optimization problems. The algorithm begins by randomly initializing a "population" of "individuals". Each individual has a single binary or real-valued chromosome which is used to encode a possible solution in a given problem space. This space, referred to as the search space, comprises all possible solutions to the problem at hand. The algorithm assigns a numerical "fitness" value to each individual based on how well they perform in the "environment". The environment used to determine each individual's fitness is case-dependent. In model calibration, for example, the numerical fitness of each individual can be calculated through a goodness-of-fit function. The population is then reproduced and selection is applied to all individuals. Many selection procedures are currently in use, one of the simplest being Holland's original fitness-proportionate selection [19], where individuals are selected with a probability proportional to their relative fitness. This ensures that the expected number of times an individual is chosen is approximately proportional to its relative performance in the population. Thus, high-fitness ("good") individuals have a better chance of reproducing, whereas low-fitness ones are more likely to disappear. The population of potential solutions is allowed to evolve until a certain number of generations has passed or until a satisfactory solution has been found.

While GAs are very efficient at directing the search towards promising regions of the search space, they are also found to be relatively inefficient at finding the optimal solution once they reach such regions. This is because the configuration that is efficient at early parts of the search becomes inefficient at later stages [36]. This kind of problem can be circumvented by introducing non-traditional techniques such as dynamic and adaptive strategies

[14, 36].

GA research from the computing sciences aim at making the algorithm as efficient as possible. This is of less concern here because obtaining the optimal solution in the smallest number of generations is not necessary. The information obtained through the evolution can be used to estimate the uncertainty of the parameters being calibrated and can provide an overview of model sensitivity to those parameters. The GA has to provide an acceptable answer to the problem at hand, but not necessarily in the shortest amount of time possible. This bias towards efficiency is more suited for other real-time GA applications.

B.1 The mutation and recombination operators

Selection alone cannot introduce any new individuals into the population; it cannot find new points in the search space. These are generated by genetically-inspired operators used in the reproduction step of the GA. The most common reproduction operators are crossover and mutation. Crossover is performed with a certain probability (the "crossover probability") between two selected individuals, called parents, by exchanging parts of their chromosomes to form two new individuals, called offspring. In its simplest form, the recombination operator exchanges substrings determined by a randomly selected crossover point. This operator tends to enable the evolutionary process to move toward "promising" regions of the search space. The mutation operator is introduced to prevent premature convergence to local optima by randomly sampling new points in the search space. It is carried out by flipping bits at random, with some (small) probability. The role of the different operators is to strike a balance between exploiting promising areas of the search space while allowing the GA to continue exploring the search space. Genetic algorithms are stochastic iterative processes not guaranteed to converge; the termination condition may be specified as some fixed, maximal number of generations or as the attainment of an acceptable fitness level [18].

B.2 The elitist strategy and the standard fitness selection operator

Many ways exist for selecting individuals based on their fitness. A popular selection strategy that has empirically shown its superiority over others is the elitist strategy [26]. The

elitist strategy stipulates that the best individual of a given generation makes it to the next generation intact. This strategy circumvents the potential problem of losing the best-so-far individual through mutation and/or recombination.

Another potential pitfall is the premature convergence of the GA to a sub-optimal solution. In normal fitness-proportionate reproduction, each individual's probability of making it to the next generation is proportional to its raw fitness. It is possible that the fittest individual in a generation encodes a sub-optimal solution. If the other individuals encode solutions that are even worse, the GA will converge prematurely because of the relative advantage of the sub-optimal fittest individual. When using standard fitness (a sub-class of ranking methods), each individual receives a pre-assigned fitness value that is dependent on their *rank* within the population [4]. This way, the best individual in a given generation will have the largest pre-assigned fitness but will not get a disproportionately large fitness. There are many different ways to calculate the value of the pre-assigned fitness. An interesting aspect of this methodology is that the "evolutionary pressure", i.e. the level of exploitation of promising regions in the search space, can be controlled through an outside parameter.

Uncertainty estimation of calibrated parameters

A main difference between the GA and other optimization methods is that the algorithm searches the parameter space through a population of individuals. Most other optimization methods perform their task by handling the objective function at only one given point at a time. The GA's exploration of the parameter space can be used to obtain information such as parameter correlation and approximate $(1-\alpha)\%$ confidence region. Traditional methods return the best fit parameter values, leaving the estimation of correlation and uncertainty to a separate task using other tools such as Monte Carlo simulation.

The GA performs this task by keeping track of solutions for which the objective function is below a certain threshold value. This method only approximates the $(1-\alpha)\%$ confidence intervals of the calibrated parameters. Moreover, the information provided is uninformative about the type of distribution to which the parameters belong. For this reason, the uncertainty information is to be interpreted carefully.

A word of caution

Genetic algorithms were found to be very efficient at directing the search towards promising parts of the search space. However, they were found to be relatively inefficient at finding the

optimal solution once they reached such a promising part of the search space; the configuration efficient at early parts of the search become inefficient at later stages [36]. This kind of problem can be circumvented by introducing non-traditional techniques such as dynamic and adaptive strategies.

One of the primary objectives of GA research from computer science literature is to make the algorithm as efficient as possible. This is of little concern in our case because we are not necessarily interested in obtaining the optimal solution in the smallest number of generations. We are actually interested in the GA lingering around for many generations because we use the information obtained to estimate the uncertainty of the parameters being calibrated. It is important for the GA to provide an acceptable answer to the problem at hand but not necessarily in the shortest amount of time possible. This bias towards efficiency is more suited for other real-time GA applications.

Appendix C

Meta-level optimization and number of generations results

Each data point in Tables C.1, C.2, C.3, C.4, and C.5 is the mean value of $S1$ after 10 GA runs of 200 generations. The 10 GA runs were repeated for each combination of GA configuration parameters and used a synthetic data set of contaminants generated using $BM_{predicted}$.

The minimum, mean, and maximum values of the objective function after 10 runs is given in Tables C.6 and C.7. For each objective function, the 10 runs were repeated for 9 different number of generations (25, 50, 100, 200, 300, 500, 1000, 2000, and 5000). The GA runs were performed under strict constraints and used the empirical data set of contaminants found in Appendix E.

Table C.1: Mean value of $S1$ under different probabilities of recombination ($p(rec)$) and mutation ($p(mut)$) for $\sigma = 3$. The data in this Table was used to generate Figure C.1.

		$p(mut)$					
		0.001	0.005	0.01	0.05	0.1	0.25
$p(rec)$	0.0	31.73	7.79	4.91	14.44	32.79	89.39
	0.2	73.71	6.39	8.80	17.41	31.42	92.86
	0.4	54.89	9.39	3.81	17.38	46.82	125.54
	0.6	57.48	4.44	5.25	14.48	28.43	90.54
	0.8	111.41	5.28	6.13	14.32	35.65	133.15
	1.0	49.21	4.55	9.02	23.01	50.04	138.79

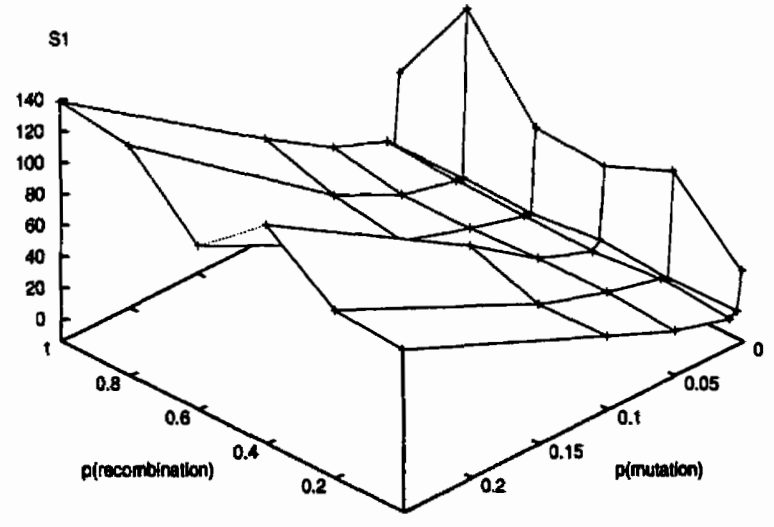

Figure C.1: Mean value of $S1$ under different probabilities of recombination and mutation for $\sigma = 3$.

