

Statistical Distributions of Compartmentalized Populations Governed by Continuous Time, Discrete State Semi-Markov Processes

Richard L. Patterson

*School of Natural Resources
The University of Michigan
Ann Arbor, Michigan 48109*

and

Zhenkui Ma

*School of Natural Resources
The University of Michigan
Ann Arbor, Michigan 48109*

ABSTRACT

Stochastic compartmental modeling theory is extended to represent nonhomogeneous Poisson immigration of an arbitrary discrete population into an open system subdivided into compartments in which residence times of individuals in each compartment are statistically independent, are identically distributed, and follow arbitrary piecewise continuous distributions. The model is fitted to time series of chloride concentrations using an L_1 metric implemented by a linear goal program, covering lakes Huron, Erie, and Ontario. Advantages of the stochastic model are: (1) fitting a multivariate model to a multivariate data set using formal methods of statistical inference, and (2) allowance for multiple sources of random variability covering input, residence times, and distribution of individuals among compartments. Feasibility of numerical implementation of the model is demonstrated.

INTRODUCTION

Equilibrium behavior of stochastic compartmental models was recently reviewed by Whittle [1]. An important characteristic of certain classes of these network models is that distributions of counts of individuals at network nodes (compartments) are insensitive to all moments of residence time distributions at these nodes except the first, i.e., the mean residence time.

Moreover, the joint equilibrium distribution of counts of individuals at nodes has the product form denoting statistical independence of counts at nodes and, in fact, is a product of Poisson distributions when the network is open and arrivals from outside the network are Poisson distributed.

Our objective here is: (1) to describe the transient joint distribution of counts of individuals at nodes of both open and closed networks in which statistically independent movements of individuals among nodes are governed by continuous time, discrete state semi-Markov processes, and (2) to illustrate a case example of the theory in the field of water quality modeling.

Let $(X(t); t > 0)$ denote a continuous time, finite state semi-Markov process in which $X(t)$ is the state of the process at its most recent change of state prior to time t . $X(t)$ is described probabilistically by an underlying Markov chain with stochastic transition matrix $P = (p_{ij})$, state residence time distribution function matrix $W = (w_{ij}(t))$, and stochastic interval transition probability matrix $F = (f_{ij}(t))$ ($i, j = 1, \dots, n$). The element $f_{ij}(t)$ is the conditional probability that $X(t) = j$ at time t , given that the initial state $X(0+)$ is i . Assuming time invariance of the process, the conditional probability that $X(t) = j$, given that $X(z) = i$ ($0 < z < t$), is $f_{ij}(t - z)$. Elements of F are related to elements of P and W by a Markov renewal equation of the Volterra type whose unique, time dependent solution is, by conditioning on the number of changes of state of the process prior to time t ,

$$\begin{aligned} f_{ij}(t) &= \sum_{l=0}^{\infty} \Pr(X(t) = j | X(0+) = i, l \text{ changes of state in } (0, t)) \\ &\quad \times \Pr(l \text{ changes of state in } (0, t) | X(0+) = i) \\ &= d_{ij}h_i(t) + p_{ij} \int_0^t w'_{ij}(z)h_j(t-z) dz \\ &\quad + \sum_{l=2}^{\infty} \left\{ \left[\sum_{q_1=1}^n p_{iq_1} \sum_{q_2=1}^n p_{q_1q_2} \cdots \sum_{q_{k-2}=1}^n p_{q_{k-3}q_{k-2}} \sum_{q_{k-1}=1}^n p_{q_{k-2}q_{k-1}} p_{q_{k-1}q_l} \right] \right. \\ &\quad \left. \times \left[\int_0^t (w_{iq_1} * w_{q_1q_2} * \cdots * w_{q_{k-2}q_{k-1}} * w_{q_{k-1}q_l})'(z)h_j(t-z) dz \right] \right\}, \end{aligned}$$

where $q_1 = j$; d_{ij} is the Kronecker delta;

$$\int_0^t (w_{iq_1} * \cdots * w_{q_{k-1}q_l})'(z)h_j(t-z) dz,$$

an l -fold convolution density, convolved with $h_j(t)$, is multiplied by the probability $(p_{iq_1} \cdots p_{q_{k-1}q_l})$ that the l -step sequence of changes of state $(i, q_1; \dots; q_{k-1}, q_l = j)$ occurs:

$$h_i(t) = 1 - \sum_{k=1}^n p_{ik} w_{ik}(t);$$

and state j is assumed to be accessible from state i , so that there exists at least one l -step sequence with positive probability, for some index $l = 1, 2, 3, \dots$.

Let C denote a discrete population for which states (i.e., conditions) of individuals in C are in one-to-one correspondence with states of the process $(X(t))$. Let $S = (P, W, F)$ denote a system governing mutually independent movements of individuals among states $1, 2, \dots, n$ once they enter S . The states are also identified with nodes $1, 2, \dots, n$ of the system S . The conditional probability that an individual is in state j at time t (i.e., in node j of the system), given that it initially entered S at time z ($0 < z < t$) in node i , is therefore $f_{ij}(t - z)$. Let $Y_{ij}(t)$ ($i, j = 1, \dots, n$) denote random variables describing counts of individuals in node j at time t which initially entered S at node i at time $t = 0+$. We shall investigate distributions of the counts of individuals $Y_{ij}(t)$ for both open and closed systems S .

CLOSED SYSTEMS

Assume that N individuals enter S initially in node i at instant $t = 0+$ and none enter thereafter. Since by the stochasticity of the matrices P and F all individuals are accounted for once they enter S , the system is effectively closed for all $t > 0$. Assuming that the behavior of all individuals in S follows the conditions described above, the joint distribution of counts in nodes $1, 2, \dots, n$ at time $t > 0$ is multinomial with parameters $N, f_{i1}(t), \dots, f_{in}(t)$. The marginal distribution of the count in node j is binomial with parameters N and $f_{ij}(t)$. All N individuals are assumed to initiate movement among nodes simultaneously at instant $t = 0+$ from node i . Otherwise, some type of departure process of individuals out of node i is indicated, and the distribution of counts in nodes $1, \dots, n$ is no longer multinomial. If N_1, N_2, \dots, N_n individuals are initially at nodes $1, 2, \dots, n$, the count of individuals at node j at time t has a mixed distribution and is binomial if and only if $f_{1j}(t) = \cdots = f_{nj}(t)$. A Poisson approximation applies in the general case.

