Estimation of Cell Proliferation Dynamics Using CFSE Data

H. T. Banks*, Karyn L. Sutton, W. Clayton Thompson
Center for Research in Scientific Computation
Center for Quantitative Science in Biomedicine
North Carolina State University
Raleigh, NC 27695-8212
and
Gennady Bocharov[1], Dirk Roose[2],
Tim Schenkel[3] and Andreas Meyerhans[3]
[1]Institute of Numerical Mathematics, RAS, Moscow, Russia
[2]Department of Computer Science, Katholieke Universiteit, Lueven, Belgium
[3]Department of Virology, University of the Saarland, Homburg, Germany

December 23, 2009

Abstract

Advances in fluorescent labeling of cells as measured by flow cytometry have allowed for quantitative studies of proliferating populations of cells. The investigations [38, 39, 40] contain a mathematical model with fluorescence intensity as a structure variable to describe the evolution in time of proliferating cells labeled by carboxyfluorescein succinimidyl ester (CFSE). Here this model and several extensions/modifications are discussed. Suggestions for improvements are presented and analyzed with respect to statistical significance for better agreement between model solutions and experimental data. These investigations suggest that the new decay/label loss and time dependent *effective proliferation and death rates* do indeed provide improved fits of the model to data. Statistical models for the observed variability/noise in the data are discussed with implications for uncertainty quantification. The resulting new cell dynamics model should prove useful in proliferation assay tracking and modeling, with numerous applications in the biomedical sciences.

Key words: Cell proliferation, CFSE, label structured population dynamics, partial differential equations, inverse problems.

*Corresponding author: htbanks@ncsu.edu

1 Introduction

Much progress in the quantification of cell population dynamics has been made in the last several years. Such improvements have allowed for the application of these methods to the investigation of questions in the life sciences in which proliferation plays a key role. For example, accurate quantification of changes in the rates at which various lymphocytes divide, differentiate and die can be used as a marker for changes in an immune response. Thus better understanding of cell proliferation can lead to improvements in treatment for disease (such as cancer, HIV and other viruses, etc.) progression. In the past several decades, proliferative assays have been carried out through incorporation of 5-Bromo-2'-deoxyuridine (BrdU) or tritiated thymine deoxyriboside (3HTdr) [42], both of which take the place of thymidine in the DNA of dividing cells, the latter of which is radioactive. In comparison to these two, carboxyfluorescein succinimidyl ester (CFSE) is more stably and evenly incorporated into cells, is detected easily by flow cytometry, and is nonradioactive [20, 42, 47]. It is not surprising, then, that CFSE has become the de facto staining method for many cell labeling studies [31, 41, 47]. The development of CFSE-based cytometry assays [43] in conjunction with a Fluorescence-Activated Cell Sorter (FACS) provides biologists with the ability to measure some of these fundamental properties of a population of cells quickly and efficiently. However, there may be distinct benefits to be gained from alternative modeling approaches. To that end, the development of models of the type discussed in this work can contribute to the quantitative understanding of cellular behavior as represented by FACS data.

CFSE is introduced into a population of cells as (membrane permeable) carboxyfluorescein diacetate succinimidyl ester (CFDA-SE). After CFDA-SE diffuses across the cell membrane, enzyme reactions with cellular esterases cleave the acetate groups, resulting in highly fluorescent and membrane impermeable CFSE. At reasonable concentrations and near neutral pH, the incorporation of CFSE does not in any way adversely affect the function of the cell. As cells divide, the CFSE fluorescence intensity (FI) is split roughly evenly between the two daughter cells. Thus, measurement of FI provides an indirect measure of the number of divisions a cell has undergone. After staining, the cell population is analyzed at regular time intervals via FACS, which returns a histogram of the number of cells as a function of CFSE FI. Typically, CFSE can be used to track up to 8 rounds of division before the CFSE FI is reduced to the autofluorescence level of unstained cells. More information regarding the biological processes and experimental protocol can be found in [31, 41, 42, 43, 47].

While the ability to track a population of cells through multiple rounds of divisions has been greatly improved by the use of CFSE labelling and quantitative methods, there are possible improvements. In addition to the work of Bocharov, et al., [40] upon which this work is based, many cell proliferation models have focused on describing growth and division dynamics of cells as a function of the number of divisions undergone, which is strongly correlated with CFSE intensity. The Smith-Martin cell cycle model [51], in which the cell cycle is divided into a stochastic resting G_1 phase and a deterministic dividing SG_2M phase, provides the basis for many of these papers [18, 22, 24, 27, 32, 37, 38]. These models vary in structure, drawing on a range of techniques including compartmental modeling, agent-based modeling and probability-based modeling. While this is by no means an exhaustive list, there are a number of areas of study in which these approaches have contributed: immunoglobulin class switching in B cells [33], cytokine regulation of T cells [29], and surface molecule expression or internal expression of cytokines [19, 28]. One notable recent effort is the paper by Lee, et al., 2009, in which a generalized Smith-Martin model with division-dependent death rate is compared to a cyton model (developed in [32]). The validated cyton model is then used to study the effects of IL-4,

thought to protect against apoptosis, on B cell population dynamics.

While these models are typically strongly biologically motivated, there are some drawbacks when considering their use with flow cytometry data. Some deconvolution is required of the CFSE data to obtain the number of cells as a function of the number of divisions undergone since initiation of the assay. In this process a distribution for the CFSE across cells of the same division number is assumed, which may be reasonable in most cases. However, it is possible that there may be certain cell populations, either normal or abnormal, that do not fit these distributions. Even if the assumed distribution is reasonable, it is possible that errors are made in counting some cells near the tails of the distributions corresponding to a given number of divisions. The alternative modeling approach we discuss in this manuscript is somewhat less restrictive in that it does not assume any distribution of label uptake of the cells. An additional benefit to this approach is that model solutions are directly compared to histograms obtained by flow cytometry, eliminating any possible misinterpretation of such data.

Building upon the work in [39, 40], we seek to update these models (in which age is a discrete variable corresponding to division number) with a hyperbolic partial differential equation (PDE) model for the label-structured population density in which CFSE FI is a continuous state variable. Because FI intensity is determined by cellular events included in the model, it is useful to study cells as a function of FI in time. Divisions of cells are seen in model solutions, just as in histograms obtained by flow cytometry, but specifying division number is unnecessary. Meanwhile, because FI can range continuously, the nonuniform uptake of CFSE (and hence the resulting nonuniform distribution of CFSE in each generation) is preserved.

Previous work [40] has already demonstrated the advantages of this type of PDE model as compared to some compartmental modeling efforts. While this model provided a reasonable way to mathematically reproduce experimentally observed proliferation dynamics, there was still room for improvement as peaks/generations of cells were incorrectly predicted by the best fit model solution. In this report, careful consideration is given to the biological and mathematical assumptions of the original model and refinements in both interpretation and parameter/mechanism formulations are proposed. First, the proliferation and death rate functions from [40] are redefined on a domain which strongly correlates with division number, and the proliferation rate function is changed so that it is time varying. It is shown via a model comparison test that the resulting model provides a statistically significant improved fit to an experimental data set over the previous model. Evidence is then offered to suggest that the death rate for cells should remain division dependent when compared to fitting the data set with a constant rate. Next, improvements to the treatment of the label loss rate are considered. It is found that the best fit to the data is not improved significantly when this function is either affine in form or taken to be a probability distribution of constant label loss rates within the population. For each version of the model, the parameters are estimated in an ordinary least squares inverse problem and the resulting cost functions are tested statistically to quantitatively assess the improvement in fitting to the data. Statistical models for the noise structure in the data are also discussed. It is shown that the noise in the data does not appear to have either constant variance (ordinary least squares formulation for absolute error) or variance proportional to the magnitude of the observation (generalized least squares formulation for relative error). Implications for the corresponding quantification of parameter uncertainty via confidence intervals are discussed.