Table C.2: Mean value of $S1$ under different probabilities of recombination ($p(rec)$) and mutation ($p(mut)$) for $\sigma = 5$. The data in this Table was used to generate Figure C.2.

		$p(mut)$					
		0.001	0.005	0.01	0.05	0.1	0.25
$p(rec)$	0.0	14.66	8.71	5.43	12.60	25.73	141.50
	0.2	22.27	5.46	6.72	11.84	33.44	99.64
	0.4	43.73	8.70	6.81	15.33	28.35	130.42
	0.6	59.19	4.45	4.68	15.95	38.42	115.78
	0.8	93.31	8.84	7.64	23.95	49.24	148.85
	1.0	80.67	8.73	7.10	26.74	53.02	121.76

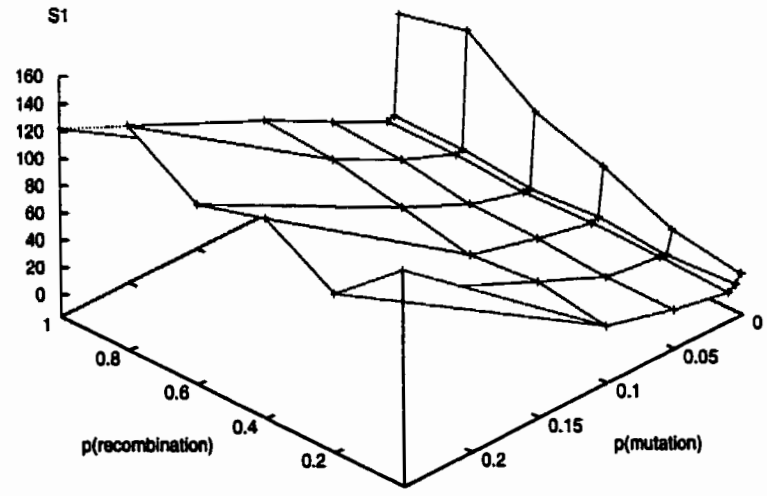

Figure C.2: Mean value of $S1$ under different probabilities of recombination and mutation for $\sigma = 5$.

Table C.3: Mean value of $S1$ under different probabilities of recombination ($p(rec)$) and mutation ($p(mut)$) for $\sigma = 10$. The data in this Table was used to generate Figure C.3.

		$p(mut)$					
		0.001	0.005	0.01	0.05	0.1	0.25
$p(rec)$	0.0	36.35	9.18	6.12	11.58	27.80	132.10
	0.2	100.68	9.91	8.18	16.34	23.38	139.07
	0.4	23.12	9.49	6.82	12.09	27.53	109.24
	0.6	29.70	10.20	5.12	17.52	46.49	138.43
	0.8	17.13	6.37	8.37	22.28	64.55	133.03
	1.0	44.46	8.77	8.21	35.95	68.06	123.68

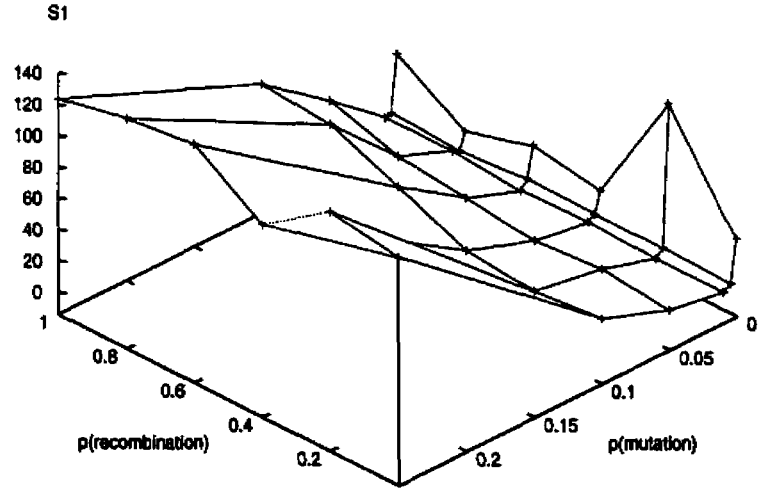

Figure C.3: Mean value of $S1$ under different probabilities of recombination and mutation for $\sigma = 10$.

Table C.4: Mean value of $S1$ under different probabilities of recombination ($p(rec)$) and mutation ($p(mut)$) for $\sigma = 25$. The data in this Table was used to generate Figure C.4.

		$p(mut)$					
		0.001	0.005	0.01	0.05	0.1	0.25
$p(rec)$	0.0	53.65	7.87	7.49	17.61	40.48	138.66
	0.2	40.03	9.88	8.63	16.98	33.15	107.89
	0.4	25.45	5.52	5.86	19.80	42.98	99.32
	0.6	40.56	9.07	8.23	32.04	55.93	107.01
	0.8	24.82	12.57	9.16	26.99	63.22	123.28
	1.0	52.96	23.22	28.30	48.38	89.53	174.09

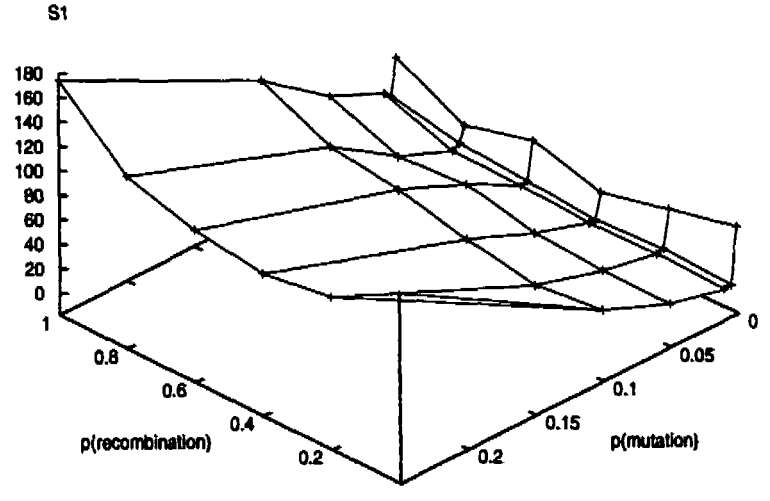

Figure C.4: Mean value of $S1$ under different probabilities of recombination and mutation for $\sigma = 25$.

Table C.5: Mean value of $S1$ under different probabilities of recombination ($p(rec)$) and mutation ($p(mut)$)for $\sigma = 50$. The data in this Table was used to generate Figure C.5.

