Probabilistic methods for addressing uncertainty and variability in biological models: Application to a toxicokinetic model

H.T. Banks [a,*], Laura K. Potter [a,1]

[a] *Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC 27695-8205 USA*

Abstract

Population variability and uncertainty are important features of biological systems that must be considered when developing mathematical models for these systems. In this paper we present probability-based parameter estimation methods that account for such variability and uncertainty. Theoretical results that establish well-posedness and stability for these methods are discussed. A probabilistic parameter estimation technique is then applied to a toxicokinetic model for trichloroethylene using several types of simulated data. Comparison with results obtained using a standard, deterministic parameter estimation method suggests that the probabilistic methods are better able to capture population variability and uncertainty in model parameters.

Key words: Parameter estimation, biological modeling, population variability, model uncertainty, toxicokinetics, trichloroethylene

1 Introduction

Uncertainty is an inherent factor in mathematical models for biological systems. Model equations themselves are an approximation of the phenomena they are designed to model, introducing a degree of uncertainty that is difficult to measure. Further simplifications and approximations of a model for

* Author to whom correspondence should be addressed.
Email addresses: htbanks@eos.ncsu.edu (H.T. Banks), laura.k.potter@gsk.com (Laura K. Potter).
[1] Current address: Scientific Computing and Mathematical Modeling, GlaxoSmithKline, Research Triangle Park, NC 27709 USA.

theoretical and computational purposes result in additional layers of uncertainty. Moreover, many biological processes are subject to variability that may not be incorporated into a mathematical model. Experimental observations also introduce uncertainty when data are used with a model to estimate parameters.

Two types of variability that are common in biological models and are well-known in the statistical literature are intra-individual and inter-individual variability. *Intra-individual* variability is defined as variability that occurs within a given individual organism or biological process. This type of variability may result in time-dependent and/or spatially-dependent variation within an individual. Biological examples of such variability include parameters such as body weight, blood pressure, fat content and cell membrane permeabilities.

A second type of variability that is commonly found in biological modeling is *inter-individual* variability. This type of variability results from variations in individuals across a population. Biological models that are based on behavior or phenomena over a population are almost always subject to inter-individual variability. This is especially the case when a model is designed to predict or explain experimental observations that are collected from multiple individuals.

It is reasonable to expect that different individuals of a population would possess different values for biological, physical and chemical parameters. These parameters would then take on a range of values across the population, so that each parameter would be associated with a probability distribution that would mathematically describe this variation. Using data from multiple individuals, the resulting probability distributions can be estimated with inverse problem techniques.

Examples of biological parameters that are often subject to inter-individual variability include growth and death rates, susceptibility to infection, efficacy of vaccines and other prophylactics, and age. Note that each of the examples given above for intra-individual variability may also involve inter-individual variability depending on the type of model and experimental observations. Similarly, each of the examples for inter-individual variability also may be subject to intra-individual variability.

The motivating example we consider here is a toxicokinetic model for the systemic transport of the environmental contaminant trichloroethylene (TCE). TCE is a solvent that has been used widely in industry as a metal degreasing agent, and is now a common soil and groundwater contaminant. This highly fat-soluble compound is rapidly absorbed into the bloodstream, and has been shown to accumulate in the adipose (fat) tissue of humans and animals [1,2]. Known and suspected toxic effects of TCE and its metabolites in laboratory animals and/or humans include acute effects such as dizziness, drowsiness,

headaches and fatigue, as well as chronic effects such as developmental defects and lung, kidney and liver tumors [3–10].

Toxicokinetic models are used in the overall risk assessment process for toxic compounds to help quantify the expected risk of toxicity to humans as a function of the level of exposure to the given chemical. In particular, physiologically based pharmacokinetic (PBPK) models predict the effective dose level of a toxic compound that is delivered to the "target" tissues (i.e., tissues that experience toxic effects) for a given external exposure level. PBPK models are compartmental models that describe the systemic transport of a compound through the tissues and organs, including the dynamics of uptake, tissue distribution, metabolism and elimination. The resulting model is a system of ordinary, partial and/or delay differential equations, with each equation representing the dynamics of tissue concentrations in a particular tissue or organ.

Several PBPK models have been developed for TCE (e.g., see [11–15]). A majority of these models make use of the standard "perfusion-limited" compartmental model for each of the modeled tissues and organs (see Section 4 for a description of the perfusion-limited model). In [16] and [17], three PBPK models for TCE are developed and compared, each with a different submodel for the adipose tissue compartment.

As discussed in [16], preliminary simulations indicated that a perfusion-limited adipose tissue compartment does not appear to sufficiently capture the dynamics of TCE accumulation in fat as seen in experimental data. Moreover, adipose tissue is known to have highly heterogeneous physiological properties, including significant variations in fat cell size, lipid distribution, blood flow rates and cell membrane permeabilities [18–21]. These characteristics further suggest that the "well-mixed," rapid equilibrium assumptions of the perfusion-limited model may be inappropriate for describing the disposition of fat-accumulating compounds such as TCE in adipose tissue.

To better capture the dynamics of TCE in fat tissue, a spatially varying axial dispersion model was developed [16] to address the *intra-individual* variability that results from the heterogeneous lipid distribution and physiological characteristics of adipose tissue. This variability is built into the adipose compartmental model with a special axial dispersion term, where the "dispersion" coefficient is a measure of the degree of intra-individual variability that occurs in the fat.

In addition to the intra-individual variability that appears to affect TCE concentrations in fat tissue, *inter-individual* variability also plays a major role in toxicokinetic models in general. Current technology almost always requires that measurements of chemical concentrations in tissues over time must be

taken from multiple individuals. This experimental necessity immediately introduces inter-individual variability into the measured observations, and must be considered in the development of mathematical models.

As biological models have become more widely utilized and influential in a variety of fields, the need to account for variability and uncertainty in modeling has been recognized. Markov Chain-Monte Carlo methods have been developed to address issues of variability and uncertainty, and these methods have been applied to PBPK models as a part of the parameter estimation process. Monte Carlo methods are based on a Bayesian statistical approach that involves the use of experimental data to update estimates of a hypothesized "prior" probability distribution for one or more model parameters. Examples of Monte Carlo methods applied to PBPK models can be found in [22–31].

An alternative, probability-based method has been developed to incorporate uncertainty and variability in mathematical models. This method, which is discussed in [32–34] and is detailed in Section 2, is centered around a probabilistic parameter estimation approach that involves the estimation of probability distributions for model parameters. Well-known theoretical results from probability theory establish the theoretical soundness of this technique, which can be implemented computationally in a straightforward manner.

A distinct advantage of this approach over the Monte Carlo-based methods is an added level of flexibility in choosing the prior probability distributions. As we discuss in Section 2, these probability-based methods can be used with preselected prior distributions as with Monte Carlo methods, or they may be used with weighted sums of Dirac delta measures that do not assume a fixed form for the probability distribution functions. A version of this method has been applied to a population model for mosquitofish in rice paddies, and was used to successfully describe fish population dynamics by estimating distributed growth rate functions using aggregate experimental data [35].

In this paper we present probability-based parameter estimation methods for incorporating uncertainty and variability into biological models. These methods are general and may be applied to a wide variety of models to account for various types of model uncertainty as we have outlined here. In Section 2 we formulate these probabilistic parameter estimation methods in a general setting. We address related theoretical issues in Section 3, establishing the well-posedness of the resulting parameter estimation process. Finally, implementation of the methods in the context of a toxicokinetic model for TCE is discussed in Section 4.

2 General problem formulation

Suppose that a biological process is described by the parameter-dependent system

$$\frac{dy}{dt} = f(t, y(t), q), \tag{1}$$

where y represents the state of system, q is the vector of parameters on which the state depends, and f represents the dynamics of the system in the form of ordinary, partial or functional differential equations. Experimental data $z = \{z_i\}$, $i = 1, \ldots, N_t$ are given that correspond to complete or partial observations $\mathcal{O}y(t_i; q)$ of the state.

The model parameters q are estimated in a deterministic, least squares setting by minimizing the objective function

$$J(q, z) = \sum_{i=1}^{N_t} |\mathcal{O}y(t_i; q) - z_i|^2 \tag{2}$$

over $q \in Q$ subject to (1), where Q is the space of admissible parameters and $\mathcal{O}$ is the observation operator.

In standard least squares estimation problems, the space Q is usually defined as a compact subset of $\mathbb{R}^n$ for some positive integer n, so that $q = (q_1, q_2, \ldots, q_n)$ is a vector with real components. This assumption requires each parameter to be a constant, which may not be reasonable for parameters and experimental data that are subject to inter-individual variability. Indeed, if experimental observations are collected from multiple individuals in a population, then one must think of the constant parameters $q \in \mathbb{R}^n$ as *average* values over the sampled population. This approximation may be appropriate for some parameters that do not vary to a large extent across individuals, but in many cases these "mean" value approximations may lead to inaccurate parameter estimates and an inaccurate description of the population. This is especially true in situations when subpopulations are described by different parameter values, or means, variances, etc.

Such population-dependent variability in model parameters can be incorporated into least squares estimation problems using a probability-based formulation. We assume that the model parameters q are realizations of random variables with probability distributions P that vary over the population, so that P belongs to a probability space $\mathcal{Q}$ that *may* be infinite dimensional. As in [33], we define the set $\mathcal{P}(Q)$ of all probability distributions on the admissible

parameter space Q and seek a probability distribution function $\bar{P} \in \mathcal{Q} \subseteq \mathcal{P}(Q)$ that minimizes the objective function

$$J(P, \hat{z}) = \sum_{i=1}^{N_t} |\mathcal{E}[\mathcal{O}y(t_i; q)|P] - \hat{z}_i|^2 \tag{3}$$

over $\mathcal{Q} \subseteq \mathcal{P}(Q)$, where $\hat{z}_i$, $i = 1, \ldots, N_t$ are observations corresponding to the expected value

$$\mathcal{E}[\mathcal{O}y(t_i; q)|P^*] = \int_Q \mathcal{O}y(t_i; q) dP^*(q) \tag{4}$$

for some $P^* \in \mathcal{Q} \subseteq \mathcal{P}(Q)$. For simplicity we often choose $\mathcal{Q} = \mathcal{P}(Q)$, but this is not essential and one may readily restrict the family of admissible distributions in certain formulations.

Depending on the choice of the set $\mathcal{Q} \subseteq \mathcal{P}(Q)$ of probability distributions, this method may be implemented with pre-determined "prior" probability distributions (as with the Monte Carlo method), or it may be used without the pre-specification of a particular probability distribution. For the case when there is a reasonable expectation that a parameter varies across the population in a manner similar to a given probability distribution, the set $\mathcal{Q}$ can be chosen as the space of those distribution functions (e.g., log normal distributions) defined over the admissible parameter space Q. For this type of formulation, the distribution functions are uniquely determined by their parameterizations $\tilde{q}$ (e.g., $\tilde{q} = (\mu, \sigma)$, the mean and standard deviation), and hence may be estimated by minimizing

$$J(\tilde{q}, \hat{z}) = \sum_{i=1}^{N_t} |\mathcal{E}[\mathcal{O}y(t_i; q)|P(\tilde{q})] - \hat{z}_i|^2 \tag{5}$$

over the space $\tilde{Q}$ of admissible parameterizations $\tilde{q}$, where $\mathcal{Q}$ is given as the set of probability distributions of the pre-specified form with the parameterizations $\tilde{q}$.

For example, if it is believed that a parameter can be approximated by a log normal distribution, then $\mathcal{Q}$ is given as the set of all *log normal distributions* defined over Q. For this particular formulation, the estimation problem is solved by minimizing (5) over the space $\tilde{Q}$ of admissible mean and standard deviation parameters $\tilde{q} = (\mu, \sigma)$. This approach has been implemented in [36] where $\mathcal{Q}$ is defined as a set of bitruncated normal distributions with certain specified properties, and the estimated parameters $\tilde{q}$ are the mean and a modified standard deviation (see [36] for details).

If it is not possible to predict the expected form of the probability distributions *a priori*, this method also may be used without the specification of prior distributions. In this case, $\mathcal{Q} = \mathcal{P}(Q)$ may be chosen as the space of *all* probability distributions defined on Q. For computational purposes, the estimation problem may then be implemented using finite dimensional approximations to the original infinite dimensional problem. First we define the infinite dimensional set

$$\mathcal{P}_0(Q) \equiv \{P \in \mathcal{P}(Q) : P = \sum_{j=1}^{k} p_j \delta_{q_j}, k \in \mathbb{N}^+, q_j \in Q_0, p_j \geq 0, \sum_{j=1}^{k} p_j = 1\}, \quad (6)$$

where $Q_0 = \{q_j\}_{j=1}^{\infty}$ is a given countable, dense subset of the parameter space Q and δ_{q_j} is the Dirac delta distribution with mass at $q_j \in Q$. In other words, $\mathcal{P}_0(Q)$ is the set of probability distributions on Q that have finite support in Q_0. We then define the finite dimensional set

$$\mathcal{P}^M = \{P \in \mathcal{P}_0(Q) : P = \sum_{j=0}^{M} p_j \delta_{q_j}\}$$

which we use to define a family of finite dimensional approximation problems. That is, for fixed $\{q_0, q_2, \ldots, q_M\}$ in Q_0 with $P_M = \sum_{j=0}^{M} p_j \delta_{q_j} \in \mathcal{P}^M$, we minimize the objective function

$$J(P_M, \hat{z}) = \sum_{i=1}^{N_t} |\mathcal{E}[\mathcal{O}y(t_i; q)|P_M] - \hat{z}_i|^2 \quad (7)$$

$$= \sum_{i=1}^{N_t} \left| \sum_{j=0}^{M} \mathcal{O}y(t_i; q_j) p_j - \hat{z}_i \right|^2 \quad (8)$$

over the set $\mathcal{P}^M$. These precise definitions are necessary to obtain a well-posed estimation problem, as we discuss in Section 3. Note that the problem of minimizing the objective function (8) corresponds to solving a constrained quadratic programming problem for $(p_0, \ldots, p_M)$ with the constraints $p_j \geq 0$, $\sum_{j=0}^{M} p_j = 1$ (see Section 4.3). There currently exist a number of acceptable computational methods to solve such problems which are again special cases of choosing an *a priori* parameterization set ($\mathcal{Q} = \mathcal{P}^M$ in this case) and optimizing over admissible parameterizations

$$\tilde{Q} = \{\tilde{q} = (p_0, \ldots, p_M) : p_j \geq 0, \sum_{j=0}^{M} p_j = 1\}.$$

3 Theoretical results

In order to address theoretical issues related to the inverse problems discussed in Section 2, we need to define a suitable metric for the probability spaces $\mathcal{P}(Q)$. Using the Prohorov metric and well-known results from probability theory, we establish a theoretical framework that allows us to prove method stability for our probability-based parameter estimation problems.