The investigations in this paper demonstrate that the proposed PDE models can be used successfully in an interpretive framework for cell division dynamics and lay the groundwork for the continued refinement and extension of these models and their application to additional data sets. Specifically, two new features in the current models offer a dramatic improvement in

fits to data in the context of biologically supportable model mechanisms. First, we introduce a decay dependent *translation of intensity coordinate* resulting in a dependent variable more strongly correlated to division number with which we define new *effective proliferation and death rates*. A further extension of the effective proliferation rates (which depend implicitly on time) to depend explicitly on time accounts for differences in proliferation rates beyond division dependence also results in statistically significant improved performance of the models. The ability to estimate from data these effective rates (relative to the new coordinates) allows for their comparison in healthy cell populations versus cell populations with abnormal growth and/or proliferation dynamics. The characterization of these differences can potentially improve our detection, identification and understanding of the disease or conditions responsible for the change in dynamics.

2 Preliminaries

2.1 Data

An original data set (shown in Figure 1) containing time-series snapshots of the CFSE FI distribution of a population of dividing cells was used [40]. Briefly, this data set is the result of an in vitro proliferation assay with human peripheral blood mononuclear cells (PBMCs) from healthy blood donors. After isolation of the PBMCs from 'buffy coats' by density centrifugation, 5×10^6 to 5×10^7 were stained with 5 μM CFDA-SE (Invitrogen, Germany) in phosphate-buffered saline containing 5% fetal calf serum (FCS). The cells were stimulated with 2.5 μg/mL phytohemagglutinine (Sigma, Germany) at time $t = 0$ hrs and plated in 24 well plates at 1×10^6 cells/mL RPMI-1640/10% FCS medium. Beginning at day 3, every 24 hrs one third of the medium was exchanged with fresh medium to ensure sustained cell nutrition. To not disturb the proliferating cell populations, cells from a single well were harvested for each time point. Cells were then stained with CD4 antibodies. This staining makes it possible to distinguish the CD4+ cells from other cells in the PBMC culture while simultaneously measuring CFSE expression in individual cells through FACS. This measurement process is high-speed (thousands of cells in seconds), and provides data for individual cells [31, 47].

The CFSE FI data is reported on a logarithmic scale, $z = \log_{10} x$ where x is the CFSE FI of a given cell. The output of the measurement procedure is the counts c_{ij} of cells at time t_i with log-intensity z_j. The data set obtained for this report tracks cells from 0 to 120 hours in 24 hour intervals; the discretization of the z-axis into bins changes with each measurement in time ($j = 1, \ldots, J(i)$ for each time t_i). Data for the entire population density can then be obtained by the transformation

$$n_{ij}^d = \frac{c_{ij} T_i}{C_i},$$

where T_i is the total number of cells in the population at time t_i and C_i is the total number of cells counted at time t_i. This processing of the data was performed by the authors of [40] so that the data set obtained already consisted of the population density data n_{ij}^d.

Note in Figure 1 that an initially unimodal peak becomes multimodal as cohorts of cells begin to divide at different times. Each of these division peaks also slowly drifts to the left as FI is lost over time due to catabolic activity within the cell [41]. The noise in the data is typical for such experiments and is the result of any number of processes, from counting errors to variations in cell shape and size to the functioning of the machine itself [40, 52].

As alluded to in the introduction and discussed at greater length in [40, 47], many approaches to cell division tracking have involved the determination of the number of cells hav-

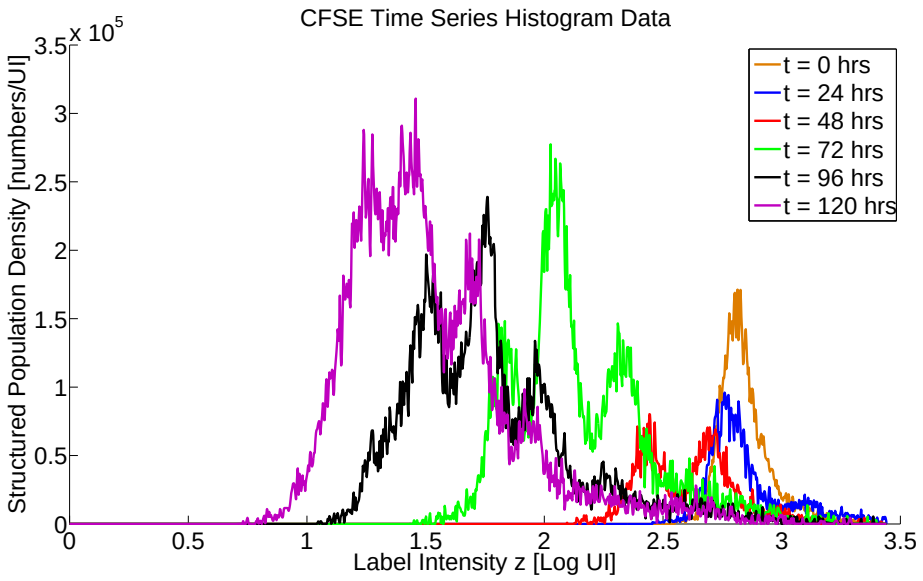

Figure 1: Original CFSE histogram data.

ing undergone a certain number of divisions. But this information is not available from the experimental data without some sort of deconvolution technique to separate out the division peaks. These techniques invariably involve the definition of FI borders between subsequent division peaks. However, such techniques do not account for the inherent heterogeneity of the cell population due to variation in initial CFSE uptake, variability across cells in catabolism, proliferation, etc. that may result in cells at a given FI having undergone a different number of divisions. Moreover, this heterogeneity can make subsequent division peaks very difficult to resolve, introducing additional error into these traditional approaches. Hence, an advantage of the current approach is that it eliminates the need for these deconvolution techniques by explicitly accounting for the heterogeneity of the population through the use of CFSE FI as a continuous state variable.

2.2 Mathematical Model

The model for the dynamics of life and death processes of a population of cells labelled with CFSE is proposed in [40] as a variation of a Bell-Anderson [16] or Sinko-Streifer [50] population model. The solutions of these models may be directly compared to time snapshots of flow cytometry histograms, which usually depict cells of multiple generations. That is, the models aim to predict the number of cells at a given fluorescence intensity and time, under specified dynamics. Let x denote the CFSE FI (in units of intensity, UI) of a cell and let $n(t,x)$ be the label-structured population density (cells/UI) of cells with FI x at time t. Then the population density is governed by a hyperbolic partial differential equation (PDE)

$$\begin{aligned} \frac{\partial n}{\partial t}(t,x) \quad &+ \quad \frac{\partial [v(x)n(t,x)]}{\partial x} = \\ &- \quad (\alpha(x)+\beta(x))n(t,x) + \chi_{[x_{\min},x_{\max}/\gamma]} 2\gamma\alpha(\gamma x)n(t,\gamma x), \end{aligned} \tag{1}$$

where $v(x)$ is the label loss rate, $\alpha(x)$ is the cell proliferation rate, $\beta(x)$ is the cell death rate, $x \in [x_{\min}, x_{\max}]$ and $t > 0$. Because cells naturally lose FI over time even in the absence

of division (due to catabolic activity [41]), the term $v(x)$ represents the natural label loss rate (UI/hr). The parameter γ is the label dilution factor, representing the ratio of FI of a mother cell to FI of a daughter cell. Division is coupled with immediate rapid growth of the new daughter cells. The observed or estimated value of γ reflects underlying dynamics (involving mechanisms regulating the growth and division) which occur on a faster time scale and have effectively been integrated over in time. A derivation of this model following the mass conservation principles of Sinko-Streifer [50] or Bell-Anderson [16] models is presented in the Appendix.