		$p(mut)$					
		0.001	0.005	0.01	0.05	0.1	0.25
$p(rec)$	0.0	116.76	15.47	20.48	23.97	43.98	130.65
	0.2	70.85	14.76	10.68	18.39	46.12	110.60
	0.4	52.99	8.39	9.93	28.65	52.15	113.22
	0.6	139.00	11.73	13.38	38.94	64.41	129.38
	0.8	20.38	17.61	27.81	35.18	73.64	163.68
	1.0	23.21	26.45	22.38	59.81	68.99	145.27

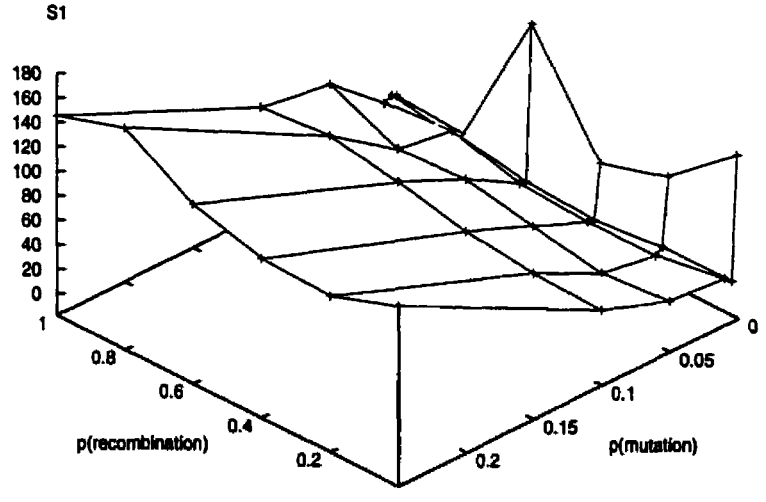

Figure C.5: Mean value of $S1$ under different probabilities of recombination and mutation for $\sigma = 50$.

Table C.6: Minimum $((S1 + penalty)_{min})$, mean $(\overline{(S1 + penalty)})$, and maximum $((S1 + penalty)_{max})$ value of $(S1 + penalty)$ after 10 runs. The GA runs were performed under strict constraints and used the empirical data set of contaminants found in Table E.1. This data was used to generate Figure 3.1.

Number of generations	$(S1 + penalty)_{min}$	$\overline{(S1 + penalty)}$	$(S1 + penalty)_{max}$
25	4129.04	6080.41	8217.48
50	3535.43	4655.80	6234.25
100	2895.06	4088.91	5684.56
200	2563.54	3573.25	4649.41
300	2370.86	3300.52	4105.18
500	2262.68	3091.40	3796.88
1000	2073.39	2929.11	3781.25
2000	2070.20	2689.55	3185.06
5000	2055.91	2506.21	2943.05

$S1 = 5063.38$ with $\Theta_{Gobas93}$

Table C.7: Minimum $((S2 + penalty)_{min})$, mean $(\overline{(S2 + penalty)})$, and maximum $((S2 + penalty)_{max})$ value of $(S2 + penalty)$ after 10 runs. The GA runs were performed under strict constraints and used the empirical data set of contaminants found in Table E.1. This data was used to generate Figure 3.2.

Number of generations	$(S2 + penalty)_{min}$	$\overline{(S2 + penalty)}$	$(S2 + penalty)_{max}$
25	1.4403	2.1177	3.9043
50	1.1303	1.3942	1.9813
100	0.9613	1.1457	1.5563
200	0.9361	1.0845	1.5128
300	0.9347	1.0781	1.4758
500	0.9300	1.0677	1.4633
1000	0.8491	1.0351	1.3950
2000	0.8440	0.9608	1.2602
5000	0.8358	0.8828	0.9125

$S2 = 0.9745$ with $\Theta_{Gobas93}$

Appendix D

Calibration results of food web structure parameters

The graphs of BM fit to the available data under the different calibration solutions (Figures D.5, D.7, D.9, and D.11) all have 44 data points per fish species for a total of 176 data points per plot. The x and y axes both show the log_e of the predictions and observations in $\frac{ng}{g}$ wet weight. The solid line represents the ideal fit. The thick dotted lines represent predictions that are within a factor of 2 from the data and the thin dashed lines represent predictions that are within an order of magnitude.

The figures of food web structure (Figures D.4, D.6, D.8, and D.10) show the diet composition of each fish species under the different calibration results along with their trophic position. For each trophic link (P_i), the GA's best answer is given along with its mean value and standard deviation for the 10 best runs.

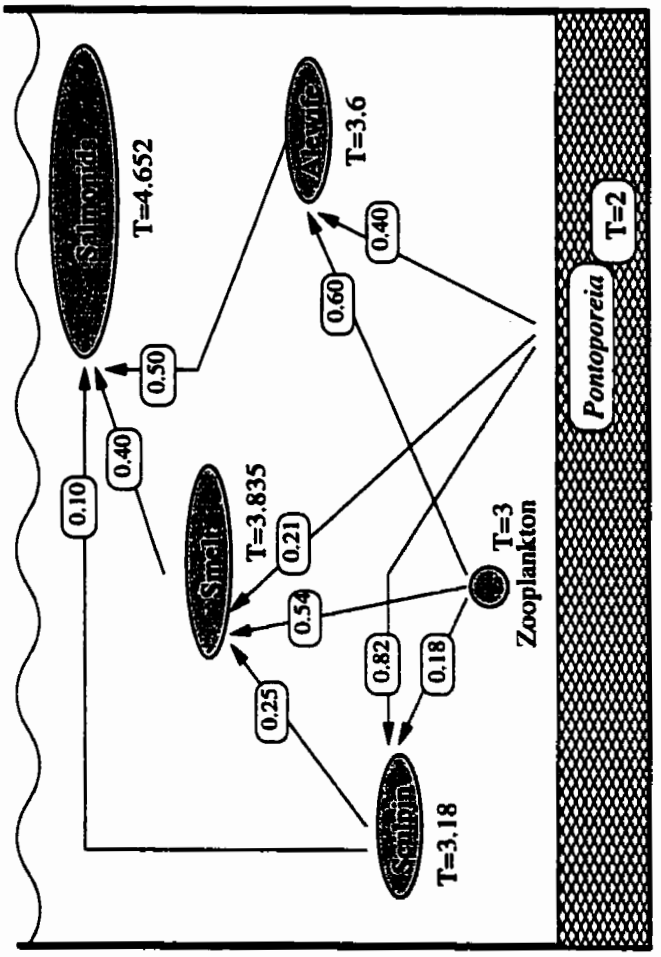

Figure D.1: Conceptual diagram of the Lake Ontario food web as estimated by [9] and used by [11].