3.1 The Prohorov metric

As in [37], the Prohorov metric is defined on the space of probability measures $\mathcal{P}(Q)$ on the Borel subsets of Q, where Q is a complete metric space with metric d. The Prohorov metric $\rho : \mathcal{P}(Q) \times \mathcal{P}(Q) \to \mathbb{R}^+$ is defined by

$$\rho(P_1, P_2) = \inf\{\varepsilon > 0 : P_1[F] \leq P_2[F^\varepsilon] + \varepsilon;\ F \text{ closed};\ F \subset Q\},$$

where

$$F^\varepsilon = \{q \in Q : d(\bar{q}, q) < \varepsilon;\ \bar{q} \in F\}.$$

It is well known that ρ is a metric on the space $\mathcal{P}(Q)$, and that this metric space $(\mathcal{P}(Q), \rho)$ is complete. Moreover, $(\mathcal{P}(Q), \rho)$ is compact for all compact sets Q.

Another well-known result [37] establishes equivalent conditions for convergence of probability distributions in the Prohorov metric. Assuming that (Q, d) is complete, then the following convergence statements for $P_k, P \in \mathcal{P}(Q)$ are equivalent:

(i) $\rho(P_k, P) \to 0$.
(ii) $\int_Q f(q)dP_k(q) \to \int_Q f(q)dP(q)$ for all bounded and uniformly continuous functions $f : Q \to \mathbb{R}$.
(iii) $P_k[A] \to P[A]$ for all Borel sets $A \subset Q$ with $P[\partial A] = 0$, where ∂A is the boundary of A.

3.2 Stability of the general parameter estimation problem

Banks and Bihari [33] have addressed theoretical issues related to probability-based estimation problems. Using the Prohorov metric, they studied the convergence properties of sequences of probability distributions in $\mathcal{P}(Q)$. These results were then applied to a sequence of minimizers for finite dimensional approximations to the estimation problem for (3). Here we summarize their findings as they relate to the inverse problems described in Section 2.

As discussed in [33], it follows that if the mapping $q \to \mathcal{O}y(t_i; q)$ is continuous, then the convergence $\rho(P_k, P) \to 0$ in the Prohorov metric is equivalent to

$$\mathcal{E}[\mathcal{O}y(t_i; q)|P_k] \to \mathcal{E}[\mathcal{O}y(t_i; q)|P],$$

and hence the map

$$P \to J(P) = \sum_{i=1}^{N_t} |\mathcal{E}[\mathcal{O}y(t_i; q)|P] - \hat{z}_i|^2$$

is continuous in the ρ topology. Moreover, if the space Q is compact we have that $(\mathcal{P}(Q), \rho)$ is a compact metric space, which along with the continuity of the map $P \to J(P)$ guarantees the existence of a minimizer over $\mathcal{P}(Q)$ for the estimation problem associated with

$$\min_{P \in \mathcal{P}(Q)} J(P, \hat{z}) = \sum_{i=1}^{N_t} |\mathcal{E}[\mathcal{O}y(t_i; q)|P] - \hat{z}_i|^2 . \tag{9}$$

In addition to establishing the existence of a solution for the inverse problem (9), Banks and Bihari [33] developed results related to method stability for this problem. Using finite dimensional approximation techniques, they show in Theorem 4.1 that the solutions for (9) depend continuously on the data (see [33] for a complete discussion). Moreover, any sequence of minimizers of the finite dimensional problems for (7) converge in the Prohorov metric to a minimizer for the original infinite dimensional problem (9). This theorem makes use of the result they prove in Theorem 3.1 [33] that the set $\mathcal{P}_0(Q)$ as in (6) is dense the space $\mathcal{P}(Q)$ with respect to the Prohorov metric ρ.

In demonstrating the convergence of solutions for the family of finite dimensional problems (7), the result established in Theorem 4.1 of [33] also provides a computational framework for solving the general parameter estimation problem (9) without specifying prior probability distributions. Using discrete Dirac delta measures, for sufficiently large M we may approximate

$$\int_Q \mathcal{O}y(t; q) dP(q) \approx \int_Q \sum_{j=0}^{M} \mathcal{O}y(t; q) p_j \delta_{q_j^M}(q) dq = \sum_{j=0}^{M} \mathcal{O}y(t; q_j^M) p_j,$$

which then allows us to approximate the infinite dimensional inverse problem (9) by the finite dimensional approximation (7).

4 Application to a toxicokinetic model for TCE

In this section we apply the methods developed in Section 2 to a toxicokinetic model for trichloroethylene. We utilize the probability-based estimation

technique involving (3) and we compare the results to those obtained with a traditional deterministic method for (2).

Here we develop and test several estimation problems for the TCE model with simulated data that qualitatively and quantitatively match the experimental data in [38], and we demonstrate which strategies and observations are most successful at capturing and predicting the dynamics of TCE in adipose tissue. This in-depth study of the parameter estimation process with simulated data is important for *testing and understanding the capability of these estimation techniques*, and is a necessary first step before use of the methods with experimental data containing additional and generally unknown sources of variability.

The results that we present here demonstrate clear advantages for the probability-based method when the experimental observations are subject to inter-individual variability. In addition, our studies of the parameter estimation problem for the TCE model illustrate ways in which both the quantity and the quality of experimental data have a major impact on the effectiveness of parameter estimation techniques.

4.1 Overview of the TCE model

Here we provide an overview of the PBPK-hybrid model for TCE as developed in [16,36]. This model utilizes standard physiologically based pharmacokinetic compartmental equations for various non-fat tissues. The fat tissue compartment is described with a spatially varying dispersion model, and is designed specifically to capture the *intra-individual* variability that results from the heterogeneous physiology of fat.

The most commonly used compartmental model in PBPK modeling is the perfusion-limited, or flow-limited compartment. This model is based on simple mass balance principles and assumptions of rapid equilibrium and spatial uniformity. Moreover, it is assumed that the blood flow rate to the tissue is much slower than the rate of transport of the compound across cell membranes. The resulting equation for the tissue concentration C of the compound is given by

$$V\frac{dC(t)}{dt} = Q_{,}bl(C_{in}(t) - C_{out}(t)),$$

where V is the volume of the tissue, Q_{bl} is the volumetric blood flow rate to the tissue, and C_{in} and C_{out} are the concentrations of compound entering and leaving the tissue respectively (see [39]). Under standard assumptions, the concentration C_{out} is equal to the concentration C of compound in the tissue divided by the blood:tissue partition coefficient [39].

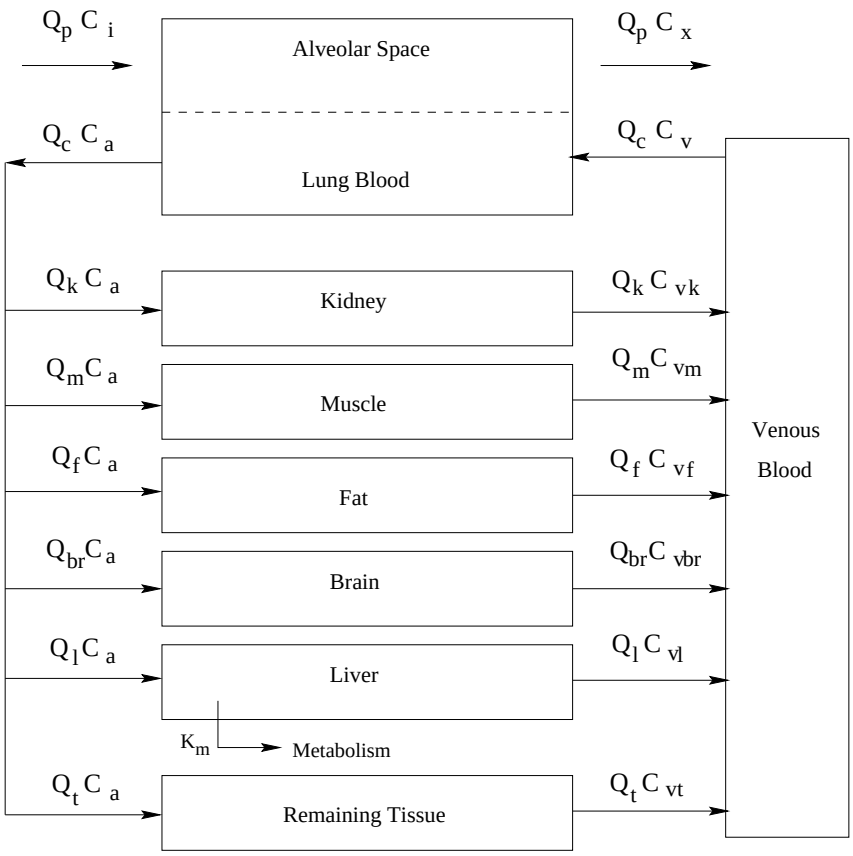

Fig. 1. Schematic of PBPK model for inhaled TCE in Long-Evans rats.

For many tissues and compounds of interest the perfusion-limited compartmental model is adequate to describe the dynamics of such compounds inside the tissues. In the case of highly lipophilic substances such as TCE, however, the standard models may not accurately capture the transport of these chemicals in the adipose tissue. As discussed in Section 1, the highly heterogeneous physiology of fat tissue appears to have a major influence on the behavior of TCE in fat. Using a PBPK model for TCE in Long-Evans rats with a perfusion-limited fat compartment [38] (see Figure 1 for a model schematic), model simulations suggested that the standard model indeed does not capture the concentration profile of TCE in adipose tissue as seen in experimental data [16].

To account for the spatial variation in TCE fat concentrations as suggested by the physiology of adipose tissue, an axial dispersion model was developed to replace the perfusion-limited fat tissue compartment. This model is based

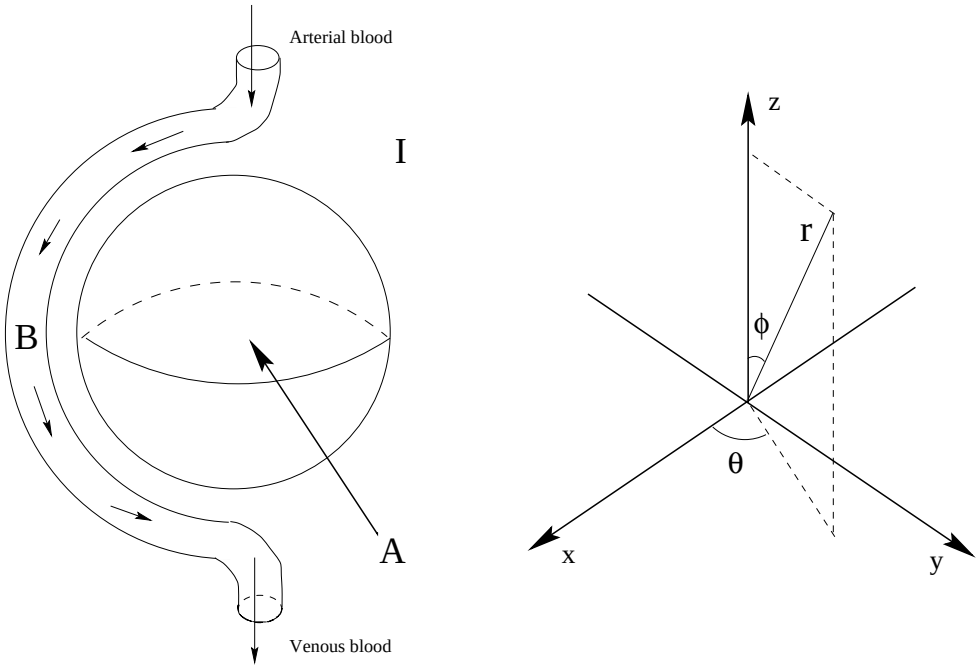

Fig. 2. Geometric representation of an adipocyte-capillary unit in adipose tissue. The adipocyte region (A) is represented by a sphere and is immersed in the interstitial fluid (I). The capillary or blood region (B) is a cylindrical tube that wraps around the adipocyte. Coordinates are in spherical coordinates (r, θ, ϕ).

directly on the structure of fat tissue, which consists primarily of spherical, lipid-containing cells called adipocytes. Each adipocyte is in contact with one or more capillaries [40] and is immersed in interstitial fluid. Figure 2 depicts the geometric representation of an adipocyte-capillary unit as used in the dispersion model. There are three subcompartments in the model that represent the adipocyte region (A), the capillary or blood region (B) and the interstitial fluid (I).