OPEN SYSTEMS

Poisson Arrivals to S

Let $(A_i(t); t > 0; a_i(t) \geq 0)$ denote Poisson process of arrivals of individuals at initial node i . The function $a_i(t)$ is a piecewise continuous arrival intensity, where $E(A_i(t)) = \int_0^t a_i(z) dz$. By applying a simple argument the following powerful result is obtained:

$$\begin{aligned} \Pr(Y_{i1}(t) = y_{i1}, \dots, Y_{in}(t) = y_{in}) \\ = \frac{[m_{i1}(t)]^{y_{i1}} \dots [m_{in}(t)]^{y_{in}}}{y_{i1}! \dots y_{in}!} e^{-m_{i1}(t)} \dots e^{-m_{in}(t)} \end{aligned} \quad (1)$$

where

$$m_{ij}(t) = \int_0^t a_i(z) f_{ij}(t-z) dz.$$

Equation (1) for the joint distribution of counts at nodes $1, 2, \dots, n$ is characteristic of open networks, of which the system S is an example. Equation (1) further demonstrates:

- (i) statistical independence of numbers of individuals at nodes $1, 2, \dots, n$;
- (ii) a Poisson marginal distribution of the count of individuals at node j ($j = 1, \dots, n$) at time t ($t > 0$) with mean

$$E(Y_{ij}(t)) = m_{ij}(t),$$

(iii) a Poisson distribution of counts of individuals across arbitrary subsets of nodes;

(iv) that the total number of individuals in S at time t is

$$E(Y_{i1}(t) + \dots + Y_{in}(t)) = E(A_i(t)),$$

and an equilibrium distribution of the count at node j will exist at $t = \infty$ provided $m_{ij}(\infty) < \infty$ and $a_i(t)$ is appropriately restricted;

(v) that if independent Poisson processes of arrivals at nodes $1, 2, \dots, n$ feed the system S , then the distribution of counts of individuals at nodes $1, 2, \dots, n$ remains Poisson distributed.

We next consider some generalizations of Poisson arrival processes of individuals at an open system S .

Randomized Poisson Arrival Intensity a_i

Assume that individuals arriving at initial node i are distributed as a Poisson process with arrival intensity $a_i(t) = a_i$, where a_i is a random variable with a Gamma distribution:

$$\Pr(a < a_i < a + da) = \frac{h}{\Gamma(r)} (ha)^{r-1} e^{-ha} da \quad (r, a, h > 0).$$

It follows that the marginal distribution of $Y_{ij}(t)$ is

$$\Pr(Y_{ij}(t) = m) = \binom{r+m-1}{m} p^r (1-p)^m \quad (m = 0, 1, 2, \dots), \quad (2)$$

where

$$p = \frac{h}{h + \int_0^t f_{ij}(z) dz}.$$

The mean and variance of $Y_{ij}(t)$ are

$$E(Y_{ij}(t)) = \frac{r(1-p)}{p} \quad \text{and} \quad V(Y_{ij}(t)) = \frac{1}{p} E(Y_{ij}(t))$$

when r is a positive integer. $Y_{ij}(t)$ has the negative binomial distribution with parameters r and p . In contrast with the Poisson distribution, the variance of the negative binomial exceeds the mean, due to the additional source of variation contributed by randomization of the arrival intensity of individuals to an initial node. Subscripts i and j have been suppressed on the parameters r and h of the Gamma distribution of the arrival intensity a_i .

Poisson Arrivals of Batches at an Initial Node

Assume a Poisson process of arrivals of random sized batches of individuals to initial node i :

$$(B_i(t); t > 0; b_i(t) \geq 0),$$

where

$$m_i = E(B_i(t)) = \int_0^t b_i(z) dz$$

is the expected number of batches arriving at initial node i in $(0, t)$. Assume that arriving batches have a common distribution of count with mean m_i and variance v_i . The counts of individuals in nodes are now Poisson distributed if and only if (i) $m_i = 1$ and (ii) $v_i = 0$. Under certain conditions, however, counts may have an approximate Poisson distribution, as demonstrated below. The mean and variance of the marginal distribution of $Y_{ij}(t)$ are

$$E(Y_{ij}(t)) = m_i \int_0^t b_i(z) f_{ij}(t-z) dz \quad (i, j = 1, \dots, n) \quad (3.1)$$

and

$$\begin{aligned} V(Y_{ij}(t)) = & m_i \int_0^t b_i(z) f_{ij}(t-z) [1 - f_{ij}(t-z)] dz \\ & + (m_i^2 + v_i) \int_0^t b_i(z) [f_{ij}(t-z)]^2 dz \quad (i, j = 1, \dots, n). \end{aligned} \quad (3.2)$$

Under appropriate conditions on m_i , v_i , and $f_{ij}(t)$ the random count $Y_{ij}(t)$ of individuals at node j at time t will be approximated by a random variable $Y_{ij}(t)$ having a Poisson distribution with mean

$$E(Y_{ij}(t)) = m_i \int_0^t b_i(z) f_{ij}(t-z) dz. \quad (4.1)$$

To investigate, consider the ratios

$$\frac{E(Y_{ij}(t))}{E(Y_{ij}(t))} \quad \text{and} \quad \frac{V(Y_{ij}(t))}{E(Y_{ij}(t))}.$$

The former is always equal to unity, but

$$\frac{V(Y_{ij}(t))}{E(Y_{ij}(t))} = 1 + \left(m_i + \frac{v_i}{m_i} - 1 \right) \frac{\int_0^t f_{ij}^2(z) dz}{\int_0^t f_{ij}(z) dz}. \tag{4.2}$$

Equation (4.2) shows that the variance of the approximating random variable $Y_{ij}(t)$ is less than the variance of $Y_{ij}(t)$ if $m_i > 1$. If $m_i < 1$ (implying fractions of individuals), the variance of $Y_{ij}(t)$ may be greater or smaller than that of the approximation, depending upon magnitudes of m_i , v_i , and $f_{ij}(t)$. The approximation may be accurate if $\int_0^t f_{ij}(z) dz$ is small.

Regular Arrivals of Batches at an Initial Node

Assume a process of arrivals of random sized batches where arrivals occur at fixed instants $0 < t_1 < t_2 < \dots < t_k < t$. Counts of individuals in arriving batches have a common distribution with mean m_i and variance v_i . All arrivals are assumed to occur at initial node i . The marginal distribution of the count of individuals at node j at time t then has mean and variance

$$E(Y_{ij}(t)) = m_i \sum_{k=1}^{[t]} f_{ij}(t - t_k) \tag{5.1}$$

and

$$V(Y_{ij}(t)) = m_i \sum_{k=1}^{[t]} f_{ij}(t - t_k) [1 - f_{ij}(t - t_k)] + v_i \sum_{k=1}^{[t]} f_{ij}^2(t - t_k), \tag{5.2}$$

where $[t]$ is the largest integer not exceeding t ($i, j = 1, \dots, n$). Comparison of Equations (3.2) and (5.2) show that the loss of the source of variation in the number of arriving batches in an interval of fixed length reduces the variability of the count of individuals at node j by a factor proportional to the square of the mean size of an arriving batch. The models of regular and random arrivals at an initial node provide bounds for the variability of counts at a node, given the constancy of other factors of the system S .

CASE EXAMPLE

Transport of conservative dissolved chlorides through lakes Huron, Erie, and Ontario was modeled as a four node (compartment) open system in which the discharge from Lake Ontario fed the fourth node, assumed to be a sink. An "individual" is a single part per million of the dissolved species in a standard volume of water (one liter). An individual chloride particle is in state (node) 1, 2, 3, or 4 if it is in Lake Huron, Lake Erie, Lake Ontario, or the discharge from Lake Ontario. No other states of the species are permitted. Movement through the system is unidirectional: Huron to Erie to Ontario to sink. Individuals may initially enter S at node 1, 2, or 3 only. Movements of particles are assumed to occur independently of each other after entry into the system. Two systems of causes are assumed to exist, independently of each other, which generate discharges of chlorides into each lake from watershed activities:

(i) background or natural physiochemical processes active in each watershed that produce chloride discharges at a constant intensity, randomly in time;

(ii) artificial processes that produce discharges at a linearly increasing rate, randomly in time.