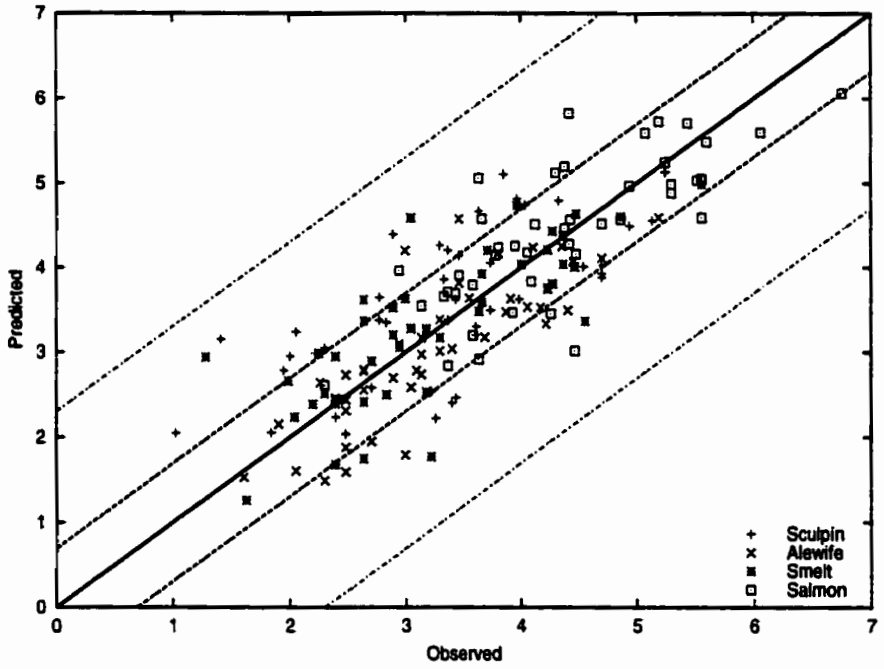

Figure D.2: Relationship between predicted and observed concentrations for the four Lake Ontario fish species when using $BM_{predicted}$ with $\Theta_{Gobas93}$. The solid line represents the ideal fit. The thick dotted lines represent predictions that are within a factor of 2 from the data and the thin dashed lines represent predictions that are within an order of magnitude.

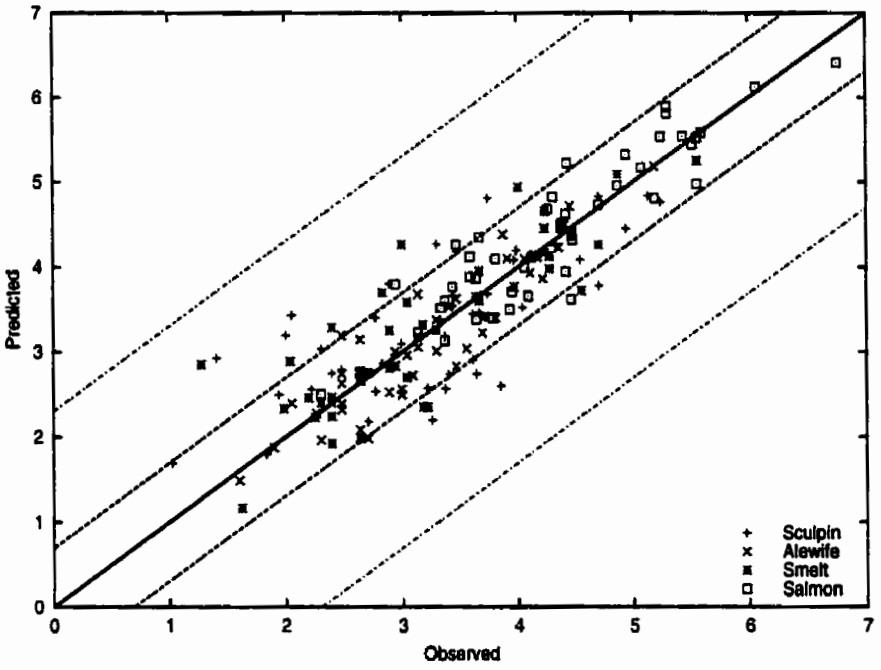

Figure D.3: Relationship between predicted and observed concentrations for the four Lake Ontario fish species when using $BM_{empirical}$ with $\Theta_{Gobas93}$. The solid line represents the ideal fit. The thick dotted lines represent predictions that are within a factor of 2 from the data and the thin dashed lines represent predictions that are within an order of magnitude.

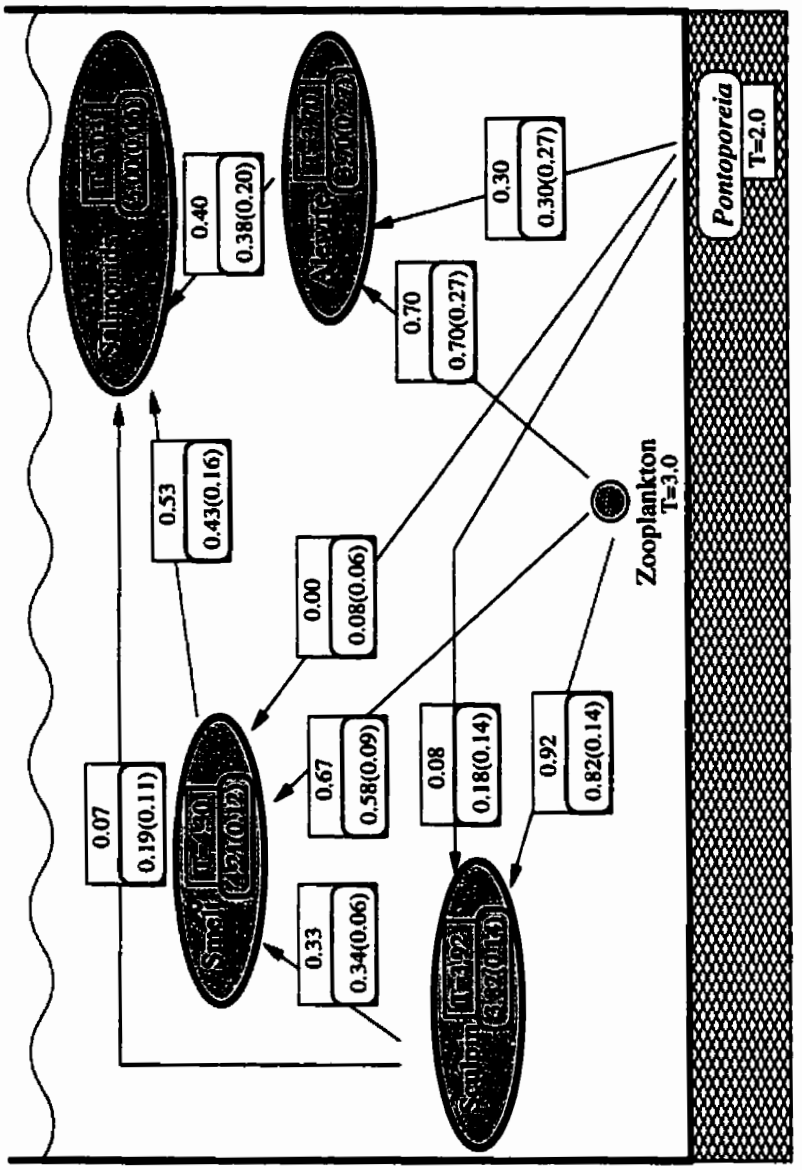

Figure D.4: Conceptual diagram of the Lake Ontario food web determined by GA calibrations of $BM_{empirical}$ through minimization of $S1$ with strict constraints. The diet composition of each fish species and their realized trophic position are presented for the GA's best answer (top of box, single value) and for the 10 best runs (bottom of box, mean and standard deviation).