The model equations are based on an axial dispersion model developed by Roberts and Rowland [41] for the liver. A key feature of their model is its *aggregate structure*, using a single cellular unit with the dispersion term to represent the intra-individual variability that occurs across the millions of cells in the tissue. As detailed in [16], we have adapted their model to describe the geometry of adipose tissue and the transport of TCE within the fat. The resulting system of partial differential equations is given by

$$\begin{aligned} V_B \frac{\partial C_B}{\partial t} = & \frac{V_B}{r_2 \sin\phi} \frac{\partial}{\partial \phi} \left[\sin\phi \left(\frac{\mathcal{D}_B}{r_2} \frac{\partial C_B}{\partial \phi} - v C_B \right) \right] \\ & + \lambda_I \mu_{BI} (f_I C_I(\theta_0) - f_B C_B) \\ & + \lambda_A \mu_{BA} (f_A C_A(\theta_0) - f_B C_B) \end{aligned} \tag{10}$$

$$-\frac{\mathcal{D}_B}{r_2}\frac{\partial C_B}{\partial \phi}(t,\phi) + vC_B(t,\phi)\bigg|_{\phi=\varepsilon_1} = \frac{Q_c}{1000A_B}C_a(t) \tag{11}$$

$$-\frac{\mathcal{D}_B}{r_2}\frac{\partial C_B}{\partial \phi}(t,\phi) + vC_B(t,\phi)\bigg|_{\phi=\pi-\varepsilon_2} = \frac{Q_c}{1000A_B}C_v(t) \tag{12}$$

$$\begin{aligned} V_I\frac{\partial C_I}{\partial t} = {} & \frac{V_I D_I}{r_1^2}\left[\frac{1}{\sin^2\phi}\frac{\partial^2 C_I}{\partial \theta^2} + \frac{1}{\sin\phi}\frac{\partial}{\partial\phi}\left(\sin\phi\frac{\partial C_I}{\partial\phi}\right)\right] \\ & + \delta_{\theta_0}(\theta)\chi_B(\phi)\lambda_I\mu_{BI}(f_B C_B - f_I C_I) \\ & + \mu_{IA}(f_A C_A - f_I C_I) \end{aligned} \tag{13}$$

$$C_I(t,\theta,\phi) = C_I(t,\theta+2\pi,\phi) \tag{14}$$

$$\frac{\partial C_I}{\partial\theta}(t,\theta,\phi) = \frac{\partial C_I}{\partial\theta}(t,\theta+2\pi,\phi) \tag{15}$$

$$C_I(t,\theta,0) < \infty \tag{16}$$

$$C_I(t,\theta,\pi) < \infty \tag{17}$$

$$\begin{aligned} V_A\frac{\partial C_A}{\partial t} = {} & \frac{V_A D_A}{r_0^2}\left[\frac{1}{\sin^2\phi}\frac{\partial^2 C_A}{\partial \theta^2} + \frac{1}{\sin\phi}\frac{\partial}{\partial\phi}\left(\sin\phi\frac{\partial C_A}{\partial\phi}\right)\right] \\ & + \delta_{\theta_0}(\theta)\chi_B(\phi)\lambda_A\mu_{BA}(f_B C_B - f_A C_A) \\ & + \mu_{IA}(f_I C_I - f_A C_A) \end{aligned} \tag{18}$$

$$C_A(t,\theta,\phi) = C_A(t,\theta+2\pi,\phi) \tag{19}$$

$$\frac{\partial C_A}{\partial\theta}(t,\theta,\phi) = \frac{\partial C_A}{\partial\theta}(t,\theta+2\pi,\phi) \tag{20}$$

$$C_A(t,\theta,0) < \infty \tag{21}$$

$$C_A(t,\theta,\pi) < \infty. \tag{22}$$

The capillary equation (10) describes the transport of TCE in the capillary region of the adipose tissue and utilizes the dispersion term

$$\frac{V_B}{r_2\sin\phi}\frac{\partial}{\partial\phi}\left[\sin\phi\frac{\mathcal{D}_B}{r_2}\frac{\partial C_B}{\partial\phi}\right]$$

with dispersion coefficient $\mathcal{D}_B$. This term accounts for the variability in physiological properties that occurs across the population of fat cells, with a large dispersion coefficient indicating a high degree of variability. Mathematically, the dispersion term is equivalent to a standard diffusion term, although the dispersion term is used specifically to approximate the observed physiological phenomena of varying path lengths, flow velocities and compound transit times that occur within a tissue.

The boundary conditions (11) and (12) connect the adipose capillary region

to the systemic arterial and venous blood compartments using flux balance. Transport of TCE between the capillary region and the other two adipose subcompartments (interstitial and adipocyte) is modeled in the PDE (10). The variables $C_B(t)$, $C_I(t)$ and $C_A(t)$ denote concentrations of TCE in the capillary, interstitial and adipocyte regions respectively, while $C_a(t)$ and $C_v(t)$ represent the systemic arterial and venous blood concentrations of TCE.

The interstitial region is modeled with the two-dimensional PDE (13) and boundary conditions (14) – (17). The adipocyte region equations (18) – (22) are similar in structure to the interstitial equations, and describe the diffusion of TCE around the surface of the adipocyte as well as the transport of TCE between the three adipose subcompartments. The boundary conditions are standard periodic and finiteness boundary conditions that are commonly used for the diffusion equation on a spherical domain. A detailed derivation and description of the dispersion model is given in [16,36].

Adipose model parameters include the dispersion coefficient $\mathcal{D}_B$ (m^2/hour); diffusion coefficients D_I and D_A (m^2/hour); the fractions f_B, f_I, f_A of unbound TCE in each adipose region; cell membrane permeability coefficients μ_{BA}, μ_{IA}, μ_{BI} (liters/hour); blood flow parameters v (m/hour) and $\mathcal{F}$; and inter-region transport parameters λ_I and λ_A.

The adipose model equations (10) – (22) are coupled with standard compartmental equations for the lung, arterial blood, venous blood, liver, brain, kidney, muscle and remaining non-fat tissue to obtain a whole-body PBPK-hybrid model. Uptake of TCE is via inhalation into the lungs, and metabolism is modeled with a Michaelis-Menten term in the liver. The resulting equations are given by

$$\begin{aligned} V_v\frac{dC_v}{dt} &= Q_mC_m/P_m + Q_tC_t/P_t + Q_fC_B(\cdot,\pi-\varepsilon_2) + Q_{br}C_{br}/P_{br} \\ &\quad + Q_lC_l/P_l + Q_kC_k/P_k - Q_cC_v \qquad (23) \\ C_a &= \frac{Q_cC_v + Q_pC_c}{Q_c + \frac{Q_p}{P_b}} \qquad (24) \\ V_m\frac{dC_m}{dt} &= Q_m(C_a - C_m/P_m) \qquad (25) \\ V_t\frac{dC_t}{dt} &= Q_t(C_a - C_t/P_t) \qquad (26) \\ V_{br}\frac{dC_{br}}{dt} &= Q_{br}(C_a - C_{br}/P_{br}) \qquad (27) \\ V_l\frac{dC_l}{dt} &= Q_l(C_a - C_l/P_l) - \frac{v_{max}C_l/P_l}{k_M + C_l/P_l} \qquad (28) \\ V_k\frac{dC_k}{dt} &= Q_k(C_a - C_k/P_k), \qquad (29) \end{aligned}$$

where $C_v(t)$, $C_{br}(t)$, $C_k(t)$, $C_m(t)$, $C_l(t)$ and $C_t(t)$ denote TCE concentrations in the venous blood, brain, kidney, muscle, liver and remaining tissue compartments, respectively. The chamber air concentration $C_c(t)$ is specified as part of the experiment and is used as a forcing function in the arterial blood equation (24). For the results we present in this paper, we set the chamber air concentration to 2000 parts per million TCE for one hour, followed by zero ppm TCE until the final time t_f (in hours).

Model parameters include tissue volumes V (in liters), volumetric blood flow rates to the tissues Q (liters/hour), and blood:tissue partition coefficients P, each labeled with a subscript corresponding to the appropriate tissue. The cardiac output and ventilation rates (in liters/hour) are denoted by Q_c and Q_b respectively, and the blood:air partition coefficient is denoted as P_b. The standard Michaelis-Menten metabolic parameters are denoted by v_{max} (mg/hour) and k_M (mg/liter). See [16,36] for complete discussion of the model equations and parameters.

Theoretical results relating to well-posedness of the whole-body PBPK-hybrid model are presented in [42]. In particular, we have shown the existence of a unique weak solution for a general class of nonlinear parabolic equations that includes the TCE model as a special case. Moreover, we established the well-posedness of the deterministic estimation problem for the TCE model, and in [36] we have addressed the well-posedness of probability-based parameter estimation methods applied to the TCE model. Numerical methods and simulations for this model with deterministic parameters are given in [17], and results for the standard PBPK models are compared to those for the PBPK-hybrid model.

4.2 Deterministic parameter estimation methods for the TCE model

In this section we present results for the standard deterministic parameter estimation problem

$$\min_{q \in Q} J(q, z) = \sum_{i=1}^{N_t} |\mathcal{O}y(t_i; q) - z_i|^2 \tag{30}$$

applied to the TCE model, where

$$y(t) = [C_B(t),\ C_I(t),\ C_A(t),\ C_v(t),\ C_{br}(t),\ C_k(t),\ C_m(t),\ C_t(t),\ C_l(t)]^{\mathrm{T}},$$

q denotes the vector of unknown parameters in the admissible parameter space Q, the observations are denoted by z_i, $i = 1, \ldots, N_t$, and $\mathcal{O}$ is the observation

operator. Here the variable $y(t)$ is subject to

$$\frac{\partial}{\partial t} y(t, \cdot) = f(t, y(t, \cdot); q),$$

which we use as abbreviated notation for the TCE PBPK-hybrid model (10) – (29).

4.2.1 Simulated observations

The observations z_i, $i = 1, \ldots, N_t$ and the observation operator $\mathcal{O}$ are defined so that they correspond to the types of experimental observations used in the experiments conducted by Evans et al. [38]. The data that they collected include measurements of TCE concentrations in homogenized samples of fat tissue. Therefore we define the observation operator

$$\mathcal{O}y(t_i; q) = \bar{C}_A(t_i; q) \tag{31}$$

where $\bar{C}_A(t_i; q)$ is the mean concentration of TCE over the adipocyte region (see [17] for a precise definition and finite-dimensional approximation). Similarly, we define the simulated observations

$$z_i = \bar{C}_A(t_i; q^*) \tag{32}$$

for $i = 1, \ldots, N_t$ and for some $q^* \in Q$.

Note that the observations in (32) are defined to simulate measurements from a single individual. The experimental data [38], however, include measurements of TCE concentrations in several different rats. To approximate this inter-individual variability within observations, we generate two additional types of simulated data. The first type of data (Type I) represents the case where measurements are collected from several individuals over time, so that each individual is measured at each time point. Note that this type of inter-individual data includes trajectories over time for each individual in the group.

The second type of data (Type II) represents the case where measurements are collected from multiple individuals so that each individual is utilized only once. That is, at each time point there is a separate group of individuals that is measured. This is the type of data that emerges from experimental measurements of tissue concentrations when the animal must be sacrificed and the entire tissue is removed for analysis. The experimental data collected by Evans et al. [38] are of this type.

For each of these types of data, we generate simulated observations by assuming that the model parameters vary across the population so that the param-

eter vector q has a probability distribution function $P \in \mathcal{P}(Q)$. We assume there are a total of N_s individuals sampled at each time point $i = 1, \ldots, N_t$. Thus for Type I data there is a total of N_s individuals, while for Type II data there is a total of $N_s N_t$ individuals.

For Type I data we define the observations

$$\hat{z}_i^I = \frac{1}{N_s} \sum_{j=1}^{N_s} \bar{C}_A(t_i; q_j^*) \tag{33}$$

for $i = 1, \ldots, N_t$, where q_j^* is sampled from a given $P^* \in \mathcal{P}(Q)$ for $j = 1, \ldots, N_s$. These observations represent the expected value of measurements taken over the N_s individuals in the group, and are realizations of the observations used in (3) and described in (4).

The Type II observations are given by

$$\hat{z}_i^{II} = \frac{1}{N_s} \sum_{j=1}^{N_s} \bar{C}_A(t_i; q_{ij}^*) \tag{34}$$

for $i = 1, \ldots, N_t$, where q_{ij}^* is sampled from P^* for $j = 1, \ldots, N_s$ and $i = 1, \ldots, N_t$.

4.2.2 Parameter estimation results

We have conducted a thorough study in [36] of the deterministic parameter estimation problem (30) for the adipose model parameters

$$q = [\mathcal{D}_B,\ D_I,\ D_A,\ \mu_{IA},\ \mu_{BA},\ \mu_{BI},\ r_1,\ A_B]. \tag{35}$$

We considered estimation problems for single parameters as well as pairs, triples, and the entire 8-vector q. The data used in the estimation problems included simulated data approximating a single individual (32) and Type II data (34). The objective function given in (30) was minimized in Matlab with `fminsearch`, which uses a Nelder-Mead direct search algorithm. A complete description of the estimation problems and results is presented in [36]. As the probabilistic estimation methods are the focus of this paper, here we merely summarize the results from the deterministic estimation method for comparative purposes.

Using observations (32) simulating a single individual, we obtained optimized parameters that were significantly close to the data-generating parameters q^*, with more accurate estimates for pairs and triples of parameters than for

the entire eight-dimensional vector q. We studied the effect of varying the time points $t_i, i = 1, \ldots, N_t$ at which the observations are "collected" and the number of samples N_s taken at each time point, determining that the accuracy of our parameter estimates increased as the number of time points N_t increased, as the range of time included in the observations increased, and as the number of samples N_s at each time point increased.

When we introduced inter-individual variability into the observations by using Type II data (34), the deterministic estimation method yielded parameters that were less accurate as the degree of population variability increased. These results, which are presented in full detail in [36], suggested that the deterministic parameter estimation technique may not be the best method to use when the observations are subject to significant inter-individual variability.

To further illustrate this point, here we present results for the deterministic parameter estimation problem with observations that simulate parameters with a bimodal probability distribution. In this case we utilize the estimation problem (30) with $q = \mathcal{D}_B$, the dispersion coefficient. This parameter is a measure of the degree of variability that occurs within an individual's fat tissue, and it is plausible to assume that $\mathcal{D}_B$ may be bimodally distributed over a population with male and female subpopulations.

The observations used in these estimation problems were generated with a bimodal distribution $P^* = P_{bi}$ composed of two normal distributions with means $\mu_1 = 1$, $\mu_2 = 3$, standard deviations $\sigma_1 = 0.1667$, $\sigma_2 = 0.2$ and mixing parameter 0.5 (i.e., equal weighting between the two gaussians). See Figure 3 for a graph of the probability density functions that are combined to create the bimodal density function. This probability density is used to generate observations of Type I (33) and Type II (34); single individual observations (32) are generated using $q^* = 1$. We used three different time vectors $t_i, i = 1, \ldots, N_t$ for our simulated observations to study the effect of the quantity of data on the quality of the estimation results. The vectors we used included the following:

$$\begin{aligned}
\vec{t_1} &= [0,\ 5,\ 20,\ 40,\ 60,\ 120] \\
\vec{t_2} &= [0,\ 5,\ 20,\ 40,\ 60,\ 120,\ 180,\ 240,\ 300] \\
\vec{t_3} &= [0,\ 5,\ 20,\ 40,\ 60,\ 90,\ 105,\ 110,\ 115,\ 120,\ 125,\ 130, \\
&\qquad 135,\ 150,\ 180,\ 210,\ 240,\ 270,\ 275,\ 285,\ 290,\ 295,\ 300]
\end{aligned}$$

(in minutes) with $t_f = 2$ hours, $N_t = 6$ for $\vec{t_1}$, $t_f = 5$ hours, $N_t = 9$ for $\vec{t_2}$, and $t_f = 5$ hours, $N_t = 23$ for $\vec{t_3}$. Simulated observations (32) – (34) were then generated using these time vectors. The initial iterate q_0 used in the optimizer was sampled from a uniform distribution on $[0.75, 1.25]$ and has a value of $q_0 = 0.9577$.