Processes of discharges are assumed to be Poisson with intensities

$$a_1(t) = c_{11} + c_{12}t \quad (\text{Huron}),$$

$$a_2(t) = c_{21} + c_{22}t \quad (\text{Erie}),$$

$$a_3(t) = c_{31} + c_{32}t \quad (\text{Ontario}).$$

While Lake Huron receives inputs of chlorides only from its watershed discharges, Lakes Erie and Ontario receive inputs from the additional source of discharges from the lake immediately upstream. Consequently, discharges into a lake will influence concentrations at all lakes downstream.

The matrix P describing movement of individual particles among nodes is

$$P = \begin{array}{c} \begin{array}{cccc} & 1 & 2 & 3 & 4 \\ \begin{array}{l} 1 \\ 2 \\ 3 \\ 4 \end{array} & \left[\begin{array}{cccc} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{array} \right] \end{array} \end{array}$$

Assuming that individual chloride particles could have been discharged at any location along shorelines of the lakes and hence that residence times are randomly distributed, a residence time distribution function matrix of exponentials is assumed:

$$W = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{bmatrix} w_{11}(t) & w_{12}(t) & w_{13}(t) & w_{14}(t) \\ w_{21}(t) & w_{22}(t) & w_{23}(t) & w_{24}(t) \\ w_{31}(t) & w_{32}(t) & w_{33}(t) & w_{34}(t) \\ w_{41}(t) & w_{42}(t) & w_{43}(t) & w_{44}(t) \end{bmatrix} \end{matrix},$$

where

$$w_{12}(t) = 1 - e^{-u_1 t}; \quad w_{23}(t) = 1 - e^{-u_2 t}, \quad w_{34}(t) = 1 - e^{-u_3 t};$$

all other entries are identically zero; and the parameters u_1 , u_2 , and u_3 are reciprocals of mean residence times. Interval transition probabilities are computed following the formula given above:

$$f_{11}(t) = e^{-u_1 t}, \quad f_{12}(t) + \frac{u_1}{u_1 - u_2} (e^{-u_2 t} - e^{-u_1 t}),$$

$$f_{13}(t) = \frac{u_1 u_2}{u_1 - u_2} \left(\frac{1}{u_2 - u_3} (e^{-u_3 t} - e^{-u_2 t}) - \frac{1}{u_1 - u_3} (e^{-u_3 t} - e^{-u_1 t}) \right),$$

$$f_{14}(t) = 1 - f_{11}(t) - f_{12}(t) - f_{13}(t),$$

$$f_{21}(t) = 0, \quad f_{22}(t) = e^{-u_2 t}, \quad f_{23}(t) = \frac{u_2}{u_2 - u_3} (e^{-u_3 t} - e^{-u_2 t}),$$

$$f_{24}(t) = \frac{u_2 u_3}{u_2 - u_3} \left(\frac{1}{u_3} (1 - e^{-u_3 t}) - \frac{1}{u_2} (1 - e^{-u_2 t}) \right),$$

$$f_{31}(t) = f_{32}(t) = 0, \quad f_{33}(t) = e^{-u_3 t};$$

$$f_{34}(t) = 1 - e^{-u_3 t}.$$

Define random variables $Y_j(t)$ ($j = 1, 2, 3, 4$) to describe the counts of individual particles in, respectively, Lakes Huron, Erie, Ontario, and the discharge from Lake Ontario. From the theory and assumptions made above, all counts are Poisson distributed. Counts are defined as parts per million or numbers of individual parts of chloride per million parts of water in a typical one liter volume of water in each of the four compartments. A typical one liter volume represents a node in the abstract system representation. Then

$$Y_1(t) = Y_{11}(t),$$

$$Y_2(t) = Y_{12}(t) + Y_{22}(t),$$

$$Y_3(t) = Y_{13}(t) + Y_{23}(t) + Y_{33}(t), \quad Y_4(t) = Y_{14}(t) + Y_{24}(t) + Y_{34}(t) + Y_{44}(t).$$

Net balances of individual particles at each of the four nodes are Poisson distributed with expectations

$$E(Y_1(t)) = c_{11} \frac{1}{u_1} (1 - e^{-u_1 t}) + c_{12} \frac{1}{u_1^2} [(u_1 t - 1) + e^{-u_1 t}],$$

$$\begin{aligned} E(Y_2(t)) = & c_{11} \left[\frac{u_1}{u_1 - u_2} \left(\frac{1}{u_2} (1 - e^{-u_2 t}) - \frac{1}{u_1} (1 - e^{-u_1 t}) \right) \right] \\ & + c_{12} \left[\frac{u_1}{u_1 - u_2} \left(\frac{1}{u_2^2} [(u_2 t - 1) + e^{-u_2 t}] - \frac{1}{u_1^2} [(u_1 t - 1) + e^{-u_1 t}] \right) \right] \\ & + c_{21} \left(\frac{1}{u_2} (1 - e^{-u_2 t}) \right) \\ & + c_{22} \left(\frac{1}{u_2^2} [(u_2 t - 1) + e^{-u_2 t}] \right), \end{aligned}$$

$$E(Y_3(t)) = c_{11} g_{11}(t) + c_{12} g_{12}(t) + c_{21} g_{21}(t) + c_{22} g_{22}(t)$$

$$+ c_{31} g_{31}(t) + c_{32} g_{32}(t),$$

where

$$g_{11}(t) = \frac{u_1 u_2}{u_1 - u_2} \left[\frac{1}{u_2 - u_3} \left(\frac{1}{u_3} (1 - e^{-u_3 t}) - \frac{1}{u_2} (1 - e^{-u_2 t}) \right) - \frac{1}{u_1 - u_3} \left(\frac{1}{u_3} (1 - e^{-u_3 t}) - \frac{1}{u_1} (1 - e^{-u_1 t}) \right) \right],$$

$$g_{12}(t) = \frac{u_1 u_2}{u_1 - u_2} \left[\frac{1}{u_2 - u_3} \left(\frac{1}{u_3^2} [(u_3 t - 1) + e^{-u_3 t}] - \frac{1}{u_2^2} [(u_2 t - 1) + e^{-u_2 t}] \right) - \frac{1}{u_1 - u_3} \left(\frac{1}{u_3^2} [(u_3 t - 1) + e^{-u_3 t}] - \frac{1}{u_1^2} [(u_1 t - 1) + e^{-u_1 t}] \right) \right],$$

$$g_{21}(t) = \frac{u_2}{u_2 - u_3} \left(\frac{1}{u_3} (1 - e^{-u_3 t}) - \frac{1}{u_2} (1 - e^{-u_2 t}) \right),$$

$$g_{22}(t) = \frac{u_2}{u_2 - u_3} \left(\frac{1}{u_3^2} [(u_3 t - 1) + e^{-u_3 t}] - \frac{1}{u_2^2} [(u_2 t - 1) + e^{-u_2 t}] \right);$$

$$g_{31}(t) = \frac{1}{u_3} (1 - e^{-u_3 t}),$$

$$g_{32}(t) = \frac{1}{u_3^2} [(u_3 t - 1) + e^{-u_3 t}],$$

$$E(Y_4(t)) = \sum_{i=1}^3 \left(\int_0^t a_i(z) dz - E(Y_i(t)) \right).$$

Table 1 exhibits a set of chloride measurements (parts per million) randomly sampled from a larger data set accumulated over the period 1902-1960 (Hydroscience [2]). Using estimates of mean residence times of conservative particles obtained from a hydrological balance model of the Great Lakes basin ($1/u_1 = 22.6$ yr, $1/u_2 = 2.6$ yr, $1/u_3 = 7.9$ yr), chloride discharge intensities $a_1(t)$, $a_2(t)$, and $a_3(t)$ were estimated by computing