Because FACS returns data on a logarithmic scale, it is convenient to make the change of variables $z = \log_{10} x$. Assuming the natural label loss is proportional to the amount of label, we take $v(x) = -cx$, a form which has been seen [40] to better fit this data set than constant label loss assumptions. The resulting model, along with the change of variable $\tilde{n}(t,z) \equiv n(t,10^z)$ gives

$$\begin{aligned} \frac{\partial \tilde{n}}{\partial t}(t,z) \quad &+ \quad \frac{\partial[\tilde{v}(z)\tilde{n}(t,z)]}{\partial z} = \\ &- \quad (\tilde{\alpha}(z)+\tilde{\beta}(z))\tilde{n}(t,z) + \chi_{[z_{\min}, z_{\max}-\log_{10}\gamma]} 2\gamma\tilde{\alpha}(z+\log_{10}\gamma)\tilde{n}(t,z+\log_{10}\gamma), \end{aligned} \tag{2}$$

where $\tilde{v}(z) = -\tilde{c} = -c/ln10$, and $\tilde{\alpha}, \tilde{\beta}$ are appropriately defined cell proliferation and death rates, respectively. The initial and boundary conditions for the model are

$$\begin{aligned} \tilde{n}(0,z) &= \tilde{n}_0(z) \\ \tilde{n}(t,z_{\max}) &= 0. \end{aligned}$$

In our subsequent discussions, we will drop the tildes on the parameters α, β, c, v and the states n in (2) and take this equation as our *fundamental model* to be investigated and modified. From the structure of the above model, we deduce that some key tacit assumptions are:

i.) Division numbers are strongly correlated with FI;

ii.) FI is proportional to total CFSE content (amount);

iii.) Total CFSE is divided equally among daughter cells with each division;

iv.) The rate of label loss $v(z)$, the proliferation rate $\alpha(z)$, and the death rate $\beta(z)$ do not depend on time.

Assumption ii.) appears to be tacitly made in the discussions of CFSE content or amount, CFSE FI and the definition of the parameter γ on page 4 of [40] which we will understand here as being defined by the mother/daughter ratio of CFSE FI. Assumptions i.) - iii.) would imply that the state variable z is strongly correlated with, although not exactly equal to, division number. Hence assumption iv.) would then be equivalent to stating that birth, death, and label loss rates, largely depending on division number, can be determined as functions of label intensity z. (It will later be shown, however, that we can modify some of these assumptions in their interpretation and implementation to obtain model extensions to produce significantly improved model agreement with the data.) Given the formulation (2), the goal is to use the CFSE data described in Section 2.1 in order to estimate the functions $\alpha(z)$, $\beta(z)$ and $v(z)$ as well as the parameter γ. Following by now standard inverse problem procedures [10, 14], these functions are parameterized by finite-dimensional approximations so that the estimation is computationally tractable (and theoretically sound - see the convergence arguments of [10,

12, 14] in such inverse problem approximations). Both $\alpha(z)$ and $\beta(z)$ are approximated by linear splines

$$\alpha(z) = \sum_{i=1}^{M_\alpha} a_i \phi_i(z), \qquad \beta(z) = \sum_{i=1}^{M_\beta} b_i \phi_i(z),$$

where $\phi_i(z)$ are piecewise linear spline functions satisfying

$$\phi_i(z_j) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} .$$

Note that $a_i = 0$ or $b_i = 0$ indicates zero birth or death respectively while $a_i = 1$ or $b_i = 1$ indicates that all cells at the given FI are dividing or dying on average once per hour, which is clearly a greater rate than these events actually occur. Thus the parameters a_i and b_i are constrained to be between 0 and 1. As already noted above, it is assumed that the rate of label loss is proportional to label intensity ($v(x) = -\bar{c}x$, where $\bar{c} \in \mathbb{R}^+$). Thus after the change of variables, we have $v(z) = -c = -\bar{c}/ln10$. Simulations of the model (2) demonstrate that $0.0025 \leq c \leq 0.0055$ provides a reasonable range of possible values for the label loss rate. Similarly, γ is reasonably constrained to $\gamma \in [1, 2]$. For biological interpretation of this range, see Section 3.2.

Parameter	Minimum	Maximum	Units
a_i	0	1	hr^{-1}
b_i	0	1	hr^{-1}
c	0.0025	0.0055	UI/hr
γ	1	2	[none]

Table 1: Summary of parameters to be estimated, with minimum values, maximum values, and units. The a_i and b_i are coefficients of the effective proliferation and death functions, respectively. The parameter c scales the label loss as a function of fluorescence intensity, and γ is the *observed* ratio of mother to daughter cell CFSE concentration.

The set of parameters to be estimated in the inverse problem is given in Table 1. Given a set of these parameters, the forward problem is solved numerically over a specified time interval using a publicly available vectorized version of the Lax-Wendroff method with a nonlinear filter developed by L.F. Shampine [49] for solving hyperbolic PDEs in MATLAB. In order to obtain solution points on the non-uniform z-grid of the data points, different with each time t_j, the solution was calculated (for all time points) on a grid of 500 evenly spaced points for each time point and then interpolated onto the data grid using linear interpolation with MATLAB's `interp1` routine. Throughout this report, it is assumed that $[z_{\min}, z_{\max}] = [0, 3.5]$.

After a brief discussion of the theory and implementation of the inverse problem in Section 2.3, computational results which attempt to fit the above model to the data are presented in Section 3. Then, in an attempt to improve the fit of the model to the data, a transformed intensity variable depending on time and label loss is introduced into the model so that new *effective* proliferation and death rate functions are defined on a domain which is strongly correlated with division number. Next, the effective proliferation rate function (depending implicitly on time) will also be allowed to vary explicitly in time. Finally, further considerations regarding the use of a constant death rate and alternative parameterizations of the label loss rate are discussed. In each case, biological and physical justifications for the changes are given.

2.3 Mathematical & Statistical Aspects of the Inverse Problem

2.3.1 Ordinary Least Squares (OLS)

Given the mathematical model (2) for the label-structured population density $n(t, z; \vec{\theta})$ at time t, log intensity z and parameters $\vec{\theta}$ (see Table 1), the CFSE time-series histogram data constitutes a direct observation of the physical process. A statistical model for the observation process is given by

$$N_{ij}^d = n(t_i, z_j; \vec{\theta}_0) + \mathcal{E}_{ij}, \tag{3}$$

where $n(t_i, z_j; \vec{\theta}_0)$ is the solution to the model (2) at time t_i ($t_i \in [t_1, \ldots, t_I]$) and log FI z_j ($z_j \in [z_1, \ldots, z_{J(i)}]$) given the assumed true parameters $\vec{\theta}_0 \in \mathbb{R}^p$ that generate observations N_{ij}^d. The statistical model (3) relies on the assumption that the random errors in the data do not depend on the magnitude of the observations themselves. The noise $\mathcal{E}_{ij}$ are independent identically distributed (*i.i.d.*) random variables with mean zero which represent the measurement error. Hence N_{ij}^d are random variables as well, and the given data represent one realization of this random variable, i.e.,

$$n_{ij}^d = n(t_i, z_j; \vec{\theta}_0) + \epsilon_{ij}.$$

In the absence of any information regarding the distribution of $\mathcal{E}_{ij}$, it is assumed only that the variance $\mathrm{var}(\mathcal{E}_{ij}) = \sigma_0^2$ does not depend on t or z (because $\mathcal{E}_{ij}$ are *i.i.d.*). The ordinary least squares (OLS) estimator is defined as

$$\vec{\theta}_{\mathrm{OLS}} = \arg\min_{\vec{\theta} \in \Theta} \sum_{i=1}^{I} \sum_{j=1}^{J(i)} \left(N_{ij}^d - n(t_i, z_j; \vec{\theta}) \right)^2, \tag{4}$$

where Θ is a set of admissible parameters for the model. Given the data as a realization for this random variable, the OLS estimate is

$$\hat{\theta}_{\mathrm{OLS}} = \arg\min_{\vec{\theta} \in \Theta} \sum_{i=1}^{I} \sum_{j=1}^{J(i)} \left(n_{ij}^d - n(t_i, z_j; \vec{\theta}) \right)^2 = \arg\min_{\vec{\theta} \in \Theta} F_{\mathrm{OLS}}(\vec{\theta}). \tag{5}$$

The MATLAB constrained global minimization package `fmincon`, a gradient-based method for problems where the objective and constraint functions are continuous and have continuous first derivatives, was used to solve for the OLS estimate $\hat{\theta}_{\mathrm{OLS}}$.