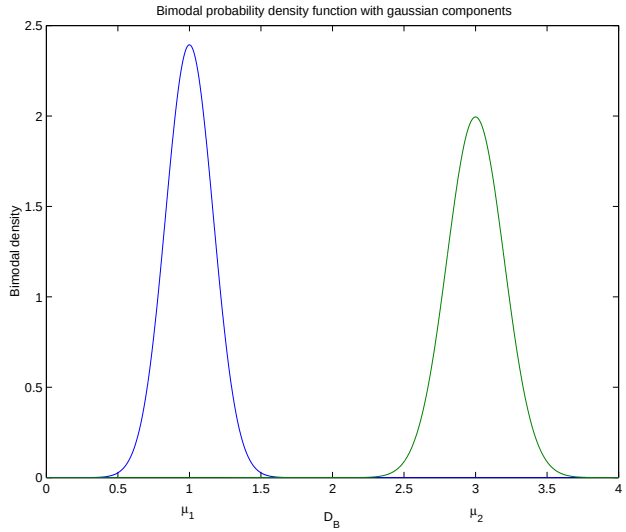

Fig. 3. Normal probability density functions with means $\mu_1 = 1$, $\mu_2 = 3$ and standard deviations $\sigma_1 = 0.1667$, $\sigma_2 = 0.2$.

	$\vec{t}_1$	$\vec{t}_2$	$\vec{t}_3$
q_{opt}	1.0000	1.0000	1.0000
$J(q_0, z)$	0.0966	0.5689	2.2229
$J(q_{opt}, z)$	$2.5014e{-9}$	$5.7127e{-9}$	$1.4811e{-8}$

Table 1
Optimized solutions q_{opt} for the deterministic estimation problem (30) with initial iterate $q_0 = 0.9577$ and data-generating parameter $q^* = 1$. The observations z were generated to simulate a single individual as in (32) at the time points $\vec{t}_1$, $\vec{t}_2$ and $\vec{t}_3$.

For observations simulating a single individual, the optimized solutions were equal to $q_{opt} = 1.0000$ for each choice of time vectors $\vec{t}_1$, $\vec{t}_2$, $\vec{t}_3$. See Table 1 for a summary of these parameter estimation results.

We also carried out the deterministic estimation problem using Type I data as in (33) with the time vectors $\vec{t}_1$, $\vec{t}_2$ and $\vec{t}_3$, where q_j^* is sampled from the bimodal distribution function $P^* = P_{bi}$ for $j = 1, \ldots, N_s$. The values of the number N_s of samples taken at each time point we considered included $N_s = 5, 10, 20$ and 50. Results are presented in Table 2, and plots of the resulting concentration profiles are given in Figures 4 – 7. In this and all following figures, the small solid dots are the individual data points and the larger open circles are the averaged values of the data at each point in time. Note that the data used in the optimization problem are the larger open circles.

As seen in the figures, as the value of N_s increases, the predicted TCE adipocyte concentrations for q_{opt} more closely approximate the response from the data-

	$\vec{t}_1$				$\vec{t}_2$				$\vec{t}_3$			
N_s	5	10	20	50	5	10	20	50	5	10	20	50
q_{opt}	2.82	2.38	1.79	1.60	2.82	2.32	1.71	1.53	2.82	2.33	1.72	1.54
$J(q_0, z)$	34.5	27.0	15.1	10.9	145	114	62.9	45.5	593	464	257	186
$J(q_{opt}, z)$	$8e{-}6$	$4e{-}3$	0.02	0.02	$2e{-}4$	0.02	0.10	0.12	$8e{-}4$	0.08	0.45	0.53

Table 2
Optimized solutions q_{opt} for the deterministic estimation problem (30) with initial iterate $q_0 = 0.9577$ and data-generating parameter $q^* = 1$. The observations z were generated to simulate Type I data as in (33) at the time points $\vec{t}_1$, $\vec{t}_2$ and $\vec{t}_3$ with $N_s = 5, 10, 20$ and 50.

generating parameters q^*. For the case $N_s = 5$, note that the individual data points all are clustered near the response for the parameter value $\mathcal{D}_B = 3$, so that observations for this particular estimation problem do not behave as bimodally distributed data. This is reflected in the optimized parameter $q_{opt} = 2.82$, which is significantly different from the results for other data sets which contain points from both gaussians. Increasing the number of time points at which the observations are collected does not improve the results for this type of data (e.g., see Figure 7). These results indicate the importance of ensuring a large enough sample size for the observations.

Similar results for Type II data are given in Table 3 and Figures 10 – 13. Note that as N_s increases and as the number N_t of time points increases, the deterministic method yields optimized parameters that generate reasonable approximations for the dynamics of TCE as seen with the data-generating parameters. It is clear from Tables 2 and 3, however, that the optimized parameters q_{opt} do not provide any information about the population variance or the distribution of parameter values. Indeed, most of the values of q_{opt} lie in the range from 1.3 to 1.8, which is in between the means of the two gaussians that form the bimodal distribution $P^* = P_{bi}$. Moreover, as constants, the optimized parameters cannot suggest the underlying bimodal nature of the observations used in the optimization problem. Therefore it appears that the deterministic approach may not be most appropriate except in situations when the form of the parameter distribution function is already known and it is necessary only to estimate the mean.

4.3 Probabilistic parameter estimation methods for the TCE model

In this section we apply the probability-based parameter estimation methods presented in Section 2 to the TCE PBPK-hybrid model. These results are then

	$\vec{t}_1$				$\vec{t}_2$				$\vec{t}_3$			
N_s	5	10	20	50	5	10	20	50	5	10	20	50
q_{opt}	1.34	1.44	1.33	1.28	1.51	1.49	1.52	1.46	1.35	1.42	1.50	1.50
$J(q_0, z)$	5.50	7.26	5.23	4.11	46.5	42.7	46.6	39.7	128	148	174	174
$J(q_{opt}, z)$	0.52	0.12	0.34	0.14	2.49	0.88	1.82	0.98	21.6	13.2	6.08	3.71

Table 3
Optimized solutions q_{opt} for the deterministic estimation problem (30) with initial iterate $q_0 = 0.9577$ and data-generating parameter $q^* = 1$. The observations z were generated to simulate Type II data as in (34) at the time points $\vec{t}_1$, $\vec{t}_2$ and $\vec{t}_3$ with $N_s = 5, 10, 20$ and 50.

compared to the results for the deterministic estimation problem as given in Section 4.2. Here we utilize the estimation problem with objective function

$$J(P, \hat{z}) = \sum_{i=1}^{N_t} |\mathcal{E}[\mathcal{O}y(t_i; q)|P] - \hat{z}_i|^2$$

and its finite-dimensional approximation

$$J(P_M, \hat{z}) = \sum_{i=1}^{N_t} \left| \sum_{j=0}^{M} \mathcal{O}y(t_i; q_j) p_j - \hat{z}_i \right|^2 \tag{36}$$

as discussed in Section 2, with observations $\hat{z}_i, i = 1, \ldots, N_t$ corresponding to the expected value (4).

As in Section 4.2 with the deterministic problem, we estimate the dispersion coefficient $\mathcal{D}_B$ using observations from a single individual and from bimodally distributed data of Types I and II. The admissible parameter space Q is defined as the closed interval $[0, 4]$, and we use the finite dimensional approximation spaces $Q_M \subset Q$ for $M = 32, 64, 128$ with

$$q_j = 4j/M, \quad j = 0, \ldots, M.$$

The constrained quadratic programming problem (36) was solved in Matlab using `quadprog` to obtain the probabilities $[p_0, \ldots, p_M]$ subject to the constraints $p_j \geq 0$ and $\sum_{j=0}^{M} p_j = 1$. That is, we minimized

$$p^{\mathrm{T}} A p + 2b^{\mathrm{T}} p + c$$

subject to $p_j \geq 0$, $j = 0, \ldots, M$ and $\sum_{j=0}^{M} p_j = 1$, where

$$A_{jk} = \sum_{i=1}^{N_t} \mathcal{O}y(t_i; q_j)\mathcal{O}y(t_i; q_k), \quad j,k = 0,\ldots,M$$
$$b_j = -\sum_{i=1}^{N_t} \mathcal{O}y(t_i; q_j)\hat{z}_i, \quad j = 0,\ldots,M$$
$$c = \sum_{i=1}^{N_t} \hat{z}_i^2$$

with observations $\hat{z}_i$ from (32), (33) or (34), respectively, and time vectors $\vec{t}_1$, $\vec{t}_2$ or $\vec{t}_3$. As in Section 4.2.1, the probability distribution P^* we used to generate the observations is the bimodal distribution P_{bi} with means $\mu_1 = 1$ and $\mu_2 = 3$, standard deviations $\sigma_1 = 0.1667$ and $\sigma_2 = 0.3333$ and mixing parameter 0.5.

For the estimation problem with observations (32) simulating a single individual, the solution vector p_{opt} obtained by the optimization routine is given by

$$p_{opt}^j = \begin{cases} 1 & \text{if} \quad q_j = 1 \\ 0 & \text{otherwise} \end{cases}$$

for $j = 0,\ldots,M$ and $M = 32, 64, 128$. This solution corresponds exactly to the data-generating parameter distribution with a single mass at $\mathcal{D}_B = 1$, suggesting that the probability-based estimation method is equally accurate in solving for parameters from a single individual as the deterministic estimation method. That is, if the data correspond to a deterministic parameter system, using the more general probability-based formulation will still provide correct results by returning a Dirac measure.

Results for the probabilistic method using Type I data are also plotted in Figures 4 – 9. Each of these figures illustrates the case with $M = 32$; the results for $M = 64$ and $M = 128$ were qualitatively similar and are not presented here. Simulated TCE adipocyte concentrations corresponding to the optimized parameters $q_{prob} = P_{opt}$ are depicted in Figures 4 – 7 in comparison with the observations and the model response corresponding to the optimized parameter q_{det} from the deterministic problem for the same observation set. Graphs of the optimized probability distributions p_{opt} are given in Figures 8 and 9 for $\vec{t}_1$ and increasing values of N_s.

Note in Figures 4, 5 and 6 that as the sample size N_s increases from 5 to 50, the predicted adipocyte concentrations for both q_{det} and q_{prob} more closely match the response corresponding to the data-generating parameter set. Similarly, the optimized probability distributions appear to converge to the data-generating distribution $P^* = P_{bi}$ as N_s increases (see Figures 8 and 9). This apparent convergence is in agreement with the established theoretical convergence of the finite-dimensional parameter estimation solutions summarized in

Section 3 and detailed in [33].

As discussed in Section 4.2, however, Figure 7 suggests that the accuracy of the predicted model responses does not appear to improve for this type of data as the number N_t of time points increases without a concurrent increase in N_s. In this case, the additional time-course information contained in the extra time points does not contribute to the richness of data since the same group of individuals is sampled at every time point.

Unlike the optimized solutions from the deterministic problem, the solutions from the probabilistic estimation method contain information about the distribution of the parameters across the population. In addition to providing an estimate of the mean and variance, the probabilistic method also yields an approximation of the probability distribution itself. It is important, however, to ensure that a large enough sample size is used for the observations so that the population is accurately represented in the data.

We present results for the probabilistic estimation problem with Type II data in Figures 10 – 15. Predicted TCE adipocyte concentrations are given in Figures 10 – 13 for $\vec{t}_1$ and $\vec{t}_3$ and various values of N_s; probability distribution functions for $\vec{t}_3$ and varying N_s are presented in Figures 14 and 15.

As illustrated in Figures 10 and 11 for $\vec{t}_1$ with $N_s = 5$ and 20 respectively, the optimized solutions for the probabilistic method can yield highly inaccurate predictions of TCE adipocyte concentrations while the deterministic method produces an accurate match to the observations. Our studies with the quadratic programming problem have indicated the existence of multiple solutions that satisfy the constraints, with some of the solutions yielding inaccurate predictions of adipocyte concentrations beyond the time period covered in the data.

This difficulty does not arise for observations that utilize the larger time vector $\vec{t}_3$ (see Figures 12 and 13), where the predicted responses for both methods are reasonably close to the observations for all four values of N_s. As seen in Figures 14 and 15, however, the optimized probability distributions P_{opt} do not appear to converge with increasing values of N_s as clearly as in the case with Type I data.

It is important to note that the quality of the Type II data is significantly different than the quality of Type I data since there is no time-course information from any individual in the Type II data. Since each individual is measured only once for the Type II data, the data points do not contain any trajectories in time from an individual. In contrast, *every* individual is measured at *each* time point for the Type I data, incorporating important time-course dynamics into the observations.

As demonstrated in our results, the deterministic method can reasonably predict the *expected value* of the model response over the population for Type I and Type II data, but can provide no information about the population distribution of the parameters. The probabilistic method, however, can successfully yield the expected value, variance and overall shape of the population distribution using Type I data. The probabilistic method has some difficulty capturing this probability distribution with Type II data since valuable time-course information is not contained in these data. Therefore it appears that Type II data is, not surprisingly, less desirable for situations when the probability distribution of parameters must be estimated with no prior knowledge of the shape of the distribution.

5 Concluding remarks

In this paper we have presented probability-based parameter estimation methods which incorporate and account for uncertainty and population variability that arise in biological models. We outlined known theoretical results that address issues of well-posedness for these estimation problems, and we applied them to a toxicokinetic model for trichloroethylene using simulated observations that exhibited differing types of variability. These results were compared to results obtained with a traditional deterministic parameter estimation method.

As one might expect, our results indicate that the performance of the deterministic and probabilistic estimation methods depends greatly on both the quantity and quality of the data used in the estimation process. For data that represent a single individual, each method produced parameter estimates that accurately matched the data-generating parameter q^*. Not surprisingly, in this case it appears that the deterministic method is sufficient for estimating parameters using data from a single individual.

When any degree of variability is introduced into the experimental data, however, the deterministic estimation method is unable to capture any of this variability in its parameter estimates. This method can in some cases reasonably predict the *expected value* of the parameter over the population, but can provide no information about the variance or the shape of the probability distribution.

For observations of the expected value that contain important time-course trajectories (as in the case of our Type I data), the probabilistic method is able to successfully predict *both* the expected value and the overall probability distribution of the parameter. This is accomplished by solving a standard quadratic programming problem with no prior knowledge of the population

distribution. As seen in our results, it is important to use a sufficiently large sample size so that the observations contain an adequate representation of the population.