TABLE 1
SAMPLED CONCENTRATIONS OF CHLORIDES IN
LOWER GREAT LAKES, 1900-1960

Date	Concentration (ppm)		
	Huron	Erie	Ontario
1906	2.5		7.5, 9.0
1907		9.0	
1909		10.5	
1927			11.0
1928			11.5
1929		12.0	
1932			13.0
1934	4.5	15.5	
1935	3.5		15.5
1936	5.5		
1937	3.5		16.0
1938			14.0
1941			17.0
1942			17.0, 15.0
1944			16.5
1947	7.0		18.0, 18.5, 19.0
1948		19.5, 21.0	17.5
1949		18.0	23.0
1950	6.0		19.5
1952	4.5, 5.5	18.5	19.0, 20.0, 21.0
1953			22.0
1954			21.0, 21.5
1955			23.0
1956	7.5	26.5	
1957		21.5, 26.5	20.0, 24.0
1958		27.0, 22.5	24.0, 25.0, 26.0
1960	7.0	23.0, 24.0	

optimized estimates of the coefficients c_{ij} from a linear goal program. The expectations $E(Y_j(t))$ are linear functions of the input intensity coefficients, and when the exponential coefficients of the c_{ij} 's are calculated for the same years for which chloride measurements are available, a set of linear combinations of the unknowns c_{ij} are obtained. By setting each linear combination equal to the appropriate measured chloride density a set of constraints, linear in the c_{ij} 's, become the constraining equations of the goal program. The objective function of the program is the sum of absolute deviations of computed concentrations from measured concentrations. Applying a simplex

algorithm, the combination of c_{ij} 's is found which minimizes the sum of deviations. The following estimates were obtained:

$$\hat{a}_1(t) = 0.158 + 0.004t,$$

$$\hat{a}_2(t) = 0.504 + 0.048t,$$

$$\hat{a}_3(t) = 0$$

Expected concentrations together with one and two sigma probability limits of the Poisson distributions were computed and plotted against measured concentrations, including both those that were used to estimate the c_{ij} 's and those that were not. Figures 1-3 show the comparisons of computed versus measured concentrations. In the first years of the series measured values exceeded computed, clearly indicating that, contrary to the model assumption of zero initial concentrations, chlorides were present in the water columns of the lakes. The computed mean concentration for Lake Erie fell consistently below the measured concentration even though the estimates of the input intensity functions had been previously optimized. Had the lakes not been coupled as a system and discharge intensities been estimated separately, such an underestimate would not have occurred. Since the lakes are coupled, however, it would be a modeling error to ignore coupling. Causes of the systematic underestimate of concentration in Lake Erie may be due to any one or combination of the following:

- (1) errors in measurements of field densities;
- (2) nonrepresentativeness of measured densities, leading to an unrealistic distribution of actual field conditions;
- (3) residence time distributions incorrectly specified, particularly with respect to the means;
- (4) unequal representation of concentrations in the lakes due to unequal numbers of constraints in the optimizing program.

Differences in the numbers of constraints for the lakes will affect estimates $\hat{a}_2(t)$ and $\hat{a}_3(t)$ if measured concentrations in the two lakes differ by a substantial amount. Since this difference did not exist, unequal numbers of constraints in the program are unlikely to have had any effect.

Incorrectly specified residence time distributions will affect computed concentrations and hence comparisons between computed and measured concentrations, particularly when means are incorrectly estimated. An error in the mean $1/u_2$ for Lake Erie will cause errors in computed concentrations for both Lakes Erie and Ontario. A sensitivity analysis by Patterson [3]