The true covariance σ_0^2 of the random variables $\mathcal{E}_{ij}$ is given by

$$\sigma_0^2 = \frac{1}{N} E \left[\sum_{i=1}^{I} \sum_{j=1}^{J(i)} \left(N_{ij}^d - n(t_i, z_j; \vec{\theta}_0) \right)^2 \right]$$

where $N = \sum_{i=1}^{I} J(i)$. The bias-corrected estimate of the variance with observations $\{n_{ij}^d\}$ is

$$\hat{\sigma}_{\mathrm{OLS}}^2 = \frac{1}{N - p} F_{\mathrm{OLS}}(\hat{\theta}_{\mathrm{OLS}}).$$

One can also use an asymptotic theory involving sensitivity matrices for the corresponding covariance matrix $\Sigma(\hat{\theta}_{\mathrm{OLS}})$ associated with the estimated parameters. This can be used to compute standard errors and confidence intervals (see the detailed discussions in [3, 15]).

Rewriting Equation (3), we have $\mathcal{E}_{ij} = N_{ij}^d - n(t_i, z_j; \vec{\theta}_0)$. Thus the residuals

$$r_{ij} = n_{ij}^d - n(t_i, z_j; \hat{\theta}_{\text{OLS}}) \tag{6}$$

are a realization of the error in the data and, given a mathematical model for a set of data with constant variance (CV), these residuals should be randomly distributed (Figure 2, Left) when plotted against the model values, $n(t_i, z_j; \hat{\theta}_{\text{OLS}})$. In many situations, the variance in the observations is not constant but is proportional to the magnitude of the observations. If OLS is used with such data, the same plot will not be random but rather will exhibit a characteristic fanning pattern (Figure 2, Right), indicating dependence of the residuals on model quantities. In that case a generalized least squares (GLS) procedure, described in the next section, is appropriate. Thus the shape of the residuals after data fitting by a least squares minimization provides a means of investigating the reliability of the assumptions of the statistical model.

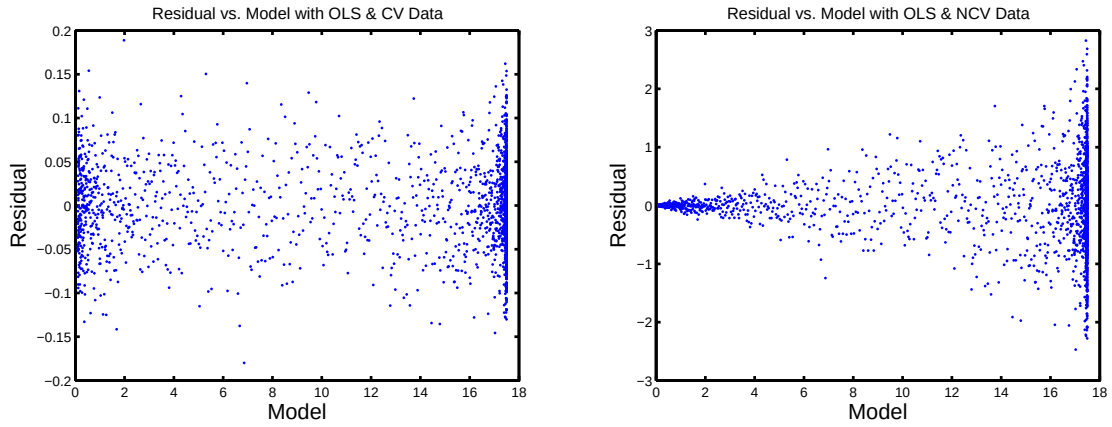

Figure 2: Left: Residuals vs Model for observations with constant variance. Right: Residuals vs Model for observations with nonconstant variance.

2.3.2 Generalized Least Squares (GLS)

When observational error is proportional to the magnitude of the observation, the statistical model is given by

$$N_{ij}^d = n(t_i, z_j; \vec{\theta}_0)(1 + \mathcal{E}_{ij}), \tag{7}$$

where $\mathcal{E}_{ij}$ are defined as before. Then $\text{var}(N_{ij}^d) = \sigma_0^2 n^2(t_i, z_j; \vec{\theta}_0)$. Accordingly, the generalized least squares (GLS) cost functional weights the observations according to their variance and the GLS estimator is

$$\vec{\theta}_{\text{GLS}} = \arg\min_{\vec{\theta}\in\Theta} \sum_{i=1}^{I} \sum_{j=1}^{J(i)} w_{ij} \left(N_{ij}^d - n(t_i, z_j; \vec{\theta}) \right)^2 = \arg\min_{\vec{\theta}\in\Theta} F_{\text{GLS}}(\vec{\theta}), \tag{8}$$

where the weights are

$$w_{ij} = \begin{cases} n^{-2}(t_i, z_j; \vec{\theta}), & n(t_i, z_j; \vec{\theta}) \geq N^* \\ 0, & n(t_i, z_j; \vec{\theta}) < N^* \end{cases} . \tag{9}$$

This definition of the weights prevents the GLS algorithm from giving unreasonably large weight to points where the model is near zero [13]. The value of N^* is chosen so that the maximum number of data points are chosen without the minimization algorithm stagnating or converging erroneously as a result of assigning little weight to larger-magnitude observations. The computation of the GLS estimate $\hat{\theta}_{\text{GLS}}$ is an iterative procedure and is given as follows:

i.) Compute the OLS estimate $\hat{\theta}^k_{\text{OLS}}$ according to Equation (5). Set $k = 0$.

ii.) Form the weights

$$w^k_{ij} = \begin{cases} n^{-2}(t_i, z_j; \hat{\theta}^k_{\text{OLS}}), & n(t_i, z_j; \hat{\theta}^k_{\text{OLS}}) \geq N^* \\ 0, & n(t_i, z_j; \hat{\theta}^k_{\text{OLS}}) < N^* \end{cases} .$$

iii.) Compute the approximation to the GLS estimate

$$\tilde{\theta}^{k+1} = \arg\min_{\vec{\theta} \in \Theta} \sum_{i=1}^{I} \sum_{j=1}^{J(i)} w^k_{ij} \left(n^d_{ij} - n(t_i, z_j; \vec{\theta}) \right)^2 .$$

iv.) Repeat steps ii.) and iii.), incrementing k by one with each iteration, until the successive estimates satisfy

$$\sum_{i=1}^{p} |\tilde{\theta}^{k+1}_i - \tilde{\theta}^k_i|^2 < \epsilon,$$

where ϵ is small, representing the predetermined desired convergence.

A more complete discussion of the GLS motivation, algorithm, and convergence can be found in [3, 15] and the references therein. The result of the algorithm is the GLS estimate $\hat{\theta}_{\text{GLS}}$. The variance coefficient may be estimated as

$$\hat{\sigma}^2_{\text{GLS}} = \frac{1}{N - p} F_{\text{GLS}}(\hat{\theta}_{\text{GLS}}). \tag{10}$$

Given these values, an asymptotic theory similar to that for the OLS estimator can be invoked to quantify uncertainty in the GLS estimate using standard errors and confidence intervals (again see [3, 15] for details).

Rewriting Equation (7), we have $\mathcal{E}_{ij} = (N^d_{ij} - n(t_i, z_j; \vec{\theta}_0))/n(t_i, z_j; \vec{\theta}_0)$. Thus the residuals

$$\tilde{r}_{ij} = (n^d_{ij} - n(t_i, z_j; \hat{\theta}_{\text{GLS}}))/n(t_i, z_j; \hat{\theta}_{\text{GLS}}) \tag{11}$$

are a realization of the error in the data and should have constant variance. When GLS estimation is used with data in which the noise is proportional to the magnitude of the observations, and hence data for which this statistical model is appropriate, these residuals should appear random when plotted as a function of the model values $n(t_i, z_j; \hat{\theta}_{\text{GLS}})$. Figure 3 depicts a typical example of these residuals for data with variance proportional to the magnitude of the observation.