The results for the probability-based estimation problem are less conclusive for Type II data, which are significantly different from the Type I data in that they contain only one observation from each individual. This results in data that include no individual time-course trajectories, which contributes to difficulties in the optimization process. There appear to be many solutions to the constrained quadratic programming problem for this type of data, and some of these solutions produce inaccurate predictions of TCE adipocyte concentrations for time periods not included in the data themselves. Increasing the number of time points N_t at which the observations are collected seems to improve the results, but the convergence of the probability distributions is less clear than for the case with Type I data. Our results therefore suggest that it is less than desirable to utilize Type II data for estimating parameters that have an unknown probability distribution. If at all possible, it is best to collect experimental data that contain multiple measurements from individuals over time, as this time-course information adds significant richness to the experimental data and significantly enhances our ability to estimate underlying inter-individual variability.

A key feature of the probability-based methods presented here is their ability to estimate population distributions for parameters without the use of priors. In situations where there is little information about the shape of the probability distribution(s), this can be a clear advantage over methods (e.g., Markov Chain-Monte Carlo) that require the specification of priors. For example, a parameter with a bimodal parameter distribution may be mistakenly estimated as a unimodal distribution if too much weight is placed on the assumption of a unimodal prior. The probability-based estimation methods avoid this problem, and the utilization of Dirac delta measures in the proper metric space guarantees the theoretical convergence of the resulting estimates of the probability distributions.

Current and future efforts related to this work include a study and comparison of other probability-based parameter estimation techniques. In particular, we plan to implement Markov Chain-Monte Carlo methods applied to a PBPK model and compare the results we obtain with the probability-based estimations outlined in this paper.

Acknowledgments

This research has been supported in part by the Air Force Office of Scientific Research under grants AFOSR F49620-98-1-0180 and AFOSR F49620-01-1-0026, as well as an NSF-GRT fellowship (grant GER-9454175) and a P.E.O. Scholar Award to L.K.P.

References

[1] United Nations Environment Programme, International Labour Organisation, World Health Organization, Trichloroethylene, World Health Organization, Geneva, 1985.

[2] K. Bergman, Application and results of whole-body autoradiography in distribution studies of organic solvents, CRC Critical Reviews in Toxicology 12 (1983) 59–118.

[3] J. V. Bruckner, B. D. Davis, J. N. Blancato, Metabolism, toxicity, and carcinogenicity of trichloroethylene, Critical Reviews in Toxicology 20 (1989) 31–50.

[4] G. M. Pastino, W. Y. Yap, M. Carroquino, Human variability and susceptibility to trichloroethylene, Environmental Health Perspectives 108 Suppl. 2 (2000) 201–215.

[5] R. J. Bull, Mode of action of liver tumor induction by trichloroethylene and its metabolites, trichloroacetate and dichloroacetate, Environmental Health Perspectives 108 Suppl. 2 (2000) 241–260.

[6] T. Green, Pulmonary toxicity and carcinogenicity of trichloroethylene: Species differences and modes of action, Environmental Health Perspectives 108 Suppl. 2 (2000) 261–264.

[7] H. Brauch, G. Weirich, M. A. Hornauer, S. Störkel, T. Wöhl, T. Brüning, Trichloroethylene exposure and specific somatic mutations in patients with renal cell carcinoma, Journal of the National Cancer Institute 91 (1999) 854–861.

[8] L. H. Lash, J. C. Parker, C. S. Scott, Modes of action of trichloroethylene for kidney tumorigenesis, Environmental Health Perspectives 108 Suppl. 2 (2000) 225–240.

[9] A. M. Saillenfait, I. Langonne, J. P. Sabate, Developmental toxicity of trichloroethylene, tetrachloroethylene and four of their metabolites in rat whole embryo culture, Archives of Toxicology 70 (1995) 71–82.

[10] H. A. Barton, H. J. Clewell III, Evaluating noncancer effects of trichloroethylene: Dosimetry, mode of action, and risk assessment, Environmental Health Perspectives 108 Suppl. 2 (2000) 323–334.

[11] R. Abbas, J. Fisher, A physiologically based pharmacokinetic model for trichloroethylene and its metabolites, chloral hydrate, trichloroacetate, dichloroacetate, trichloroethanol, and trichloroethanol glucuronide in B6C3F1 mice, Toxicology and Applied Pharmacology 147 (1997) 15–30.

[12] J. W. Fisher, M. L. Gargas, B. C. Allen, M. E. Andersen, Physiologically based pharmacokinetic modeling with trichloroethylene and its metabolite, trichloroacetic acid, in the rat and mouse, Toxicology and Applied Pharmacology 109 (1991) 183–195.

[13] J. W. Fisher, D. Mahle, R. Abbas, A human physiologically based pharmacokinetic model for trichloroethylene and its metabolites, trichloroacetic acid and free trichloroethanol, Toxicology and Applied Pharmacology 152 (1998) 339–59.

[14] M. S. Greenberg, G. A. Burton, J. W. Fisher, Physiologically based pharmacokinetic modeling of inhaled trichloroethylene and its oxidative metabolites in B6C3F1 mice, Toxicology and Applied Pharmacology 154 (1999) 264–78.

[15] R. D. Stenner, J. L. Merdink, J. W. Fisher, R. J. Bull, Physiologically-based pharmacokinetic models for trichloroethylene considering enterohepatic recirculation of major metabolites, Risk Analysis 18 (1998) 261–269.

[16] R. A. Albanese, H. T. Banks, M. V. Evans, L. K. Potter, Physiologically based pharmacokinetic models for the transport of trichloroethylene in adipose tissue, Bulletin of Mathematical Biology 64 (2002) 97–131.

[17] H. T. Banks, L. K. Potter, Model predictions and comparisons for three toxicokinetic models for the systemic transport of trichloroethylene, Mathematical and Computer Modeling 35 (2002) 1007–1032.

[18] D. Crandall, M. DiGirolamo, Hemodynamic and metabolic correlates in adipose tissue: Pathophysiologic considerations, FASEB 4 (1990) 141–147.

[19] D. Crandall, G. J. Hausman, J. G. Kral, A review of the microcirculation of adipose tissue: Anatomic, metabolic, and angiogenic perspectives, Microcirculation 4 (1997) 211–232.

[20] G. J. Hausman, The comparative anatomy of adipose tissue, in: New Perspectives in Adipose Tissue: Structure, Function and Development, Butterworths, London, 1985.

[21] S. Rosell, E. Belfrage, Blood circulation in adipose tissue, Physiological Reviews 59 (1979) 1078–1104.

[22] I. Nestorov, Modelling and simulation of variability and uncertainty in toxicokinetics and pharmacokinetics, Toxicology Letters 120 (2001) 411–420.

[23] P. Bernillon, F. Y. Bois, Statistical issues in toxicokinetic modeling: A Bayesian perspective, Environmental Health Perspectives 108 Suppl. 5 (2000) 883–893.

[24] F. Y. Bois, Applications of population approaches in toxicology, Toxicology Letters 120 (2001) 385–394.

[25] A. Gelman, F. Bois, J. Jiang, Physiological pharmacokinetic analysis using population modeling and informative prior distributions, Journal of the American Statistical Association 91 (1996) 1400–1412.

[26] F. Y. Bois, Statistical analysis of Fisher et al. PBPK model of trichloroethylene kinetics, Environmental Health Perspectives 108 Suppl. 2 (2000) 275–282.

[27] F. Y. Bois, Statistical analysis of Clewell et al. PBPK model of trichloroethylene kinetics, Environmental Health Perspectives 108 Suppl. 2 (2000) 307–316.

[28] F. Jonsson, F. Y. Bois, G. Johanson, Assessing the reliability of PBPK models using data from methyl chloride-exposed, non-conjugating human subjects, Archives of Toxicology 75 (2001) 189–199.

[29] F. Jonsson, F. Y. Bois, G. Johanson, Physiologically based pharmacokinetic modeling of inhalation exposure of humans to dichloromethane during moderate to heavy exercise, Toxicological Sciences (2001) 209–218.

[30] F. Jonsson, G. Johanson, Physiologically based modeling of the inhalation kinetics of styrene in humans using a Bayesian population approach, Toxicology and Applied Pharmacology 179 (2002) 35–49.

[31] F. Jonsson, G. Johanson, Bayesian estimation of variability in adipose tissue blood flow in man by physiologically based pharmacokinetic modeling of inhalation exposure to toluene, Toxicology 157 (2001) 177–193.

[32] H. T. Banks, Remarks on uncertainty of assessment and management in modeling and computation, Mathematical and Computer Modeling 33 (2001) 39–47.

[33] H. T. Banks, K. L. Bihari, Modeling and estimating uncertainty in parameter estimation, Inverse Problems 17 (2001) 1–17.

[34] H. T. Banks, Incorporation of uncertainty in inverse problems, Tech. Rep. CRSC-TR02-08, Center for Research in Scientific Computation, North Carolina State University, March 2002 (www.ncsu.edu/crsc/reports.html); Proceedings of the International Conference on Inverse Problems (Hong Kong, June 9-12, 2002), World Scientific, In press.

[35] H. T. Banks, B. G. Fitzpatrick, L. K. Potter, Y. Zhang, Estimation of probability distributions for individual parameters using aggregate population data, in: Stochastic Analysis, Control, Optimization and Applications: a Volume in Honor of W.H. Fleming, Birkhauser, Boston, 1999, pp. 353–371.

[36] L. K. Potter, Physiologically based pharmacokinetic models for the systemic transport of trichloroethylene, Ph.D. thesis, North Carolina State University, Raleigh, NC, www.lib.ncsu.edu (August 2001).

[37] P. Billingsley, Convergence of Probability Measures, Wiley, New York, 1968.

[38] M. V. Evans, W. K. Boyes, P. J. Bushnell, J. H. Raymer, J. E. Simmons, A physiologically based pharmacokinetic model for trichloroethylene (TCE) in Long-Evans rats, Preprint (1999).

[39] M. A. Medinsky, C. D. Klaassen, Toxicokinetics, in: Casarett and Doull's Toxicology: The Basic Science of Poisons, 5th Edition, McGraw-Hill, Health Professions Division, New York, 1996.

[40] B. G. Slavin, The morphology of adipose tissue, in: New Perspectives in Adipose Tissue: Structure, Function and Development, Butterworths, London, 1985.

[41] M. S. Roberts, M. Rowland, A dispersion model of hepatic elimination: 1. Formulation of the model and bolus considerations, Journal of Pharmacokinetics and Biopharmaceutics 14 (1986) 227–260.

[42] H. T. Banks, L. K. Potter, Well-posedness results for a class of toxicokinetic models, Tech. Rep. CRSC-TR01-18, Center for Research in Scientific Computation, North Carolina State University, July 2001 (www.ncsu.edu/crsc/reports.html); Discrete and Continuous Dynamical Systems, Submitted.

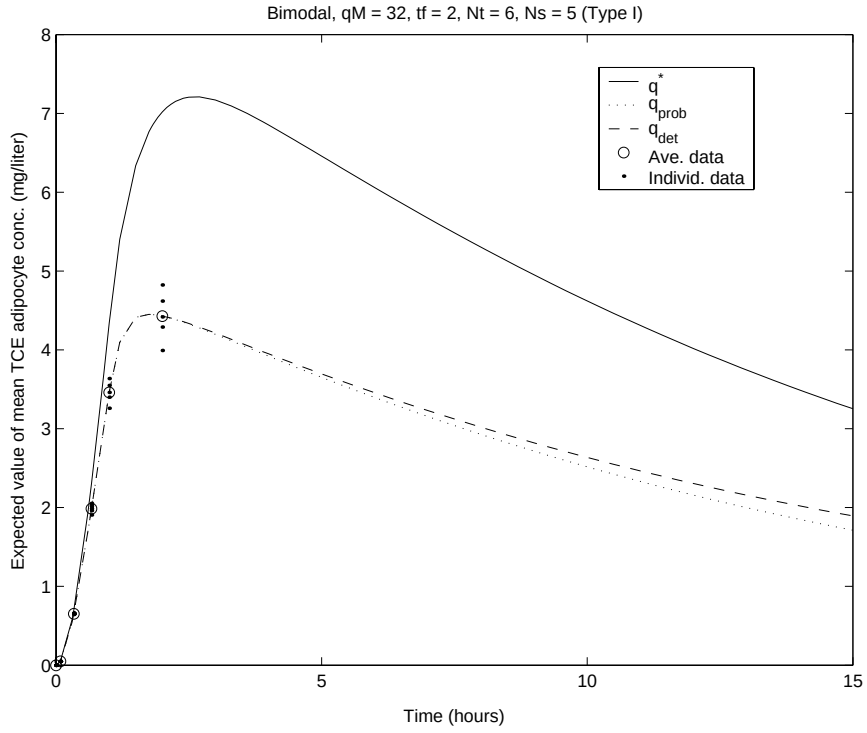

Fig. 4. Simulated observations and predicted TCE adipocyte concentrations using optimized parameters q_{det} from the deterministic estimation problem (30) and $q_{prob} = p_{opt}$ from the probabilistic problem (36) with $M = 32$. Observations used in the estimation problems were Type I data with $\vec{t}_1$ ($t_f = 2$ hours, $N_t = 6$) and with $N_s = 5$. In this and all similar figures that follow, the solid black dots are individual observations and the black open circles, which are used in the estimation problem, are the averaged values of the individual observations at each time point. The solid line is the model response corresponding to the data-generating bimodal distribution $P^* = P_{bi}$ using (4) with P^*. The dashed and dotted lines are the model responses corresponding to q_{det} and q_{prob}, respectively.