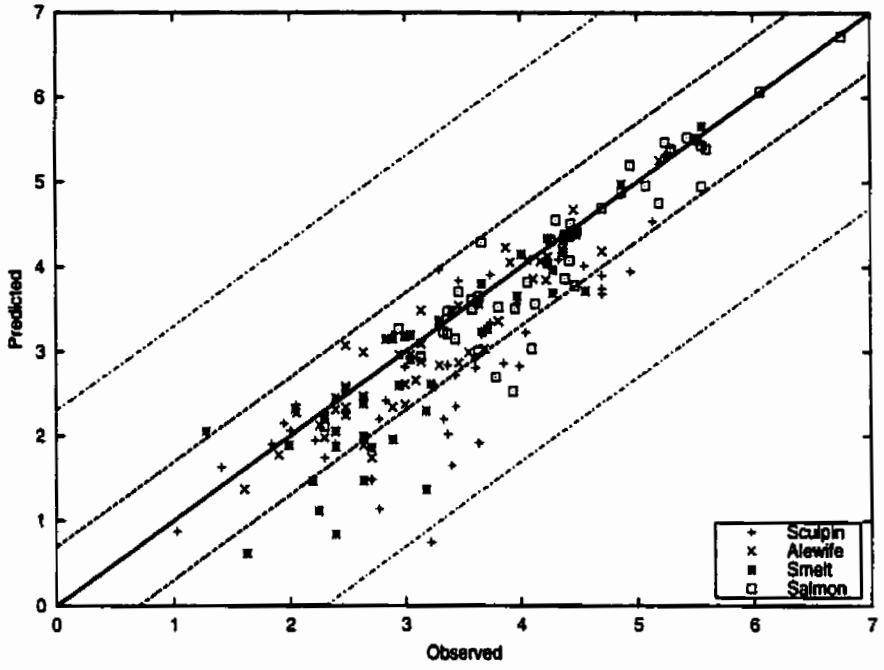

Figure D.5: Relationship between predicted and observed concentrations for the four Lake Ontario fish species when using $BM_{empirical}$ with $\Theta_{SI_{min,strict}}$. The solid line represents the ideal fit. The thick dotted lines represent predictions that are within a factor of 2 from the data and the thin dashed lines represent predictions that are within an order of magnitude.

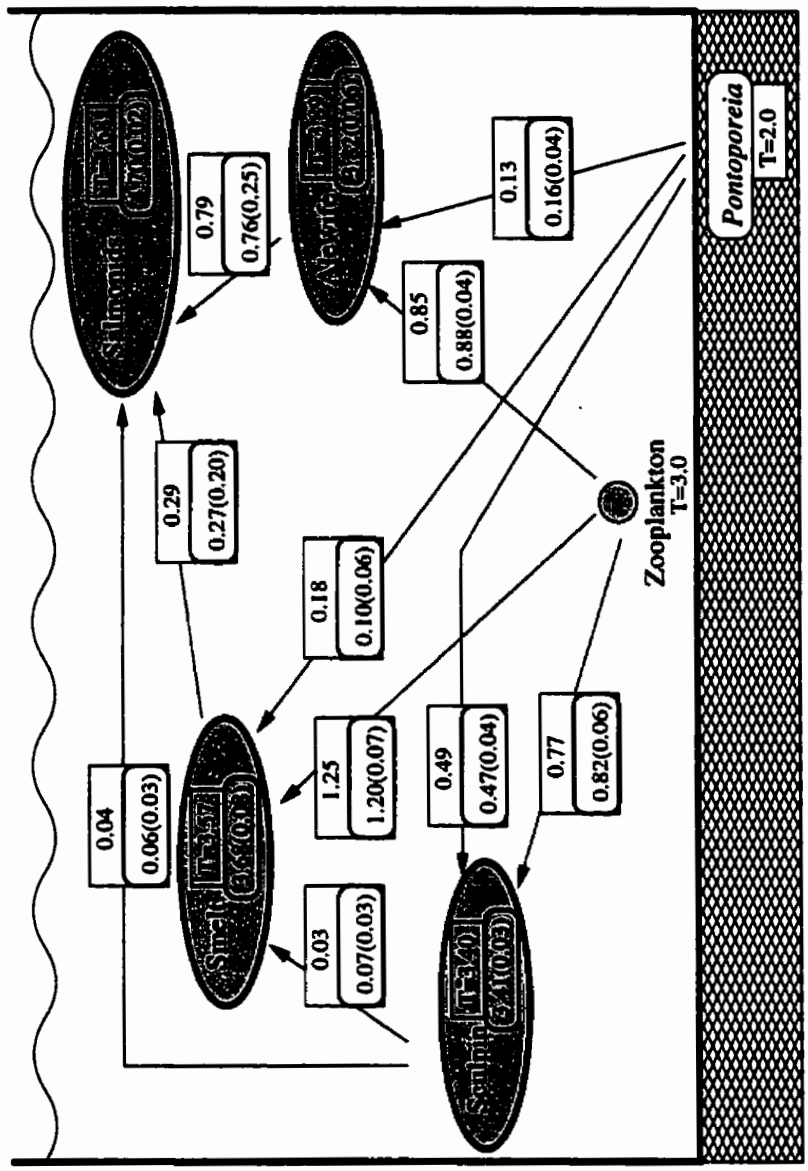

Figure D.6: Conceptual diagram of the Lake Ontario food web determined by GA calibrations of $BM_{empirical}$ through minimization of $S1$ with loose constraints. The diet composition of each fish species and their realized trophic position are presented for the GA's best answer (top of box, single value) and for the 10 best runs (bottom of box, mean and standard deviation).

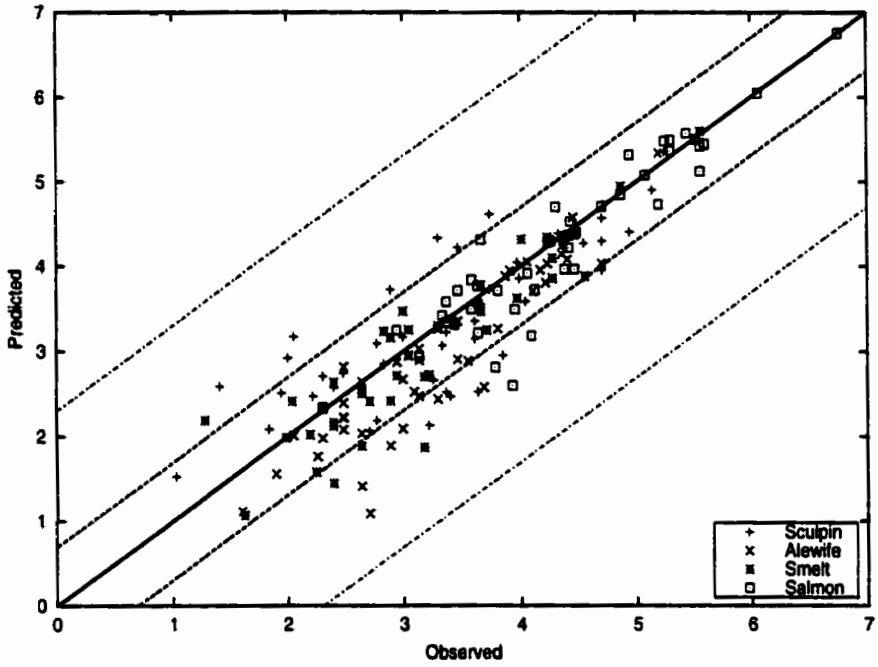

Figure D.7: Relationship between predicted and observed concentrations for the four Lake Ontario fish species when using $BM_{empirical}$ with $\Theta_{SI_{min,loose}}$. The solid line represents the ideal fit. The thick dotted lines represent predictions that are within a factor of 2 from the data and the thin dashed lines represent predictions that are within an order of magnitude.