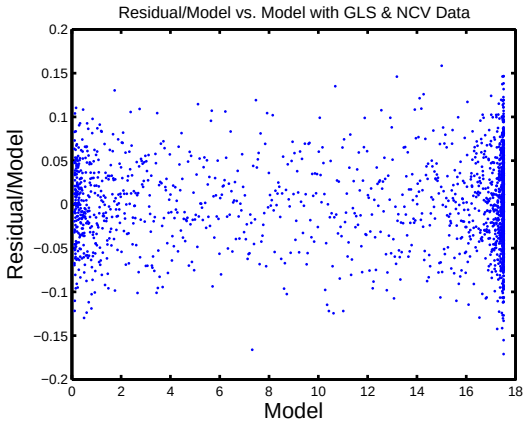

Figure 3: Modified residuals for observations with nonconstant variance.

2.3.3 Model Comparison Test

When forming a mathematical model in an effort to describe a physical, biological or sociological process, there is often the possibility of adding an additional term and/or mechanism into the model in an effort to better fit the data. In the case of such a refinement, the resulting cost function (OLS or GLS) will at least remain unchanged but will likely decrease because of the additional degrees of freedom (in essence, minimizing over a less constrained set). However, the question must be asked whether the resulting decrease in the cost function reflects a significant improvement in the model fit to data, beyond a simple increase in degrees of freedom, to warrant the addition of that mechanism into the model. Alternatively, it may be sufficient to employ the less sophisticated mathematical model with fewer parameters. In this section mathematical and statistical tools are used to help answer these questions are summarized.

In the statistical models discussed above it is assumed that the true parameter $\vec{\theta}_0$ is contained in some set Θ of admissible parameters. Following the notation of [15], consider the constrained parameter space Θ_H, a subset of the parameter space of the more complex model Θ, defined by

$$\Theta_H = \{\vec{\theta} \in \Theta | H\theta = d\} \tag{12}$$

where H is an $r \times p$ matrix with full rank and d is an r-dimensional vector. The goal is to develop a statistical test of the null hypothesis, H_0: $\vec{\theta}_0 \in \Theta_H$. Let $F(\vec{\theta})$ be the cost function (OLS or GLS) associated with a given model and data set and define

$$\hat{\theta}_H = \arg\min_{\vec{\theta} \in \Theta_H} F(\vec{\theta}), \qquad \hat{\theta} = \arg\min_{\vec{\theta} \in \Theta} F(\vec{\theta}).$$

Then it is possible to define the test statistic

$$U_N = \frac{N[F(\hat{\theta}_H) - F(\hat{\theta})]}{F(\hat{\theta})}, \tag{13}$$

where N is the total number of data points. It follows from an asymptotic theory (see [5, 6, 10, 15] and references therein), different from those mentioned above for standard errors and confidence intervals, that, if H_0 is true, U_N converges in distribution (as the number of data points go to infinity) to a χ^2 distribution with r degrees of freedom. Moreover, if $\eta = P(U > U_N)$ where $U \sim \chi^2(r)$, then we may reject H_0 with confidence $(1 - \eta)100\%$.

For example: suppose in the parameter set shown in Table 1 that four nodes are being used in the estimation of the death rate function $\beta(z)$, $b_1, \ldots, b_4$. For simplicity, it will also be assumed that these are the only four parameters being estimated; that is c, $\{a_i\}$, and γ are fixed. It may be of interest whether or not this death rate could be treated as a constant function. Then the restricted set of admissible parameters Θ_H is given by (12) where

$$H = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \qquad d = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Note in this example that $r = 3$ for the χ^2 test. Such examples along with others are discussed in [10, 15].

3 Model Modifications

We first discuss the best fit we can obtain with the original model (given in Equation (2)). The same data set was fit in [40], but exact values were not given for the proliferation and death rates (although they were depicted graphically). Also, we did not perform Tikhonov regularization, and therefore, our problem is not convexified to ensure only one minimum for a modified objective functional. Instead, we begin with the best fit parameters from [40], with the proliferation and death rates estimated to the best of our ability, and proceeded to obtain an OLS estimate of the parameters. In Section 3.2, we propose modifications to the model (2), and accept or reject these changes based on whether they result in a statistically significant improvement of the agreement between model solutions and experimental data.

3.1 Data fit with the original model

First the original model (2) is fit to data. The resulting discrepancies between the observed data and the best fit model solution are then used to provide motivation for improvements to the model. For the current discussion, focus will remain on the OLS estimation of the parameters. The alternative use of the GLS procedure is discussed in Section 4.

The death rate of cells is expected to vary little after the initial rounds of divisions. Thus only four nodes are used in the estimation of the death rate function. In the region $z \in [0, 2.5]$, the death rate is treated as a constant, $\beta(z) = b_1$. The proliferation rate function, on the other hand, should vary with division number and more nodes are needed (particularly in regions of rapid division). The placement of the nodes for the proliferation and death rate functions is given in Tables 2 and 3, respectively. These nodes were chosen after considerable trial and error running both forward simulations and inverse problem parameter estimations. Nodes chosen too close together cause instabilities due to over-discretization, while nodes chosen too far apart do not provide sufficient information regarding the behavior of the population. After OLS minimization, the best fit parameter estimate $\hat{\theta}_{\text{OLS}}$ had a cost $F_{\text{OLS}}(\hat{\theta}_{\text{OLS}}) = 3.2112 \times 10^{12}$. The best fit proliferation and death rate functions are depicted in Figure 4. The best fit values for the label loss rate and label dilution factor were $\hat{c} = 0.004421$ and $\hat{\gamma} = 1.5751$. The model solution evaluated at the best fit parameters is shown in comparison to the data in Figures 5 and 6.

z_i	a_i
1.2500	0.0020
1.5000	0.0112
1.6250	0.0169
1.7500	0.0161
1.8750	0.0091
2.0000	0.0222
2.1250	0.0015
2.2500	0.0505
2.3750	0.0117
2.5000	0.0027
2.6250	0.0231
2.7500	0.0016
2.8750	0.0076
3.0000	0.0076

Table 2: Nodes and estimated values (OLS) for the proliferation rate function in the original formulation using Equation (2) .

z_i	b_i
2.5000	0.0085
2.7500	0.0248
3.0000	0.0000
3.5000	0.0000

Table 3: Nodes and estimated values (OLS) for the death rate function in the original formulation using equation (2).

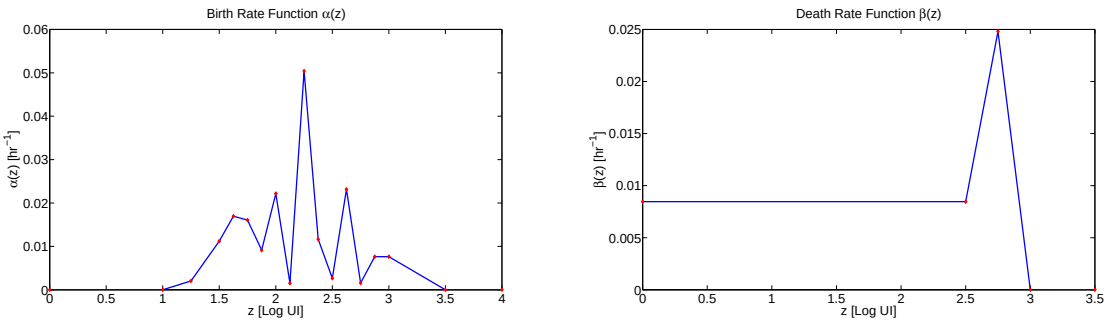

Figure 4: Graphical presentation of estimated (OLS) birth and death rate functions $\alpha(z)$ (left) and $\beta(z)$ (right).