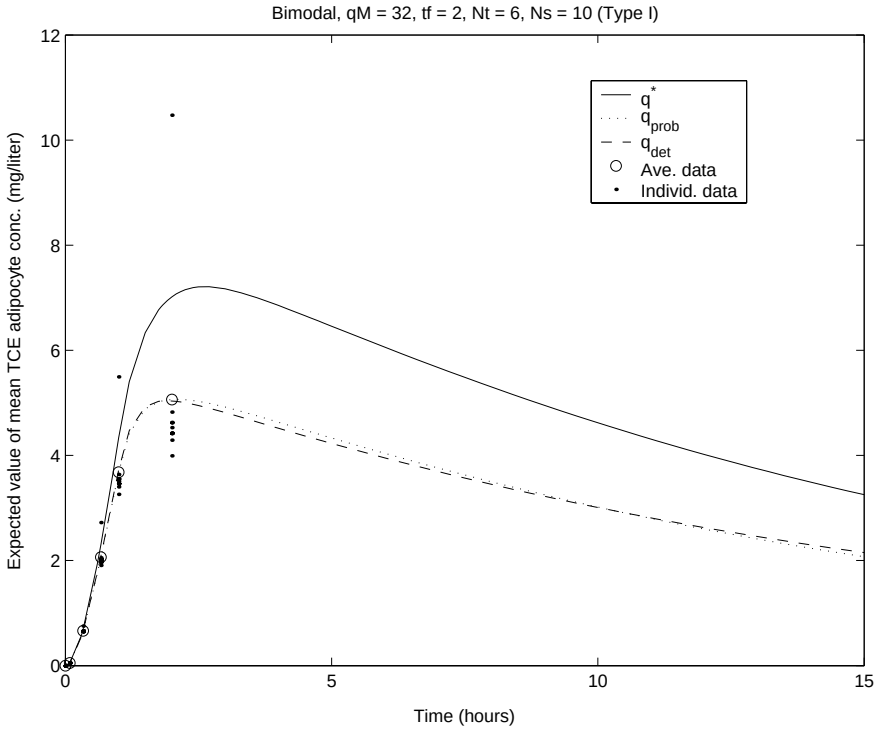

Fig. 5. Simulated observations and predicted TCE adipocyte concentrations using optimized parameters q_{det} from the deterministic estimation problem (30) and q_{prob} from the probabilistic problem (36) with $M = 32$. Observations used in the estimation problems were Type I data with $\vec{t}_1$ ($t_f = 2$ hours, $N_t = 6$) and with $N_s = 10$.

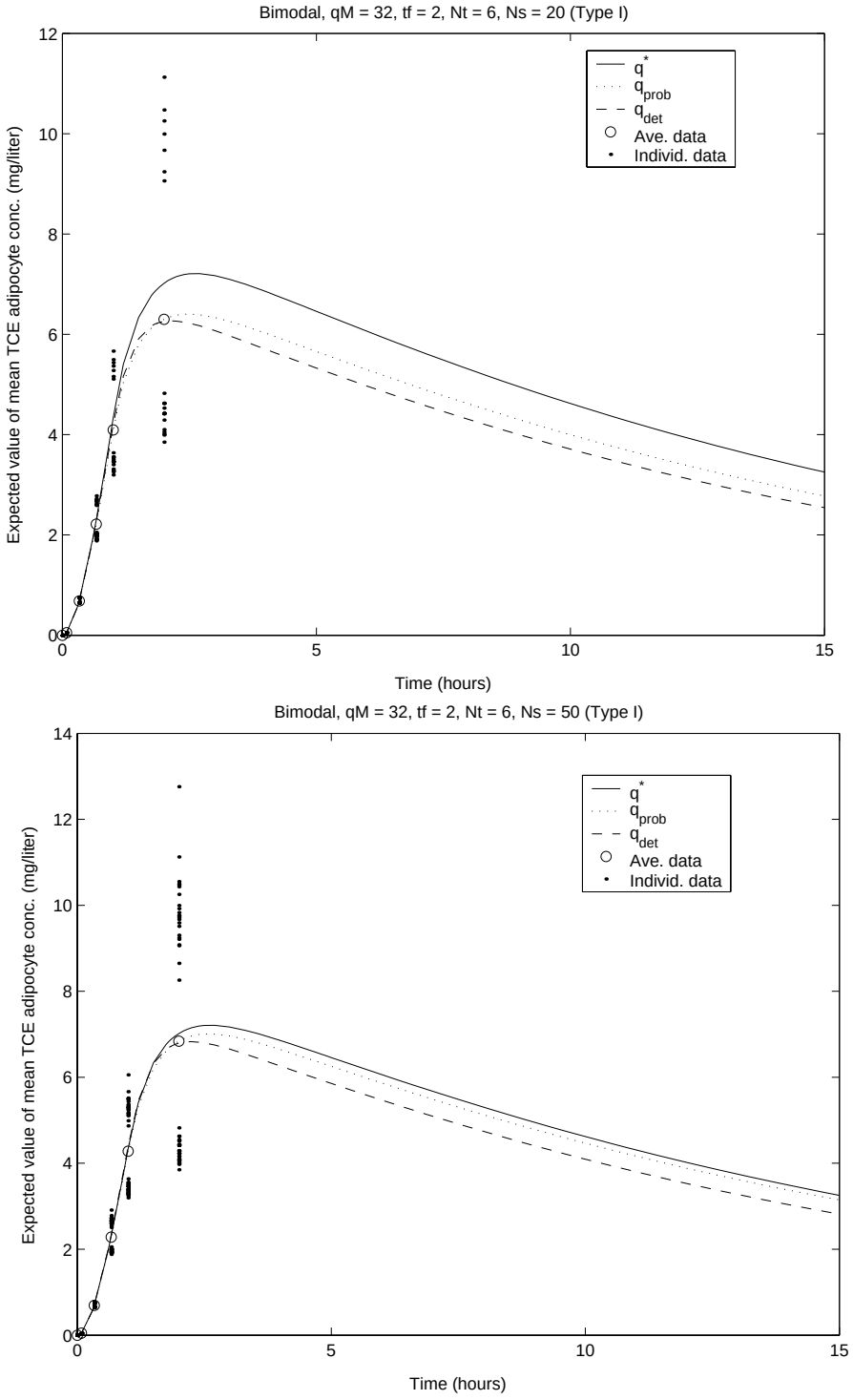

Fig. 6. Simulated observations and predicted TCE adipocyte concentrations using optimized parameters q_{det} from the deterministic estimation problem (30) and q_{prob} from the probabilistic problem (36) with $M = 32$. Observations used in the estimation problems were Type I data with $\vec{t}_1$ ($t_f = 2$ hours, $N_t = 6$) and with $N_s = 20$ (top figure) and $N_s = 50$ (bottom figure).

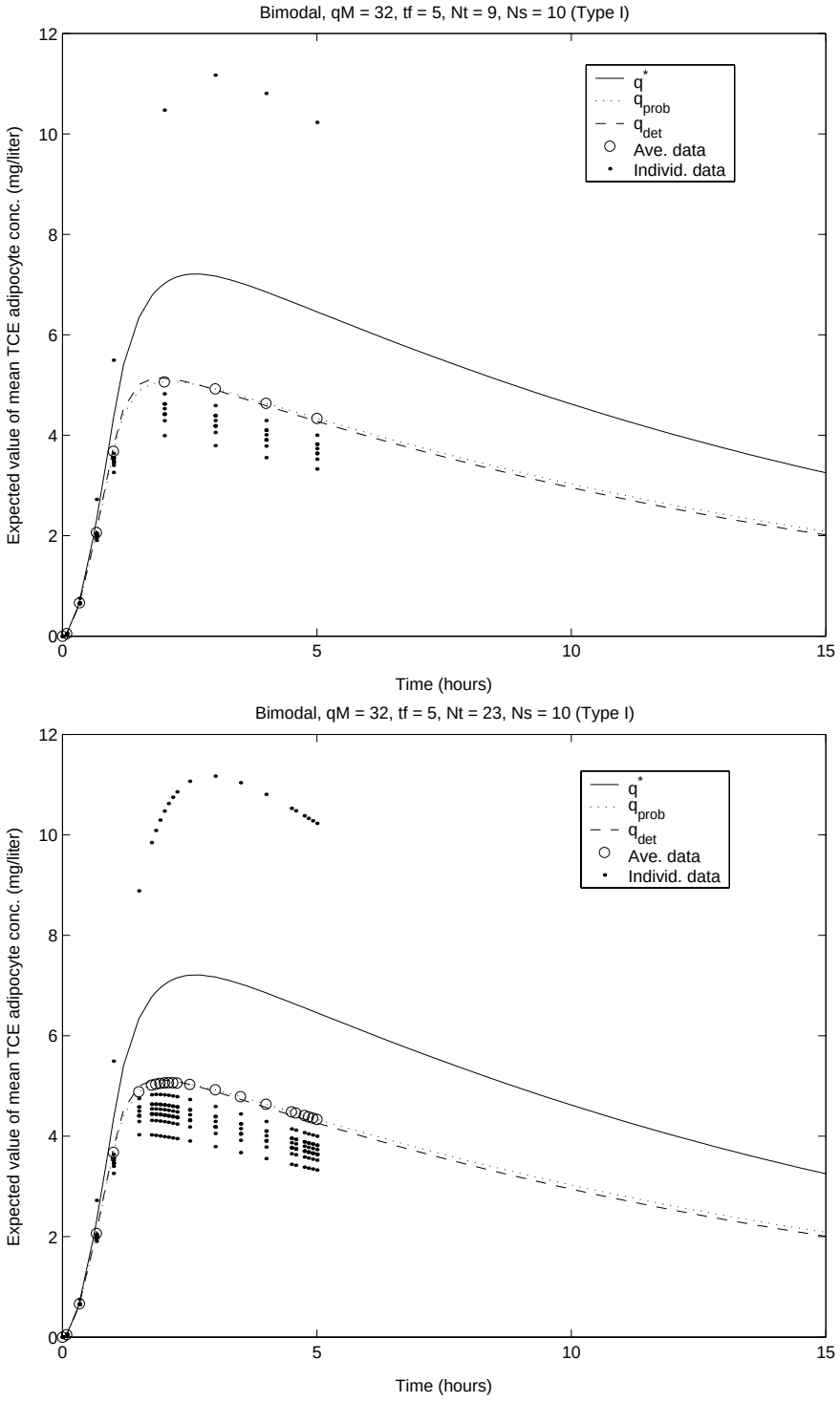

Fig. 7. Simulated observations and predicted TCE adipocyte concentrations using optimized parameters q_{det} from the deterministic estimation problem (30) and q_{prob} from the probabilistic problem (36) with $M = 32$. Observations used in the estimation problems were Type I data with $N_s = 10$ and with $\vec{t}_2$ (top figure) and $\vec{t}_3$ (bottom figure).

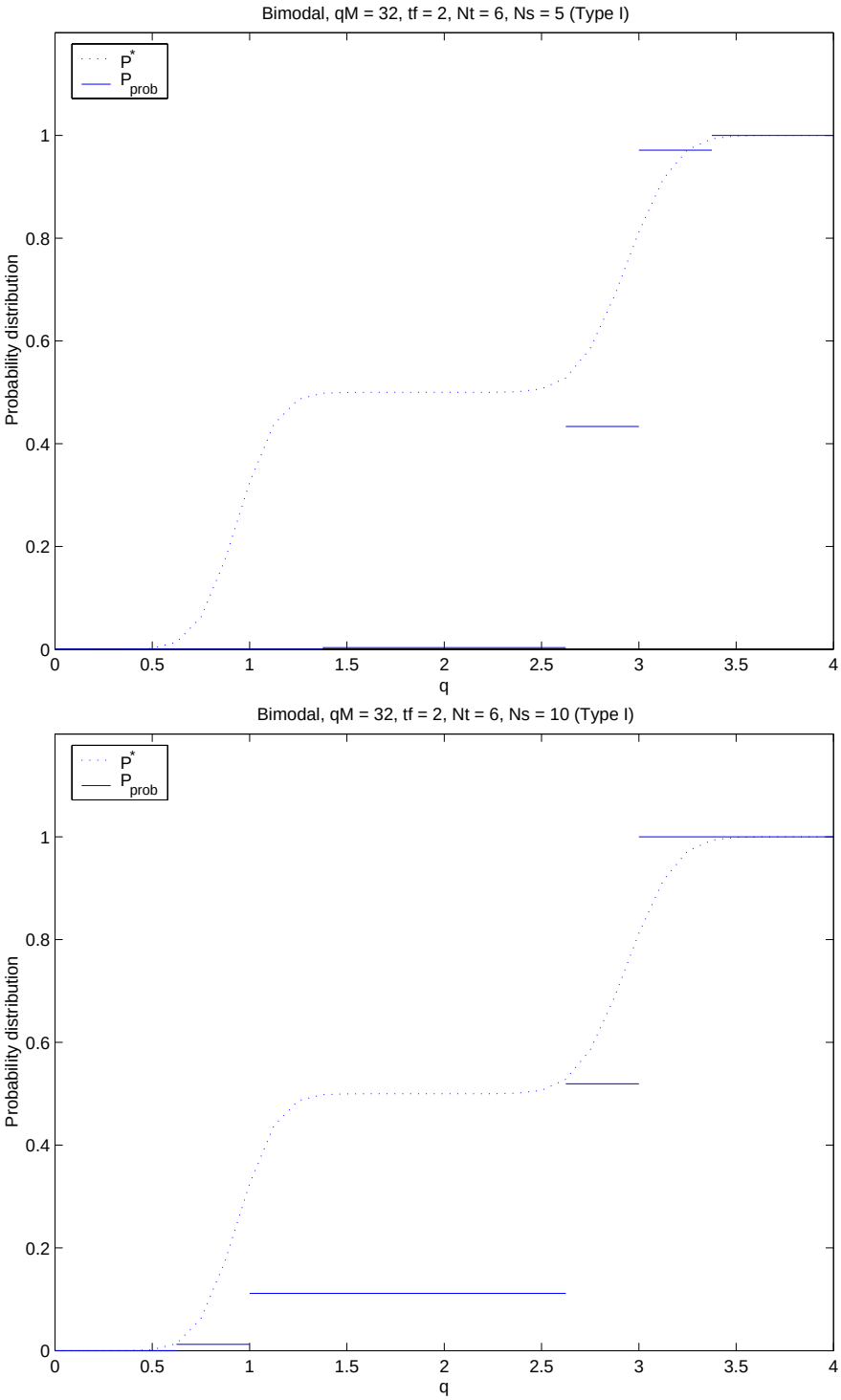

Fig. 8. Probability distribution functions P as a function of the parameter $q = \mathcal{D}_B$ for the data-generating distribution $P^* = P_{bi}$ and the optimized distribution P_{prob} from the probabilistic estimation problem (36) with Type I data, $M = 32$, $\vec{t}_1$ and with $N_s = 5$ (top figure) and $N_s = 10$ (bottom figure).