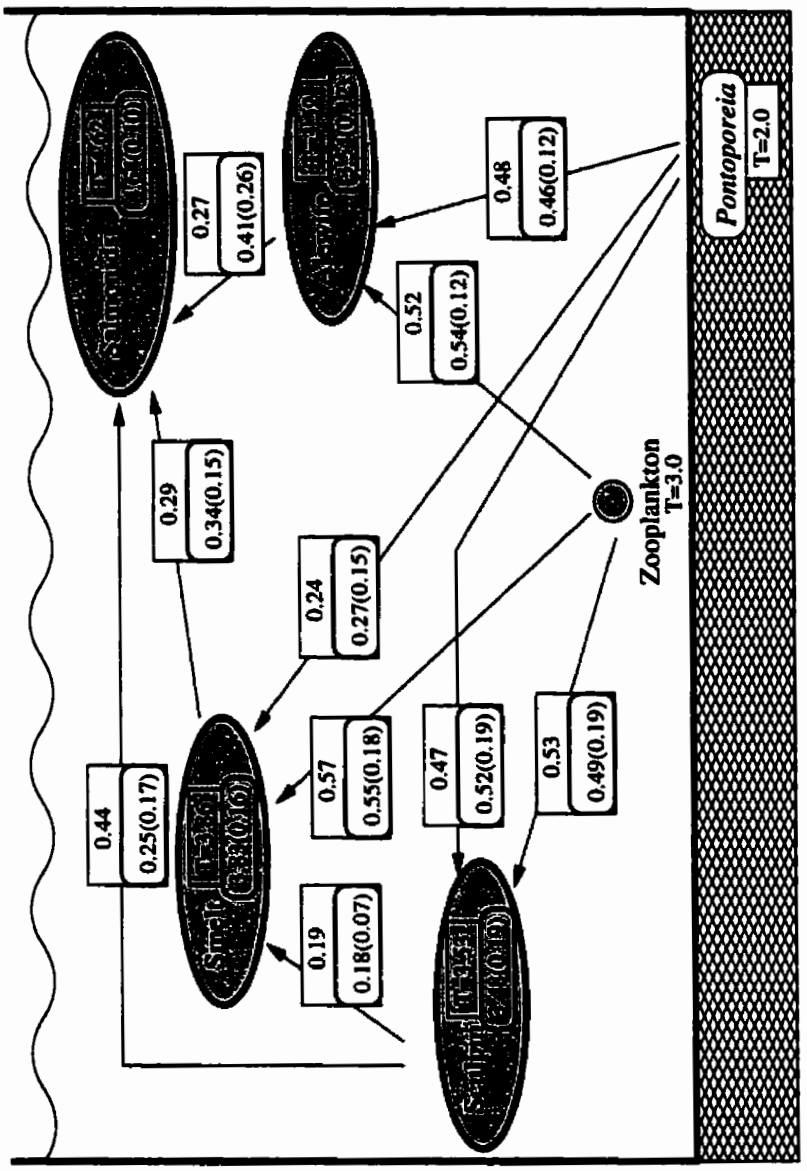

Figure D.8: Conceptual diagram of the Lake Ontario food web determined by GA calibrations of $BM_{empirical}$ through minimization of $S2$ with strict constraints. The diet composition of each fish species and their realized trophic position are presented for the GA's best answer (top of box, single value) and for the 10 best runs (bottom of box, mean and standard deviation).

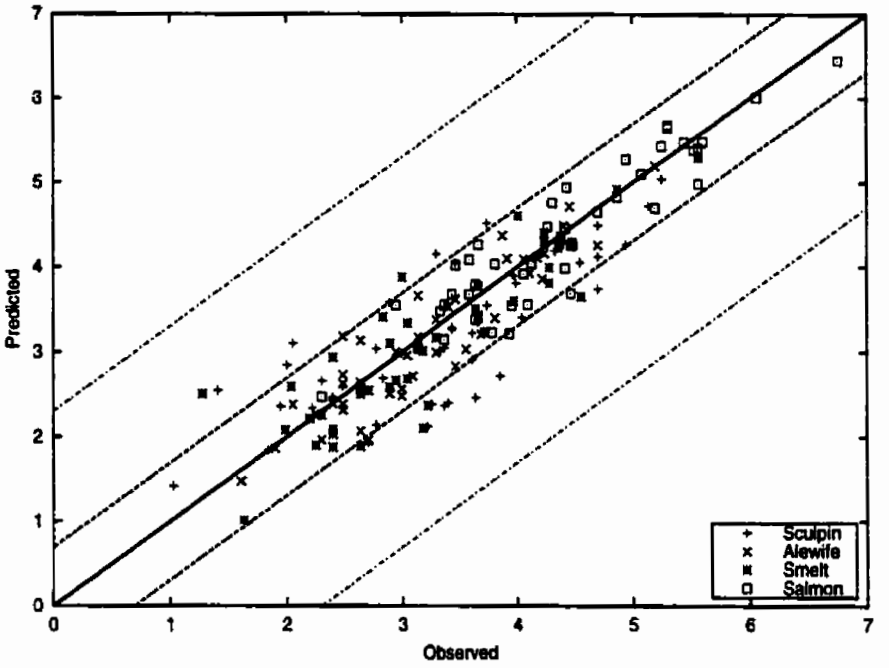

Figure D.9: Relationship between predicted and observed concentrations for the four Lake Ontario fish species when using $BM_{empirical}$ with $\Theta_{S2_{min,strict}}$. The solid line represents the ideal fit. The thick dotted lines represent predictions that are within a factor of 2 from the data and the thin dashed lines represent predictions that are within an order of magnitude.