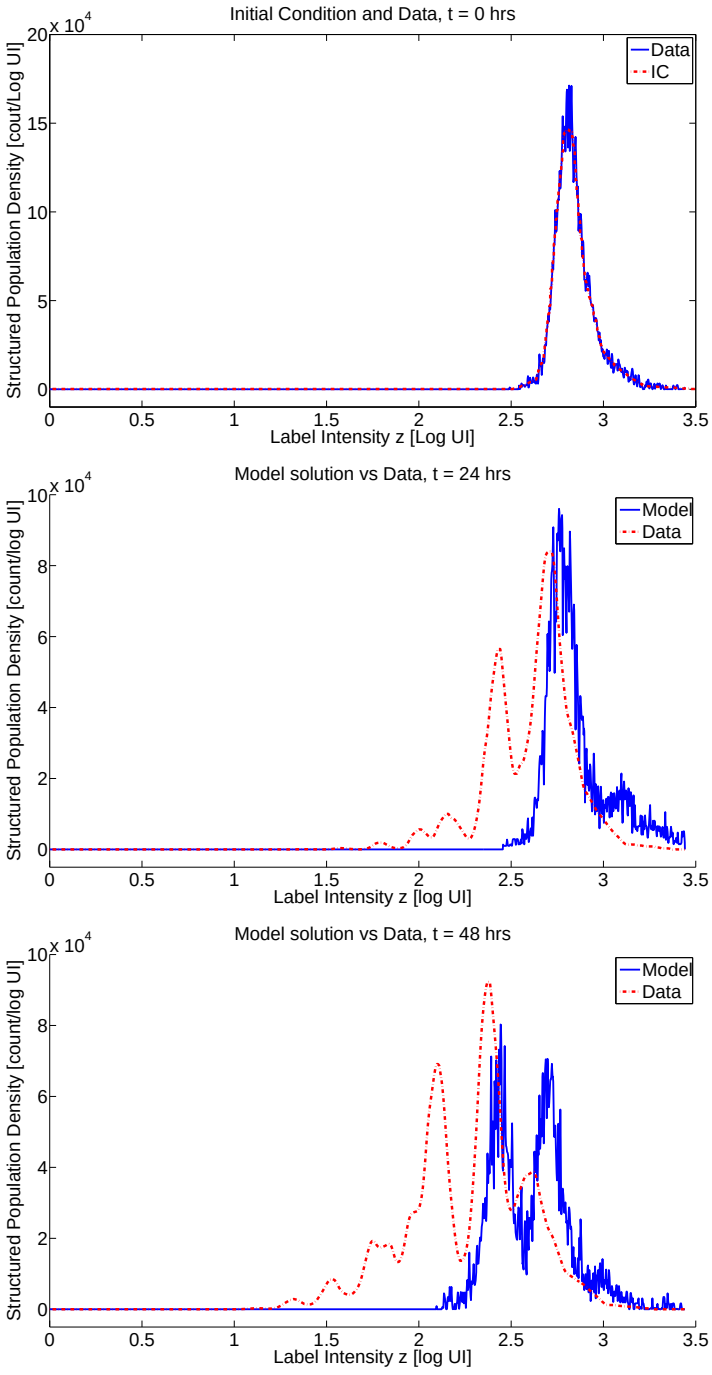

Figure 5: OLS best fit model solution to original PDE formulation with equation (2) in comparison to the data: t=0, 24, 48 hrs

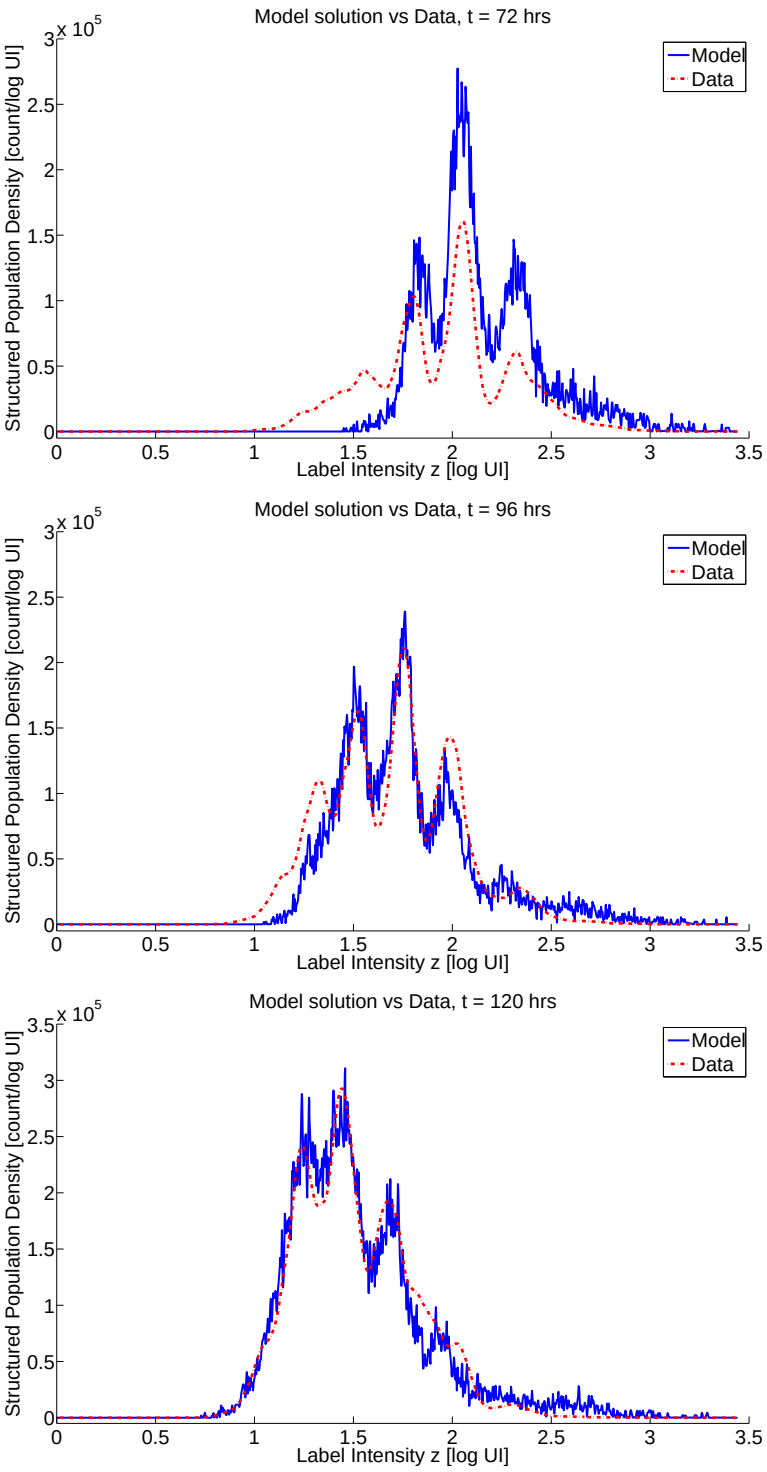

Figure 6: OLS best fit model solution to original PDE formulation with equation (2) in comparison to the data: t=72, 96, 120 hrs

3.2 Modified Model

It is clear from Figures 5 and 6 that the fit of the model to the data can be improved. It appears that the model is not quite capturing the dynamics of the cells in the proliferation assay. Moreover, the label dilution factor γ must be less than 2 in order to accurately fit the data. If we were to interpret the measurement of fluorescence as a *mass* measurement, or a measurement of total quantity (assumption ii.)), then in the model (2), this definition would require $\gamma = 2$. Values of $\gamma < 2$ would then imply a creation of label during proliferation, or that the *amount* of CFSE in the mother cell is less than twice that of the daughter cell. Clearly, this is biologically infeasible and physically implausible so the interpretation is incorrect and *FI is not a mass measurement*. However, if we interpret the fluorescence intensity as a *concentration* measurement (as discussed in [41, 43, 52]), and the same amount of CFSE in a larger volume fluoresces at a lower intensity, then a feasible range for γ is $[1, 2]$. The reason this parameter is not exactly $\gamma = 1$ is that while the concentration immediately after division of the daughter cells is identical to that of the original mother cell, a marked growth in cell volume immediately following division results in a dilution of the CFSE, and daughter cells are detected at a lower fluorescence intensity. However, the daughter cells do not fluoresce at half the original intensity (represented by $\gamma = 2$) until they reach the size of the mother cell, which occurs much later as the cell progresses through the non-mitotic phases of the cell cycle. While not actually a change in the mathematical model, this change in *interpretation* of the parameter γ provides for a more satisfactory discussion of the results of the model. The parameter γ *effectively* represents cellular mechanisms governing the timing of initial growth of daughter cells after division. These underlying dynamics likely occur on a much faster scale than the other observed processes in this model and have essentially been integrated over in time and are absorbed in the parameter γ. Thus, it is not likely that its value will be estimated in this type of experimental setting as the precise value $\gamma = 1$ which would reflect the strict biological interpretation as the ratio of mother to daughter label concentration.