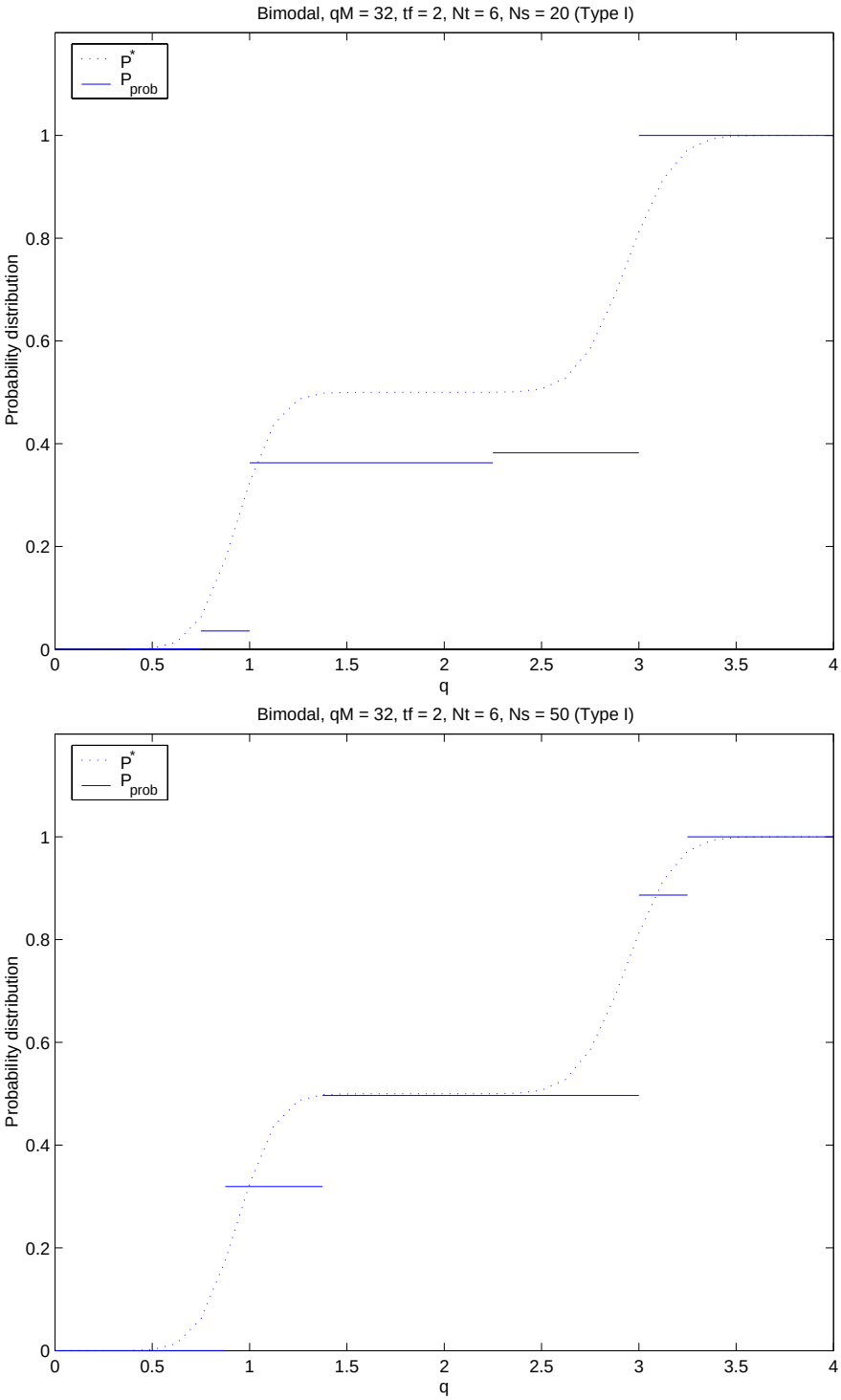

Fig. 9. Probability distribution functions P as a function of the parameter $q = \mathcal{D}_B$ for the data-generating distribution $P^* = P_{bi}$ and the optimized distribution P_{prob} from the probabilistic estimation problem (36) with Type I data, $M = 32$, $\vec{t}_1$ and with $N_s = 20$ (top figure) and $N_s = 50$ (bottom figure).

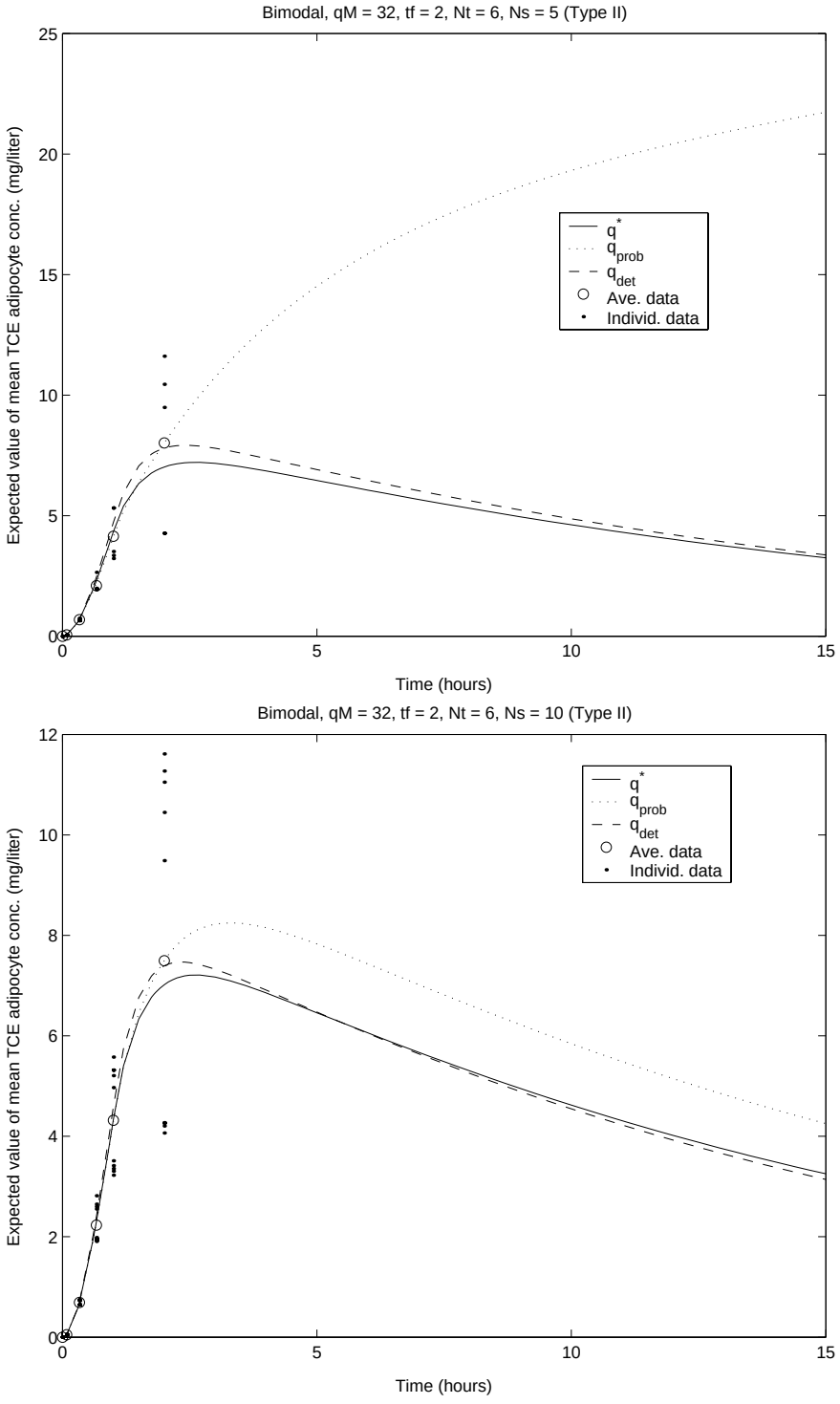

Fig. 10. Simulated observations and predicted TCE adipocyte concentrations using optimized parameters q_{det} from the deterministic estimation problem (30) and q_{prob} from the probabilistic problem (36) with $M = 32$. Observations used in the estimation problems were Type II data with $\vec{t}_1$ ($t_f = 2$ hours, $N_t = 6$) and with $N_s = 5$ (top figure) and $N_s = 10$ (bottom figure).

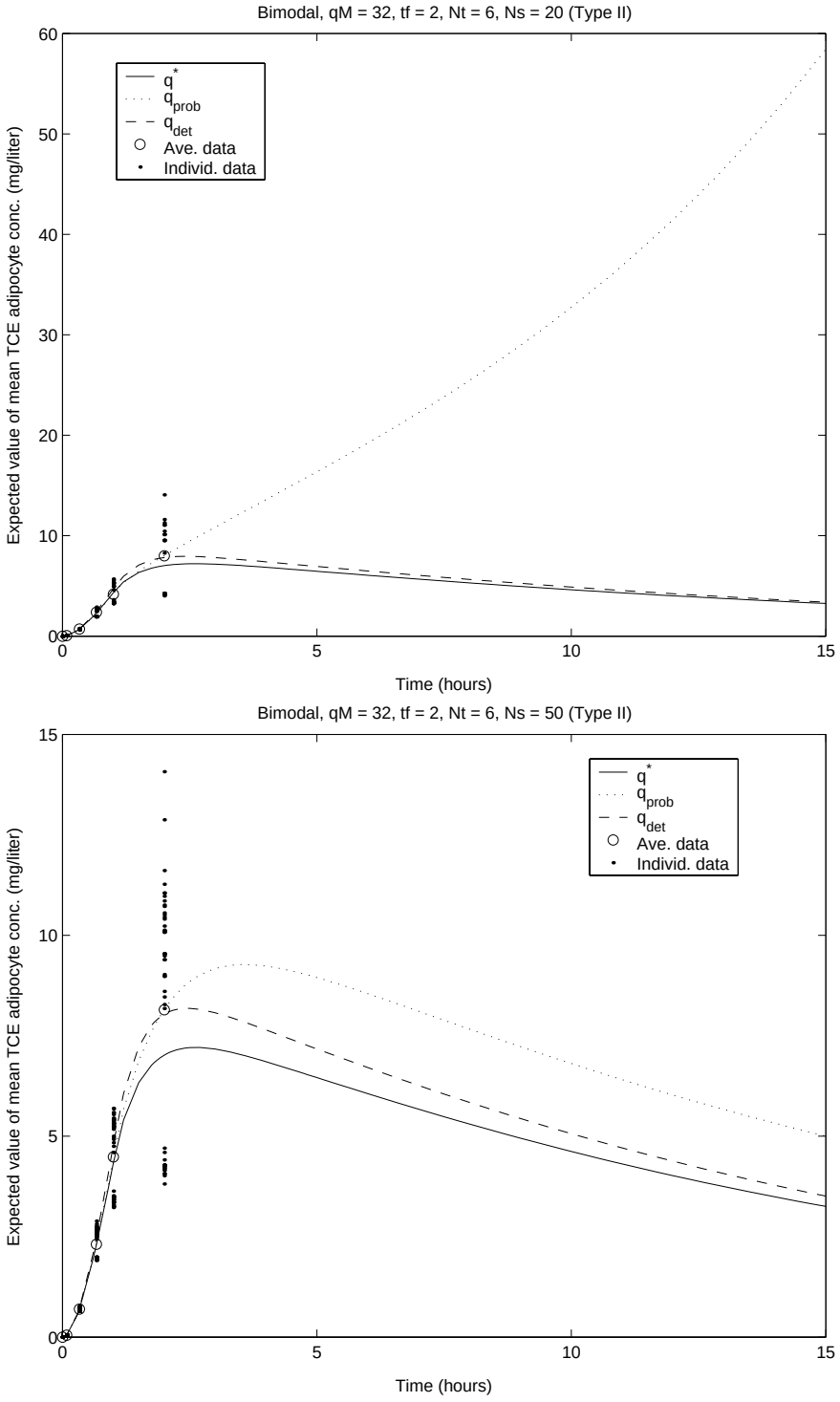

Fig. 11. Simulated observations and predicted TCE adipocyte concentrations using optimized parameters q_{det} from the deterministic estimation problem (30) and q_{prob} from the probabilistic problem (36) with $M = 32$. Observations used in the estimation problems were Type II data with $\vec{t}_1$ ($t_f = 2$ hours, $N_t = 6$) and with $N_s = 20$ (top figure) and $N_s = 50$ (bottom figure).

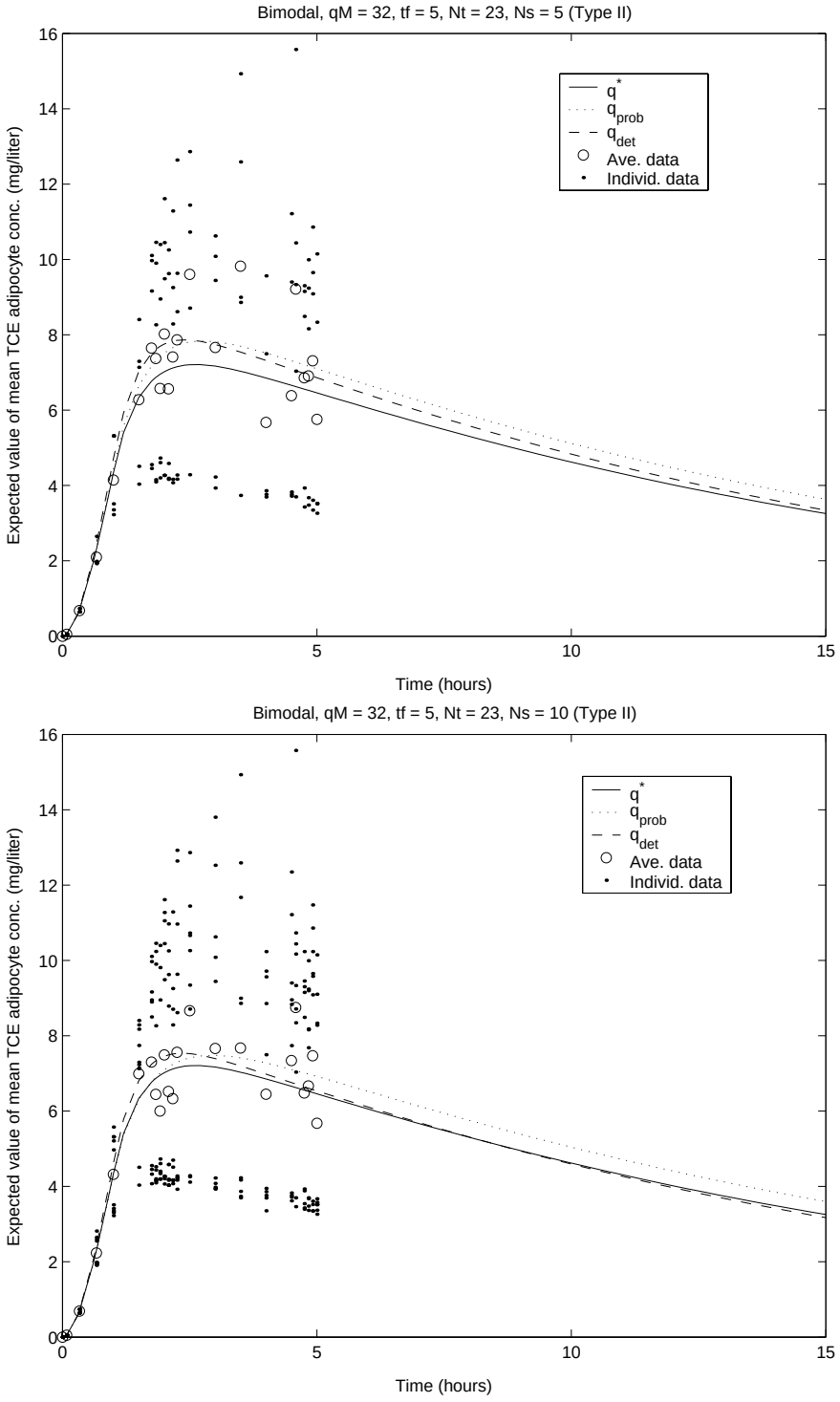

Fig. 12. Simulated observations and predicted TCE adipocyte concentrations using optimized parameters q_{det} from the deterministic estimation problem (30) and q_{prob} from the probabilistic problem (36) with $M = 32$. Observations used in the estimation problems were Type II data with $\vec{t}_3$ ($t_f = 5$ hours, $N_t = 23$) and with $N_s = 5$ (top figure) and $N_s = 10$ (bottom figure).