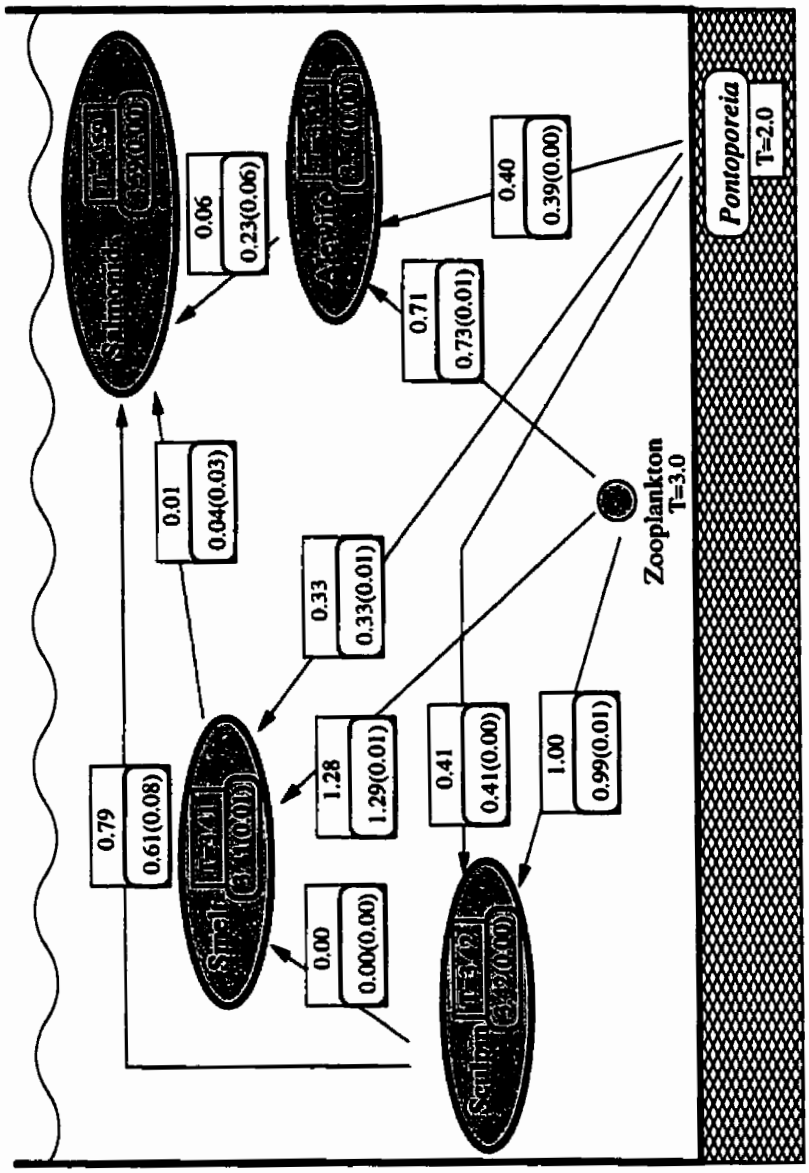

Figure D.10: Conceptual diagram of the Lake Ontario food web determined by GA calibrations of $BM_{empirical}$ through minimization of $S2$ with loose constraints. The diet composition of each fish species and their realized trophic position are presented for the GA's best answer (top of box, single value) and for the 10 best runs (bottom of box, mean and standard deviation).

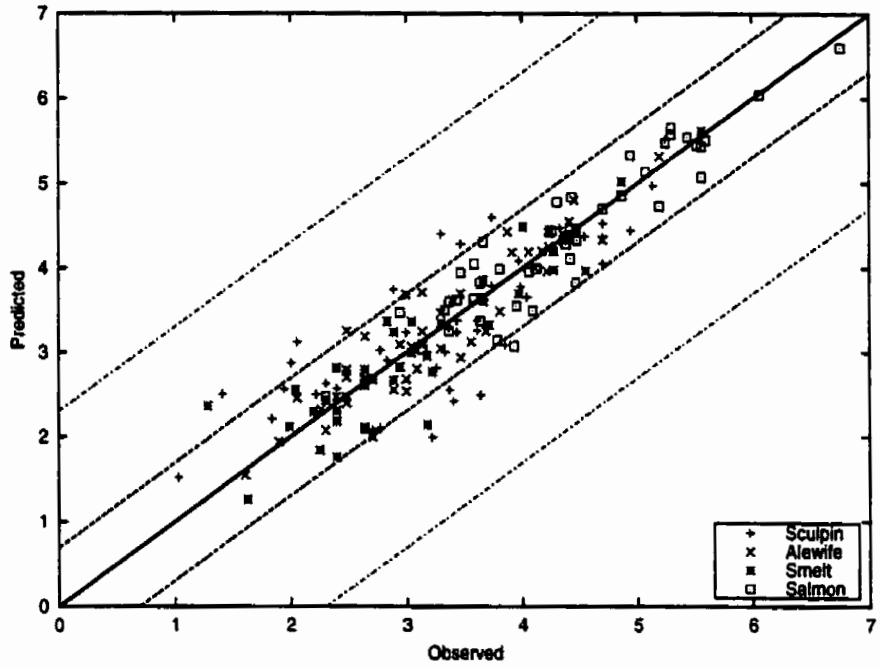

Figure D.11: Relationship between predicted and observed concentrations for the four Lake Ontario fish species when using $BM_{empirical}$ with $\Theta_{S2_{min,loose}}$. The solid line represents the ideal fit. The thick dotted lines represent predictions that are within a factor of 2 from the data and the thin dashed lines represent predictions that are within an order of magnitude.

Appendix E

Contaminants data set

The data presented here is from [30]. The fields of Table E.1 are:

C_w concentration in water ($\frac{pg}{L}$)

C_s concentration in sediments ($\frac{ng}{g}$dry weight)

C_{biota} concentration in biota ($\frac{ng}{g}$wet weight)

Table E.1: Data set of 44 hydrophobic organic contaminants used for the BM calibration.

Compound	$\log K_{ow}$	C_w	C_s	C_{zoo}	C_{ponto}	$C_{sculpin}$	$C_{alewife}$	C_{smelt}	C_{salmon}
PCB 28	5.67	46	17	6.4	30	7.8	14	14	36
PCB 66	6.2	31	46	15	30	53	61	72	160
PCB 70	6.2	45	23	20	32	32	50	72	140
PCB 56	6.11	9.7	33	11	25	18	32	39	74
PCB 52	5.84	63	25	3.5	22	28	27	18	62
PCB 47	5.85	41	12	1.7	14	4.1	18	24	60
PCB 44	5.75	50	23	4.2	26	16	23	15	45
PCB 74	6.2	10	2.7	4	8.2	12	12	14	38
PCB 49	5.85	24	11	2.3	16	10	14	9	31
PCB 64	5.95	9.3	9.4	3.6	8.4	9.2	11	11	28
PCB 42	5.76	3.3	4.7	1.5	4.6	2.8	5	5.1	10
PCB 101	6.38	130	27	18	37	140	110	79	270
PCB 84	6.04	15	21	21	24	110	68	95	260
PCB 118	6.74	34	15	19	21	94	58	87	250

Compound	$\log K_{ow}$	C_w	C_s	C_{zoo}	C_{ponto}	$C_{sculpin}$	$C_{alewife}$	C_{smelt}	C_{salmon}
PCB 110	6.48	55	37	22	28	76	78	88	230
PCB 87	6.29	21	20	18	61	42	82	69	200
PCB 105	6.65	14	10	8.5	12	39	27	38	110
PCB 95	6.13	52	14	2.8	22	31	40	24	80
PCB 85	6.3	9.4	9.8	4.4	8.1	17	22	19	58
PCB 82	6.2	2.6	2.9	3.1	2.7	6.3	10	11	29
PCB 91	6.13	40	5.7	3.4	5.8	7	12	10	29
PCB 153	6.92	50	25	30	45	170	86	130	430
PCB 138	6.83	28	15	16	26	110	65	79	260
PCB 149	6.67	34	20	18	27	27	69	69	190
PCB 146	6.89	3.8	6.7	6.3	6.2	37	21	27	88
PCB 141	6.82	8.3	7.4	5.1	12	37	23	21	83
PCB 151	6.64	2.7	3.7	0.3	5.7	25	15	11	51
PCB 132	6.58	17	11	5.8	8.5	20	19	18	39
PCB 180	7.36	27	13	11	48	110	48	55	200
PCB 187	7.17	18	8.4	8.8	14	42	30	39	130
PCB 170	7.27	2.7	10	3.8	26	54	23	20	84
PCB 183	7.2	2.5	3.1	4.3	13	31	12	17	71
PCB 177	7.08	1.1	2.5	1.9	5.9	11	7.8	7.7	36
PCB 174	7.11	1.9	5.1	2	9.4	7.4	12	11	32
PCB 203	7.65	2.6	8.2	2.4	5	29	12	14	52
PCB 194	7.8	7.8	3.7	1.4	3.6	15	6.7	7.3	23
ppDDE	6.51	76	51	81	36	190	180	260	860
ppDDD	6.02	93	72	8.3	5.3	47	32	21	83
ppDDT	6.91	19	18	5.4	8.6	29	35	41	80
mirex	6.89	31	31	8	12	57	45	53	180
photomirex	6.0	17	3.9	7.4	4.2	26	20	25	87
γ-chlordane	7.34	34	2.1	1.2	5.8	30	9.6	3.6	19
OCS	6.2	4.7	11	0.9	6.9	16	14	9.5	44
HCB	5.47	150	100	4	18	38	20	14	38

Bibliography

[1] Atwell, L., K. A. Hobson, and H. E. Welch (1998). "Biomagnification and bioaccumulation of mercury in an arctic marine food web: insights from stable isotope analysis." *Canadian Journal of Fisheries and Aquatic Sciences*, **55**:1114–1121.