Because of the natural loss of FI over time due to catabolic activity, assumption i.) is not to be taken in an exact sense – histograms of cells as a function of FI can be used to see division as it occurs, but no simple relationship between FI and division number exists. While the cells fluorescing at the intensity of the peaks seen in the data have all likely undergone the same number of divisions, cells of fluorescence intensity in the valleys in the data may have undergone different numbers of divisions since the start of the experiment. The natural label loss causes cells to slowly drift to the left on the FI scale. For that reason, no particular region of FI can be definitively linked to a particular division number in Figure 1. However, we can define a new variable $s = z + ct$ in terms of which we can more intuitively define the proliferation and death rates. As we shall see, this time dependent translation of the domain can be seen (in Figure 7 below) to result in a data plot which does, in fact, correlate well with division number. (To see this, compare Figure 7 below to Figure 1.) To better capture this in our model, we replace the rates $\alpha(z), \beta(z)$ in our model by new translated *effective* proliferation and death rates $\alpha(s) = \alpha(z + ct), \beta(s) = \beta(z + ct)$, respectively. It should be noted that the label loss rate function $v(z)$ does not undergo a similar translation/replacement because of the assumption that label loss rate depends on CFSE label intensity or FI, but not division number.

In an attempt to interpret these *effective* rates of proliferation and death introduced in the efforts here, we recall for readers that, while not common in the biological sciences, it is altogether common in the physical sciences and engineering to consider velocities (i.e., rates of change) relative to different coordinate or reference frames. For example, in mechanics and

motion of continua (elasticity and fluids) and deformable bodies [11, 25, 26, 44, 46], it is frequent to encounter velocities relative to a *fixed* coordinate system (in a *Lagrangian* formulation) or relative to a *moving* coordinate system (in an *Eulerian* formulation). More precisely, when analyzing the deformation or motion of solids, or the flow of fluids, it is necessary to describe the sequence or evolution of configurations throughout time. One description for motion is made in terms of the material or fixed referential coordinates, and is called a material description or the *Lagrangian description*–also called an *initial/referential, material, undeformed, or fixed* frame formulation. In this formulation, an observer standing in the fixed referential frame observes the changes in the position and physical properties as the material body moves in space as time progresses. In other words, this formulation focuses on individual particles as they move through space and time. The other description for motion is made in terms of the spatial or current coordinates, called a spatial description or *Eulerian description*– also called a *current/present, space, deformed, or moving* frame formulation. In this approach one focuses on the current configuration of the body, giving attention to what is occurring at a moving material point in space as time progresses. That is, the coordinate system is relative to a moving point in the body and hence is a *moving coordinate system*. An intuitive comparison of these two descriptions would be that in the Eulerian description one places the coordinate or reference system for motion of an object *on* the object as it moves through a moving fluid (e.g., on a boat in a river) while in the Lagrangian description one observes and describes the motion of the object from a fixed vantage point (e.g., motion of the boat from a fixed point on a bridge over the river or on the side of the river.).

Finally, biological evidence suggests that the proliferation rate might also depend in some way on time [31] in addition to division number, as the time to first division is clearly seen to differ from that of subsequent divisions. The assumption that the death rate depends only on division number is supported in literature [31, 38, 47], and is not investigated here. The consideration of time dependence in the proliferation rate as introduced here will later be tested (with positive affirmation) for statistical significance, which is one way of determining whether the effect might be present in the experimental data. This is also seen in the data set. At $t = 24$ hours, all cells are still in the original (undivided) generation, indicating a proliferation rate of zero in the first 24 hrs. However, these cells do divide, producing the additional peaks seen at later times. Thus it must be that α (at least as represented in the data) changes in time. In a given longitudinal in vitro data set such as we are considering here, the apparent proliferation rate does depend on time in some specific way related both to division times and label loss. Indeed, in defining the *effective* proliferation and death rates, we introduce time implicitly into the rates. Thus, permitting explicit time dependence in the effective proliferation rate α is a rather natural modification to also consider.

Taking all of these considerations into account, we modify the model (2) to obtain

$$\begin{aligned} \frac{\partial n(t,z)}{\partial t} &+ \frac{\partial [v(z)n(t,z)]}{\partial z} = -(\alpha(t, z+ct) + \beta(z+ct))n(t,z) \\ &+ \chi_{[z_{\min}, z_{\max} - \log\gamma]} 2\gamma\alpha(t, z+ct+\log\gamma)n(t, z+\log\gamma). \end{aligned} \tag{14}$$

Again, the ultimate goal is to estimate the functions $\alpha(t,s)$, $\beta(s)$, and $v(z)$ and the parameter γ such that the model best fits a given set of data. Because the death rate $\beta(s)$ is expected to be relatively constant after the initial rounds of division, it will be treated as a constant, $\beta(s) = b_1$ for all $s \in [0, 2.5]$. For $s \in [2.5, 4]$, $\beta(s)$ will be constructed as a piecewise linear

function,

$$\beta(s) = \sum_{i=1}^{M_\beta} b_i \phi_i(s), \tag{15}$$

where $\phi_i(s)$ are piecewise linear splines defined as before. The proliferation rate function $\alpha(t, s)$ will be represented by linear combinations of products of one dimensional piecewise linear splines $\xi_j,\ \psi_i$, i.e.,

$$\alpha(t, s) = \sum_{i=1}^{M_\alpha^t} \sum_{j=1}^{M_\alpha^z} a_{ij} \xi_j(t) \psi_i(s). \tag{16}$$

As before, $a_{ij} = 0$ or $b_i = 0$ indicates zero birth or death respectively while $a_{ij} = 1$ or $b_i = 1$ indicates that cells at the given location in time and/or translated log are dividing or dying once per hour. This is clearly much faster than the actual or true value, so the interval $[0, 1]$ should cover the true value for each a_{ij} and b_i. To reduce the total number of parameters, $\alpha(t, s)$ will be set to zero for $s \in [0, 1]$. No cells enter this region and hence the birth rate is arbitrary there. The s-nodes are not evenly spaced (see Table 4) but were chosen after considerable trials with both forward simulations and parameter estimations. As a general rule, the s-nodes need to be closely spaced in order to accurately model the data. Choosing too few nodes results in a poor estimation of the proliferation rate, often with the result of additional generations of cells appearing too early in the model solution. Choosing too many nodes leads to over-discretization and the additional need for some type of regularization. Here we employ so-called regularization by discretization as described in [9, 10]. The time nodes are evenly spaced every 12 hours in the region $t \in [36, 120]$. For $t \in [0, 24]$, $\alpha(t, s)$ is set to zero as it is clear from the data that no proliferation occurs during this time.

Also as before, it will be assumed that the rate of label loss is proportional to total label concentration, resulting in $v(z) = -c$ in the model (14). This form was found in [40] to provide a better fit to this data set as compared to constant label loss. Forward simulations indicate $c \in [0.0025, 0.0055]$ to be a reasonable range. The ranges of values for the parameters to be estimated in the modified model (14) are then no different from those given in Table 1, for estimating parameters in the original model (2).