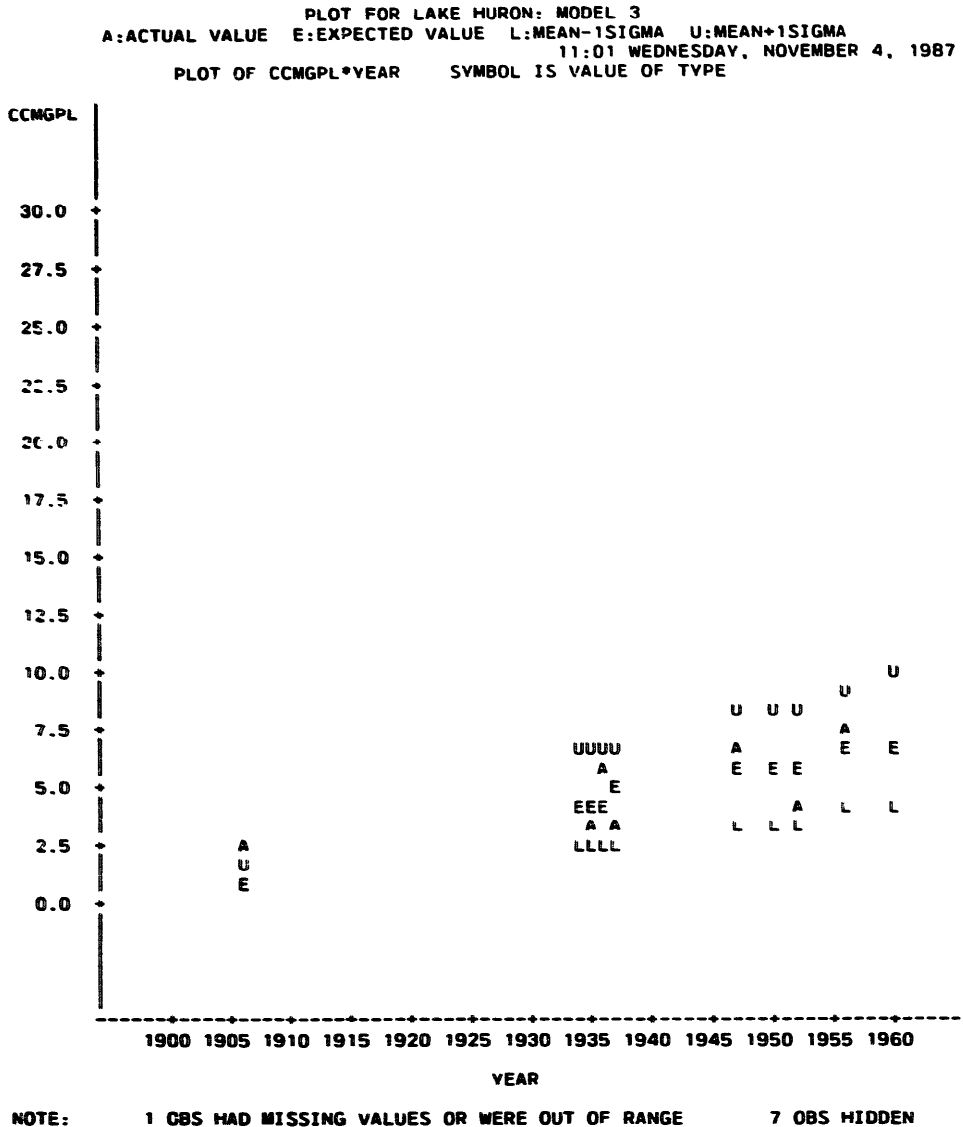
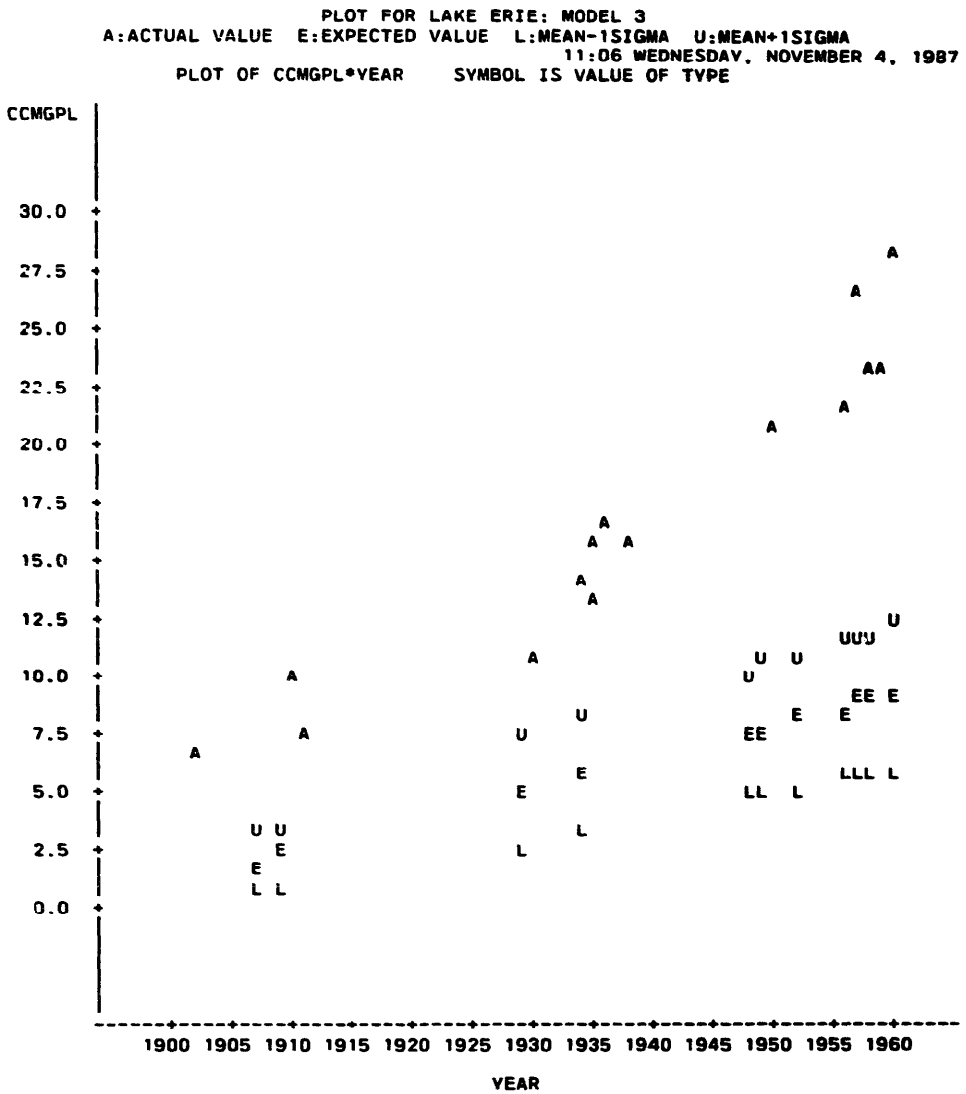


FIG. 1. Measured versus computed chloride concentration: Huron. $u_1^{-1} = 22.6$ yr. A = measured value, E = computed mean, U = computed mean + 1 std. dev., L = computed mean - 1 std. dev., CCMGPL = conc. of chloride (mg/liter).

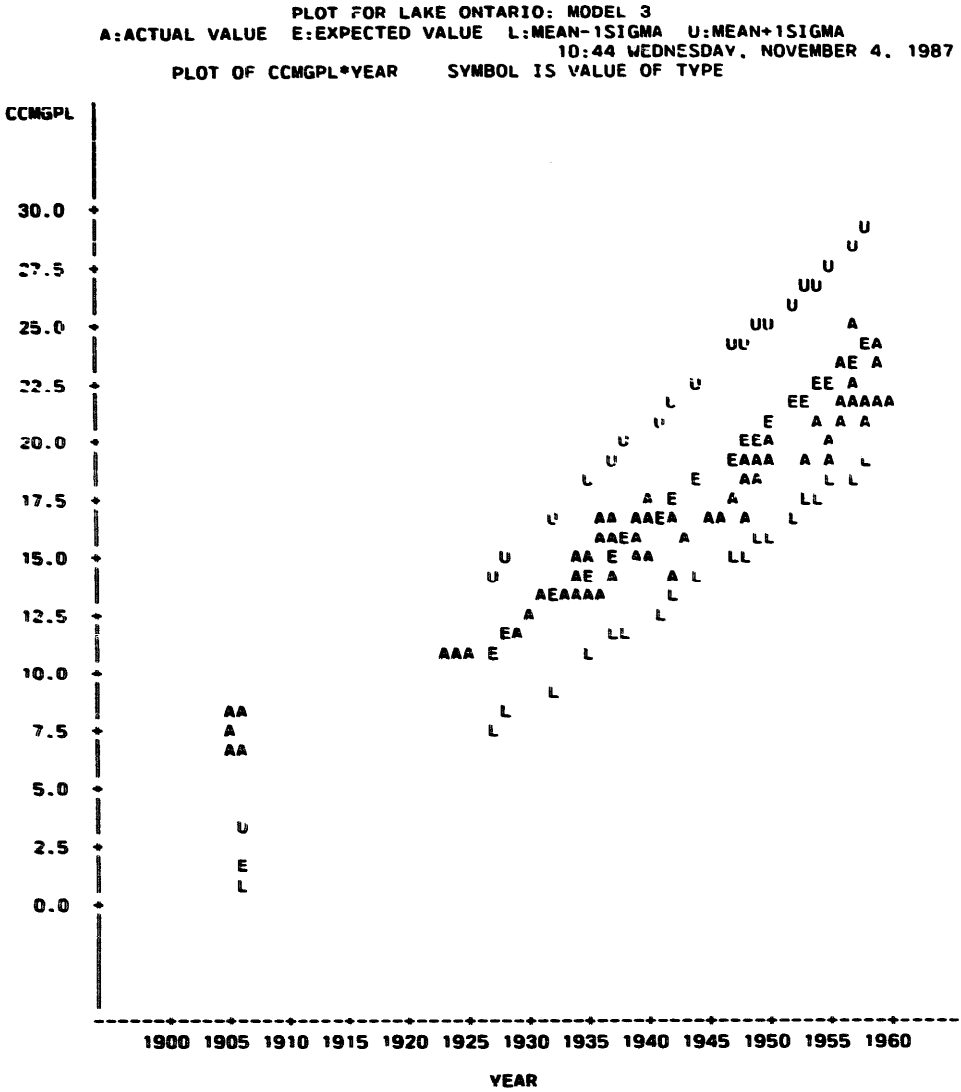
showed that computed concentrations are much more sensitive to variations in the mean residence time than in other parameters of the residence time distribution, a result that supports a theoretical finding that at equilibrium only the mean of the residence time distribution affects the actual concentration of individuals. Increases in computed values of mean concentrations in Lake Erie were achieved by increasing the mean residence time, reoptimizing