[2] Bentzen, E., D. R. Lean, W. D. Taylor, and D. Mackay (1996). "Role of food web structure on lipid and bioaccumulation of organic contaminants by lake trout (*Salvelinus namaycush*)." *Canadian Journal of Fisheries and Aquatic Sciences*, **53**:2397–2407.

[3] Berthouex, P. and L. Brown (1994). *Statistics for environmental engineers*. CRC Press.

[4] Bradshaw, J. and J. Miles (1997). "Using standard fitness with genetic algorithms." *Advances in Engineering Software*, **28**:425–435.

[5] Burkhard, L. P. (1998). "Comparison of two models for predicting bioaccumulation of hydrophobic organic chemicals in a Great Lakes food web." *Environmental Toxicology and Chemistry*, **17**(3):425–435.

[6] Cabana, G., A. Tremblay, J. Kalff, and J. B. Rasmussen (1994). "Pelagic food chain structure in Ontario lakes: A determinant of mercury levels in Lake Trout (*Salvelinus namaycush*)." *Canadian Journal of Fisheries and Aquatic Sciences*, **51**:381–389.

[7] Drever, R. (1999). *Made in the Shade: Regenrating coastal Douglas-fir and western redcedar under partial cutting treatments*. Master's thesis, School of Resource and Environmental Management, Simon Fraser University.

[8] ECS. "Kow estimation software." Free online demo version available from URL http://ecs.syrres.com/interkow/kowdemo.htm.

[9] Flint, R. W. (1986). "Hypothesized carbon flow through the deepwater lake Ontario food web." *Journal of Great Lakes research*, **12**(4):344–354.

[10] Franchini, M. and G. Galeati (1997). "Comparing several genetic algorithm schemes for the calibration of conceptual rainfall-runoff models." *Hydrological Sciences*, **42**(3):357–379.

[11] Gobas, F. A. P. C. (1993). "A model for predicting the bioaccumulation of hydrophobic organic chemicals in aquatic food-webs: Application to lake Ontario." *Ecological Modelling*, **69**:1–17.

[12] Gobas, F. A. P. C. and J. A. McCorquodal, editors (1992). *Modeling the accumulation and toxicity of organic chemicals in aquatic food chains*, chapter 6. Lewis Publishers.

[13] Gobas, F. A. P. C., M. N. Z'Graggen, and X. Zhang (1995). "Time response of the Lake Ontario ecosystem to virtual elimination of PCBs." *Environmental Science and Technology*, **29**:2038–2046.

[14] Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley Publishing Company, Inc., Reading, Mass.

[15] Grefenstette, J. J. (1986). "Optimization of control parameters for genetic algorithms." *IEEE Transactions on Systems, Man, and Cybernetics*, **16**(1):122–128.

[16] Hawker, D. W. and D. W. Connell (1988). "Octanol-water partition coefficients of polychlorinated biphenyl congeners." *Environmental Science and Technology*, **22**:382–387.

[17] Heath, F. (1972). "Origins of the binary code." *Scientific American*, **227**(2):76–83.

[18] Heitkötter, J. and D. Beasley (1999). "The hitch-hiker's guide to evolutionary computation: A list of Frequently Asked Questions (FAQ)." Available via anonymous FTP from rtfm.mit.edu:/pub/usenet/news.answers/ai-faq/genetic/.

[19] Holland, J. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI.

[20] Johnson, V. (1995). "Optimal field-scale groundwater remediation using neural networks and the genetic algorithm." *Environmental Science and Technology*, **29**:1145–1155.

[21] Kiriluk, R. M., M. R. Servos, M. D. Whittle, G. Cabana, and J. B. Rasmussen (1995). "Using ratios of stable nitrogen and carbon isotopes to characterize the biomagnification of DDE, mirex, and PCB in a Lake Ontario pelagic food web." *Canadian Journal of Fisheries and Aquatic Sciences*, **52**:2660–2674.

[22] Levine, S. (1980). "Several measures of trophic structure applicable to complex food webs." *Journal of Theoretical Biology*, **83**:195–207.

[23] Marquardt, D. W. (1963). "An algorithm for least-squares estimation of nonlinear parameters." *J. Soc. Indust. Appl. Math.*, **11**(2):431–441.

[24] Metcalfe, T. and C. Metcalfe (1997). "The trophodynamics of PCBs, including mono- and non-*ortho* congeners, in the food web of North-Central Lake Ontario." *Science of the Total Environment*, **201**:245–272.

[25] Meylan, W. and P. Howard (1995). "Atom/fragment contribution method for estimating octanol-water partition coefficients." *J. Pharm. Sci.*, pages 83–92.

[26] Michalewicz, Z. (1994). *Genetic algorithms + data structures = evolution programs*. Springer-Verlag, New York.

[27] Microsoft. "Quickbasic 4.5 for DOS."

[28] Morrison, H. A., F. A. Gobas, R. Lazar, and G. D. Haffner (1996). "Development and verification of a bioaccumulation model for organic contminants in benthic invertebrates." *Environmental Science and Technology*, **30**:3377–3384.

[29] Mulligan, A. E. and L. C. Brown (1998). "Genetic algorithms for calibrating water quality models." *Journal of environmental engineering*, **124**(3):202–211.

[30] Oliver, B. G. and A. J. Niimi (1988). "Trophodynamic analysis of polychlorinated biphenyl congeners and other chlorinated hydrocarbons in the lake Ontario ecosystem." *Environmental science and technology*, **22**:388–397.

[31] Park, L., C. Park, C. Parl, and T. Lee (1997). "Application of genetic algorithms to parameter estimation of bioprocesses." *Medical & Biological Enginnering & Computing*, **35**:47–49.

[32] Ratkowsky, D. A. (1990). *Handbook of nonlinear regression models*. M. Dekker, New York.

[33] Spears, W. M. and K. A. De Jong (1991). "An analysis of multi-point crossover." In G. Rawlins, editor, "First workshop on the foundations of genetic algorithms and classifier systems (FOGA/CS-90)," pages 301–315. Kaufmann, M.

[34] SPSS. "SPSS-statistical software."

[35] Vander Zanden, M. J. and J. B. Rasmussen (1996). "A trophic position model of pelagic food webs: impact on contaminant bioaccumulation in lake trout." *Ecological Monographs*, **66**(4):451–477.

[36] Yang, R. and I. Douglas (1998). "Simple genetic algorithm with local tuning: Efficient global optimizing technique." *Journal of Optimization Theory and Applications*, **98**(2):449–465.

[37] Yao, L. and W. A. Sethares (1994). "Nonlinear parameter estimation via the genetic algorithm." *IEEE Transactions on signal processing*, **42**(4):927–935.