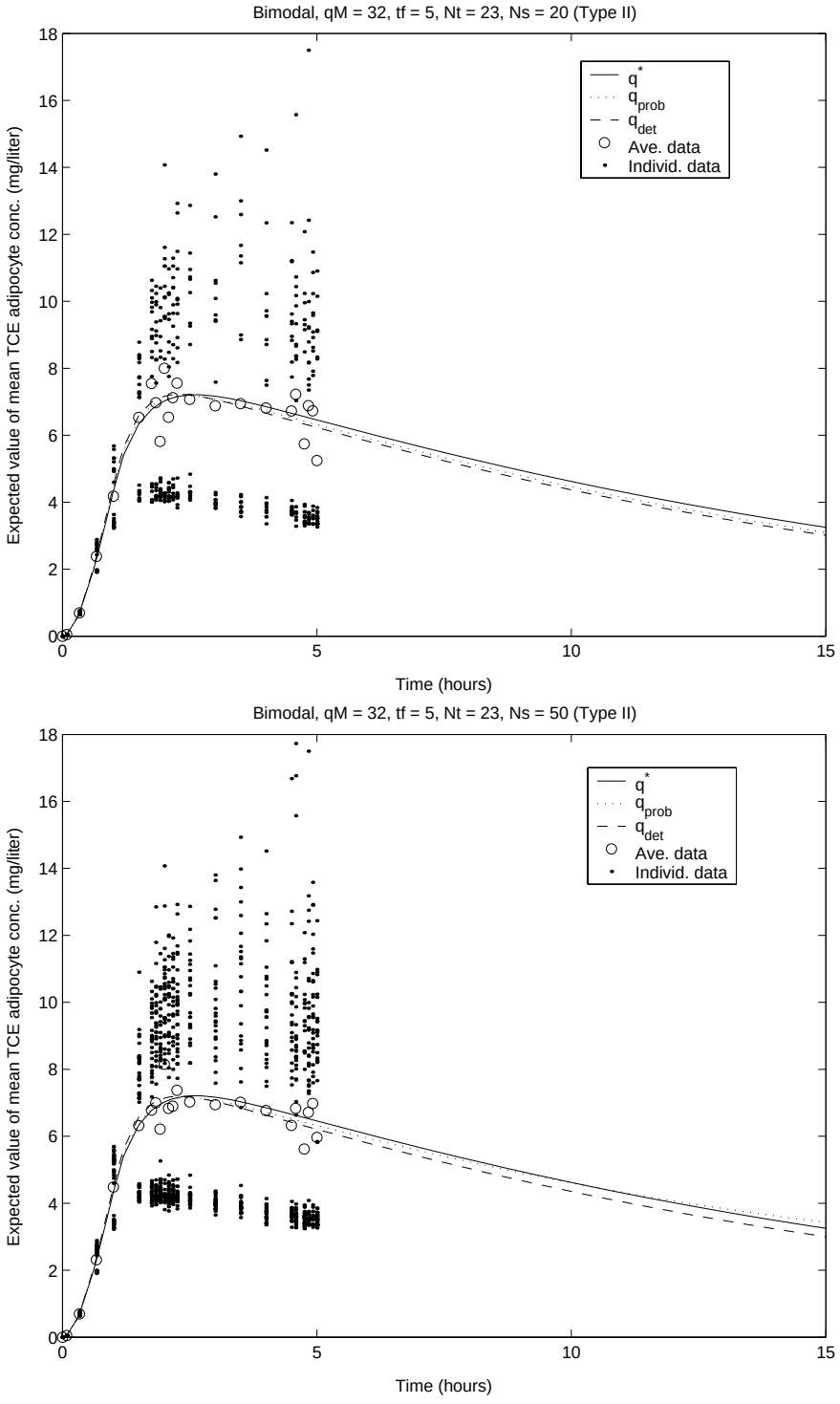

Fig. 13. Simulated observations and predicted TCE adipocyte concentrations using optimized parameters q_{det} from the deterministic estimation problem (30) and q_{prob} from the probabilistic problem (36) with $M = 32$. Observations used in the estimation problems were Type II data with $\vec{t}_3$ ($t_f = 5$ hours, $N_t = 23$) and with $N_s = 20$ (top figure) and $N_s = 50$ (bottom figure).

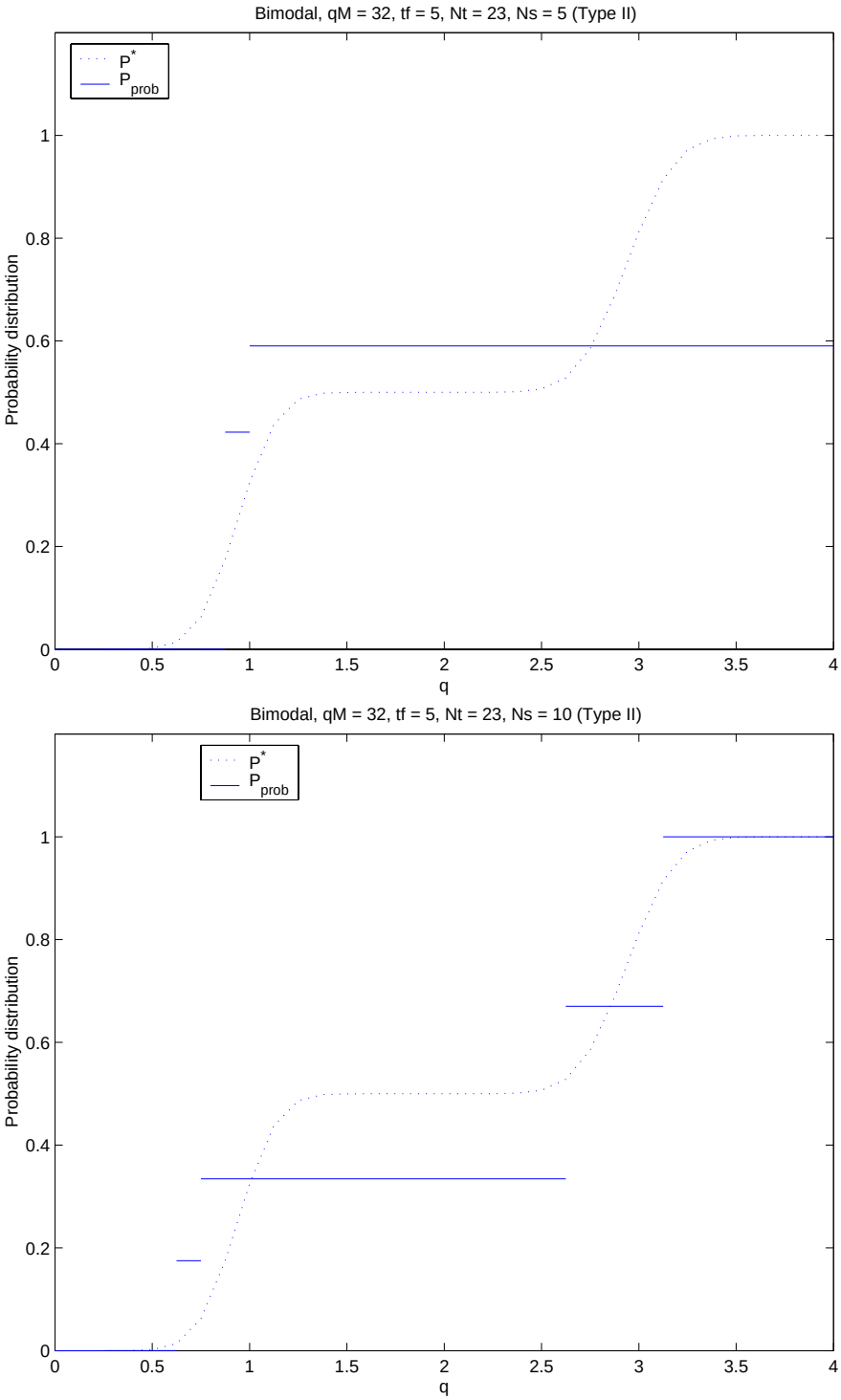

Fig. 14. Probability distribution functions P as a function of the parameter $q = \mathcal{D}_B$ for the data-generating distribution $P^* = P_{bi}$ and the optimized distribution P_{prob} from the probabilistic estimation problem (36) with Type II data, $M = 32$, $\vec{t}_3$ and with $N_s = 5$ (top figure) and $N_s = 10$ (bottom figure).

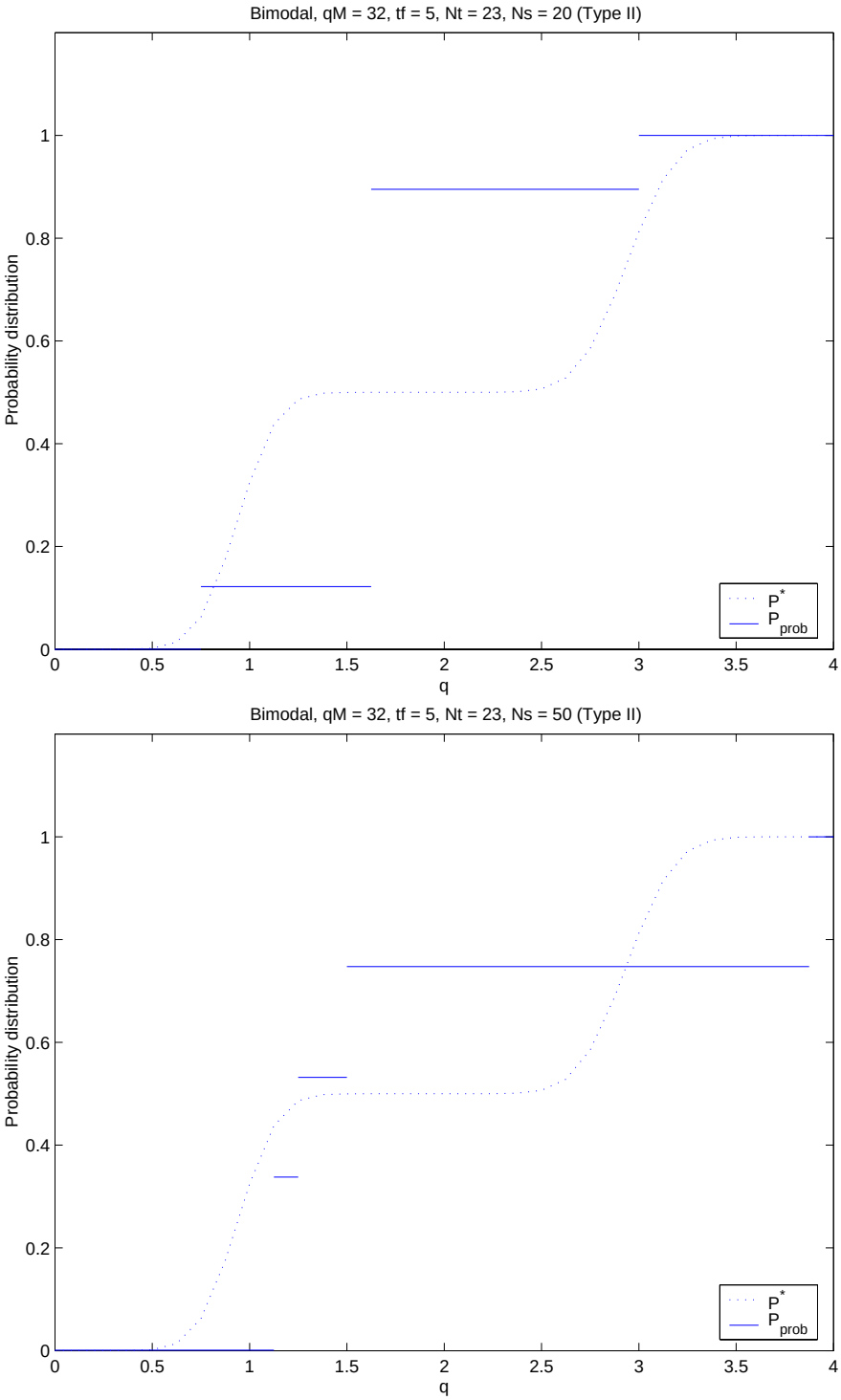

Fig. 15. Probability distribution functions P as a function of the parameter $q = \mathcal{D}_B$ for the data-generating distribution $P^* = P_{bi}$ and the optimized distribution P_{prob} from the probabilistic estimation problem (36) with Type II data, $M = 32$, $\vec{t}_3$ and with $N_s = 20$ (top figure) and $N_s = 50$ (bottom figure).