NOTE: 12 OBS HIDDEN

FIG. 2. Measured versus computed chloride concentration: Erie. $u_1^{-1} = 22.6$ yr, $u_2^{-1} = 2.6$ yr. A-measured value, E-computed mean, U = computed mean + 1 std. dev., L = computed mean - 1 std. dev., CCMGPL = conc. of chloride (mg/liter).

the fit of the discharge intensity to the data, and recomputing mean concentrations. An increase in overall fit of the four-compartment model resulted each time the mean residence time of individuals in Lake Erie was increased. When the mean residence times exceeded 1.4 times the initial estimated value of 2.6 yr, the increase in fit was slight. Similarly, a decrease in mean residence time of chloride particles in Lake Ontario below the initial esti-



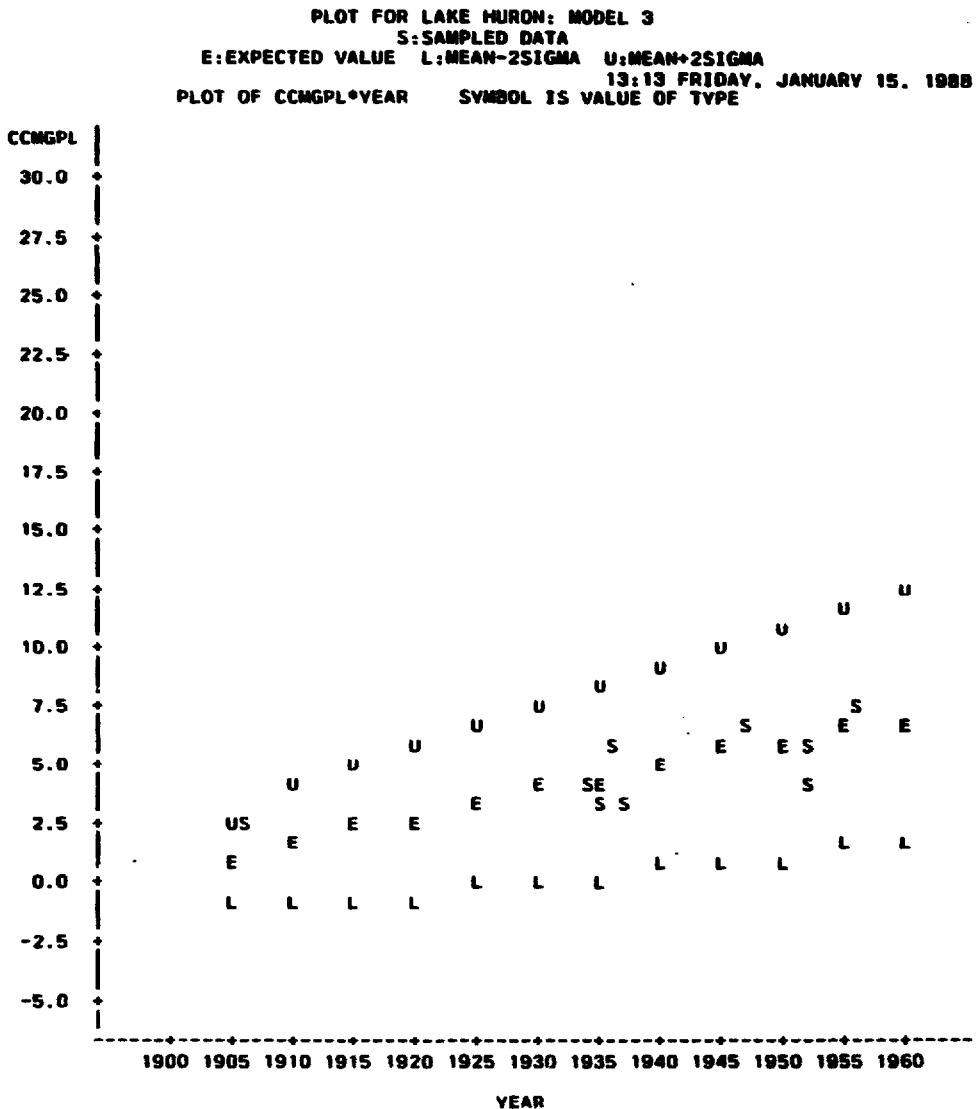
NOTE: 46 OBS HIDDEN

FIG. 3. Measured versus computed chloride concentration: Ontario. $u_1^{-1} = 22.6$ yr, $u_2^{-1} = 2.6$ yr, $u_3^{-1} = 7.9$ yr. A = measured value, E = computed mean, U = computed mean + 1 std. dev., L = computed mean - 1 std. dev., CCMGPL = conc. of chloride (mg/liter).

mated value of 7.9 yr resulted in improved fits. This result is not unexpected, as an increase in the turnover rate of water in the lake is offset by an increase in the estimated discharge intensity of chloride particles directly into Lake Ontario. Little insight is gained into the dynamics of the system by recognizing the basic tradeoff between residence time in a compartment and input rate, unless constraints can be placed on either input rates or residence times

from data not used in fitting the model, preferably data which directly measure these rates rather than measurements of net balances in concentrations.

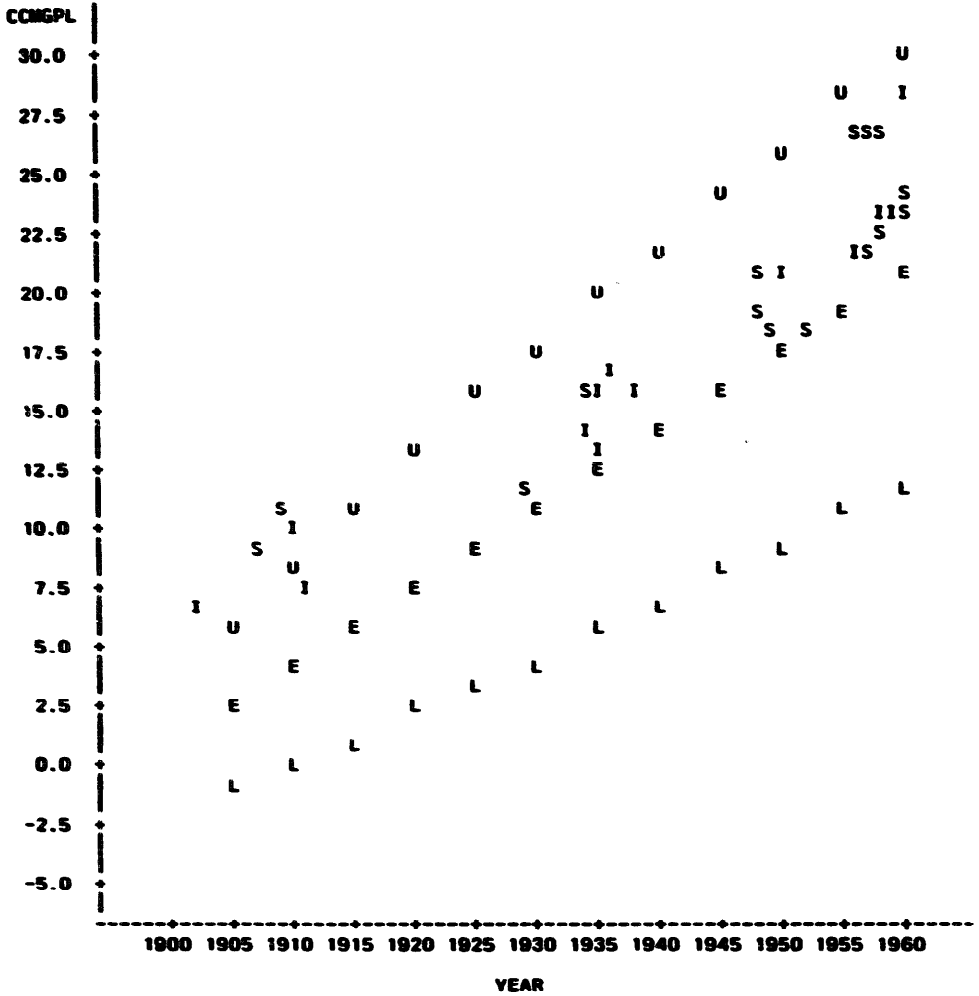
Figures 4-6 show comparisons between computed and measured concentrations when the means u_2^{-1} and u_3^{-1} were adjusted to 3.6 and 4.7 yr, respectively. At these values the model accounts reasonably well for mea-



NOTE: 2 OBS HIDDEN

FIG. 4. Measured versus computed chloride concentration: Huron. $u_1^{-1} = 22.6$ yr. S = measured value, E = computed mean, U = computed mean + 2 std. dev., L = computed mean - 2 std. dev., CCMGPL = conc. of chloride (mg/liter).

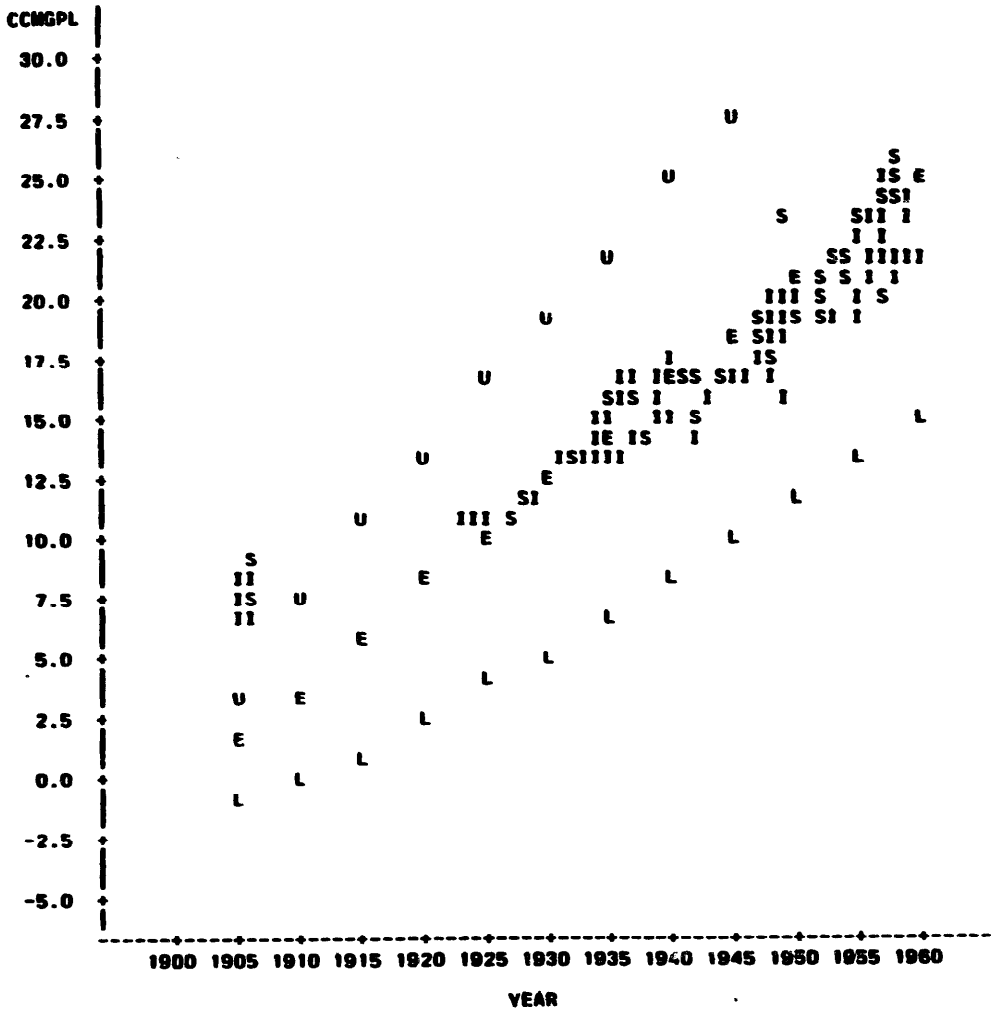
PLOT FOR LAKE ERIE: MODEL 3
 S: SAMPLED DATA I: NOT SAMPLED DATA
 E: EXPECTED VALUE L: MEAN-2SIGMA U: MEAN+2SIGMA
 13:17 FRIDAY, JANUARY 15, 1988
 PLOT OF CCMGPL*YEAR .SYMBOL IS VALUE OF TYPE



NOTE: 2 OBS HIDDEN

FIG. 5. Measured versus computed concentration: Erie. $u_1^{-1} = 22.6$ yr, $u_2^{-1} = 3.6$ yr. S = measured value included in model fit, I = measured value excluded from model fit, E = computed mean, U = computed mean + 2 std. dev., L = computed mean - 2 std. dev., CCMGPL = conc. of chloride (mg/liter).

PLOT FOR LAKE ONTARIO: MODEL 3
 S: SAMPLED DATA I: NOT SAMPLED DATA
 E: EXPECTED VALUE L: MEAN-2SIGMA U: MEAN+2SIGMA
 13:22 FRIDAY, JANUARY 15, 1988
 PLOT OF CCMGPL*YEAR SYMBOL IS VALUE OF TYPE



NOTE: 3 OBS HAD MISSING VALUES OR WERE OUT OF RANGE 19 OBS HIDDEN

FIG. 6. Measured versus computer concentration: Ontario. $u_1^{-1} = 22.6$ yr, $u_2^{-1} = 3.6$ yr, $u_3^{-1} = 4.7$ yr. S = measured value included in model fit, I = measured value excluded from model fit, E = computed mean, U = computed mean + 2 std. dev., L = computed mean - 2 std. dev., CCMGPL = conc. of chloride (mg/liter).

sured concentrations in the three lakes for the fifty year period 1910–1960. Even with these revised estimates, which are 40 percent above and below the estimates provided by the watershed balance model, respectively, no direct discharge from the Lake Ontario watershed is required to offset the increased turnover rate of water in Lake Ontario. For those values of mean residence times in the lakes, the optimized estimates of the discharge intensities of chlorides are

$$\hat{a}_1(t) = 0.155 + 0.004t,$$

$$\hat{a}_2(t) = 0.623 + 0.086t,$$

$$\hat{a}_3(t) = 0.$$

Comparison of $\hat{a}_2(t)$ for the two sets of residence times in Lakes Erie and Ontario shows that an increase in both constant and time-varying input intensity for Lake Erie is sufficient to account for an increase in computed mean balance in Lake Erie while maintaining the balance in Lake Ontario. Only when the residence time of chlorides in Lake Ontario is reduced by 95 percent of the initial value is a direct discharge into Lake Ontario required to maintain a balance that matches measured concentrations. It is clear that, instead of the model being overparametrized (a case can be made for retaining each parameter), additional data are needed to establish limits outside of which parameters should not be varied in the fitting process. If either input intensities or residence times are given, the other parameters can be more reliably estimated.

It is interesting to note that relative magnitudes of parameter estimates remained constant in the two estimates of input intensities corresponding to the two sets of mean residence times. For comparison a third set of mean residence times for Lakes Erie and Ontario were used, which were set at 95 percent above and below their original values, respectively. Table 4 summarizes the results. Significantly, relative magnitudes of coefficients changed, the constant input intensity assuming a much greater proportion of the total. A second qualitative difference between this case and the others is the substantial discharge directly into Lake Ontario from its watershed needed to balance concentrations in that lake. It is implausible that the mean residence time for chlorides in Lake Ontario is as short as 0.4 yr.

Tables 2–4 break down total discharges of chlorides into the lakes according to input component and lake, for a typical one liter volume, for the sixty year period. Bounds on total discharges may be computed corresponding to

TABLE 2
ESTIMATED CUMULATIVE CHLORIDE DISCHARGES, 1900-1960
 $u_1^{-1} = 22.6$ yr, $u_2^{-1} = 2.6$ yr, $u_3^{-1} = 7.9$ yr.

Lake	Estimated discharge intensity (ppm/yr)		Cumulative discharge (ppm)		Total
	Const.	Lin. incr.	$\int_0^{60} (\text{Const.}) dt$	$\int_0^{60} (\text{Lin. Incr.}) dt$	
Huron	$\hat{a}_1(t) = 0.158$	$+0.004t$	9.48	7.20	16.68
Erie	$\hat{a}_2(t) = 0.504$	$+0.048t$	30.24	86.40	116.64
Ontario	$\hat{a}_3(t) = 0$		0	0	0
Total			39.72	93.60	133.32

TABLE 3
ESTIMATED CUMULATIVE CHLORIDE DISCHARGES, 1900-1960
 $u_1^{-1} = 22.6$ yr, $u_2^{-1} = 3.64$ yr, $u_3^{-1} = 4.74$ yr.

Lake	Estimated discharge intensity (ppm/yr)		Cumulative discharge (ppm)		Total
	Const.	Lin. Incr.	$\int_0^{60} (\text{Const.}) dt$	$\int_0^{60} (\text{Lin. Incr.}) dt$	
Huron	$\hat{a}_1(t) = 0.155$	$+0.004t$	9.30	7.20	16.50
Erie	$\hat{a}_2(t) = 0.623$	$+0.086t$	37.38	154.80	192.18
Ontario	$\hat{a}_3(t) = 0$		0	0	0
Total			46.68	162.00	208.68

TABLE 4
ESTIMATED CUMULATIVE CHLORIDE DISCHARGES, 1900-1960
 $u_1^{-1} = 22.6$ yr, $u_2^{-1} = 5.1$ yr, $u_3^{-1} = 0.4$ yr.

Lake	Estimated discharge intensity (ppm/yr)		Cumulative discharge (ppm)		Total
	Const.	Lin. Incr.	$\int_0^{60} (\text{Const.}) dt$	$\int_0^{60} (\text{Lin. Incr.}) dt$	
Huron	$\hat{a}_1(t) = 0.155$	$+0.004t$	9.30	6.84	16.14
Erie	$\hat{a}_2(t) = 2.179$	$+0.038t$	130.74	68.40	199.14
Ontario	$\hat{a}_3(t) = 1.568$	$+0.907t$	94.08	1632.60	1726.68
Total			234.12	1707.84	1941.96

any required probability limits by applying normal approximations to the Poisson distribution.

In summary, the above-cited numerical experiment showed that the fitting criterion achieved an overall minimum when the mean residence time of chlorides in Lake Erie was increased to a large value, effectively severing the flow link between Erie and Ontario, while simultaneously decreasing the mean residence time for Lake Ontario and increasing the direct discharge rate of chlorides into Ontario. That is, flow linkages between the lakes cause optimized objective function values to be inflated with respect to the case where lake concentrations are physically independent of each other.

Nonrepresentativeness of sampling stations in the lakes is a potential cause of large differences between measured and computed concentrations even after input intensities are fitted to the field data. The condition shown in Figure 2 can be caused when a disproportionate number of sampling locations are outside the main volume of lake water (whose mean turnover time is u_2^{-1}), being situated in local pockets with low turnover rate where high concentrations accumulate.

Differences in measured versus computed variability exist in Figures 1–6, the measured variability being lower. The two modeling assumptions called into question in this case are (1) completely randomized timing of discharges represented by the Poisson distribution, and (2) complete and instantaneous mixing of chlorides upon entering the lakes. Since the Poisson distribution approximates a wide variety of distributions of discharges, it is more plausible that the second of the two assumptions is in error. Clearly, instantaneous and complete mixing of lake volumes is impossible, implying that the exponential distribution of residence times is inaccurate, particularly when lake volumes are not subcompartmentalized. Lack of subcompartmentalization is mitigated, in part, by the plausible assumption that chlorides are discharged randomly throughout shoreline waters of the lakes. A second point that provides justification for use of the exponential distribution of residence times is the theoretical result cited above concerning insensitivity of equilibrium concentrations to all moments of the residence time distribution except the first, supported by the sensitivity analysis by Patterson.

CONCLUSIONS

Stochastic compartmental models are a numerically feasible method of analyzing the distribution and fate of chemical species in aquatic environments, subject to availability of data for parameter estimation. A multivariate, linear, stochastic compartmental model can easily be fitted to a multivariate time series in the dependent variables. Three advantages of the stochastic

version of the linear compartmental model over its deterministic counterpart are:

- (1) formal methods of statistical inference can be applied when parameters are estimated from field data,
- (2) probability bounds on estimates of dependent variables are obtained, and
- (3) whereas linear, deterministic compartmental models always invoke the Markov assumption in their formulation, the stochastic version as described above by the matrices P and W is computationally feasible for general residence time distributions.

Finally, the stochastic compartmental model permits incorporation of a range of sources of uncertainty, including (i) types of external inputs, (ii) compartmental residence times, and (iii) distribution of individuals among compartments.

REFERENCES

- 1 P. Whittle, *Stochastic Networks in Equilibrium*, Wiley, 1986.
- 2 Hydrosience, Inc., *Limnological Systems Analysis of the Great Lakes. Phase I—Preliminary Model Design*, 1973.
- 3 R. L. Patterson, Sensitivity analyses of a $M(t)/G/\infty$ stochastic service system, *Appl. Math. Comput.*, 26:169–177 (1988).