We used the modified model (14)–(16) in an OLS procedure with the data. The corresponding OLS estimation results in $\hat{c} = 0.003288$ and $\hat{\gamma} = 1.5169$ with a cost $F_{\text{OLS}}(\hat{\theta}_{\text{OLS}}) = 5.3181 \times 10^{11}$. Using this value of c in the translated log intensity coordinate $s = z + ct$, we plotted the experimental data relative to this coordinate. As seen in Figure 7, this results in a data plot that correlates remarkably well with division number. The corresponding estimated effective proliferation rate function $\hat{\alpha}(t, s)$ is depicted graphically in Figure 8 with node values given in Table 4. Similarly, the estimated effective death rate function $\hat{\beta}(s)$ is shown in Figure 8 with nodal values given in Table 5. The model solution evaluated at the best-fit parameter vector is compared to the data in Figures 9 and 10.

The improvement of the fit to the data is substantial both visually and in terms of lowering the OLS cost function value. The assumption of time dependence for the birth rate function appears justified not only biologically but also by the model/data fits. To verify that this is not merely due to an increase in the number of degrees of freedom, we use the model comparison statistic to test if the reduction in residual sum of squares is statistically significant. When the original model (of Equation (2) with α not time dependent and only 20 unknown parameters: $c, \gamma, 4$ nodes for $\beta(z), 14$ nodes for $\alpha(z)$) is only modified by allowing for time dependence in the proliferation rate $\alpha(t, z)$ with now 102 unknown parameters, the statistic corresponding to

$s_j \backslash t_i$	36	48	60	72	84	96	108	120
1.2500	0.0000 *	0.0000 *	0.0000 *	0.0000 *	0.2023	0.0158	0.0004	0.0002
1.5000	0.0000 *	0.0000 *	0.0000 *	0.0000 *	0.0187	0.0020	0.0014	0.0159
1.6250	0.0000 *	0.0000 *	0.7755	0.0045	0.0000	0.0022	0.0152	0.0226
1.7500	0.0000 *	0.0000 *	0.0468	0.1303	0.0130	0.0179	0.0342	0.0684
1.8750	0.0000 *	0.0000 *	0.0309	0.0231	0.0496	0.0143	0.0079	0.0796
2.0000	0.0000 *	0.0000 *	0.0000	0.0000	0.0435	0.0588	0.0415	0.1467
2.1250	0.1103	0.4173	0.0876	0.0000	0.1019	0.0001	0.0447	0.0715
2.2500	0.0082	0.0001	0.0014	0.0390	0.0001	0.1491	0.1954	0.2255
2.3750	0.1354	0.0002	0.1775	0.0055	0.0806	0.0000	0.0868	0.1318
2.5000	0.0111	0.0003	0.0020	0.2003	0.0000	0.2007	0.1694	0.3725
2.6250	0.0015	0.0148	0.2894	0.0000	0.1658	0.0006	0.0761	0.2036
2.7500	0.0008	0.1159	0.2415	0.3985	0.0011	0.1409	0.4961	0.2110
2.8750	0.0159	0.0013	0.0230	0.1588	0.0000	0.0514	0.0005	0.4284
3.0000	0.0000	0.0378	0.0005	0.2418	0.0064	0.0870	0.0035	0.9661

Table 4: Best fit (OLS) parameters shown along with the s (left column) and t (top row) nodes for the proliferation rate $\alpha(t, s)$. The function $\alpha(t, s)$ is shown graphically in Figure 8. Note that $\alpha(0, s) = \alpha(12, s) = \alpha(24, s) = 0$ and was not estimated. * Parameter was not estimated but was set to zero as there are no cells observed in our data set at these fluorescent levels at the given times.

s_i	b_i
0.0000	0.1003
2.5000	0.1003
2.7500	0.0237
3.0000	0.0000
4.0000	0.0000

Table 5: Best fit (OLS) parameters shown along with the s nodes (left column) for the death rate $\beta(s)$. In the region $s \in [0, 2.5]$, $\beta(s) = b_1 = 0.0665$. The functions $\beta(s)$ is shown graphically in Figure 8.

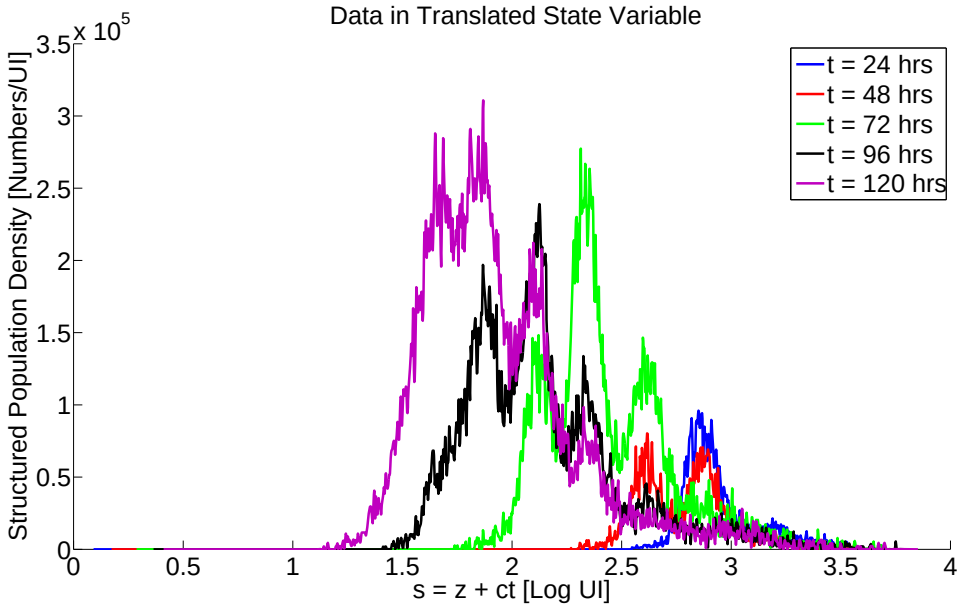

Figure 7: Original data sets shown in translated log intensity $s = z + ct$, with $c = \hat{c} = .0032888$, as estimated with the OLS procedure for the modified model (14). Note that subsequent division peaks are now strongly correlated with specific regions in the state variable, unlike the original log intensity variable z (see Figure 1).

the resulting additional $r = 82$ degrees of freedom is given by

$$U_N = \frac{N(3.2112 \times 10^{12} - 9.8423 \times 10^{11})}{9.8423 \times 10^{11}} = 12110$$

for $N = 5352$. This statistic suggests that it is unlikely at extremely high levels of confidence that this improvement in residual would have occurred by chance or by simply the act of increasing the number of degrees of freedom in the model, and supports the inclusion of explicit time dependence in the proliferation rates. We remark that in any such efforts with nonconstant parameters, one could (as is often done in general parametric estimation) substantially reduce the number of parameters to be estimated by employing a distributional form (a reduced order parameter shape or representation) instead of the general parameter representations as in (15)–(16). However, just as in using a Maximum Likelihood Estimator (where assumption of a distributional representation for measurement error is required) or in general Bayesian approaches to estimation, one often loses information by an incorrect assumption regarding an a priori form for the distribution or shape function being estimated. In any case, this part of the model analysis strongly suggests that explicit time dependence in proliferation rates is important in accurately representing and understanding such data sets.

The translation of the intensity variable, in addition to its justification by the experimental setup, provides greater insight into how the proliferation rate varies both in time and with subsequent generations of cells. This modification does not change the number of parameters so we can directly look to the value of the cost functional, which is lower at $F_{\text{OLS}}(\hat{\theta}_{OLS}) = 5.3181 \times 10^{11}$. As a means of reference, the model solution in the translated intensity coordinate is graphed in Figure 11. Note that the subsequent generation peaks align very closely. Compare also the data in the translated coordinate (Figure 7).

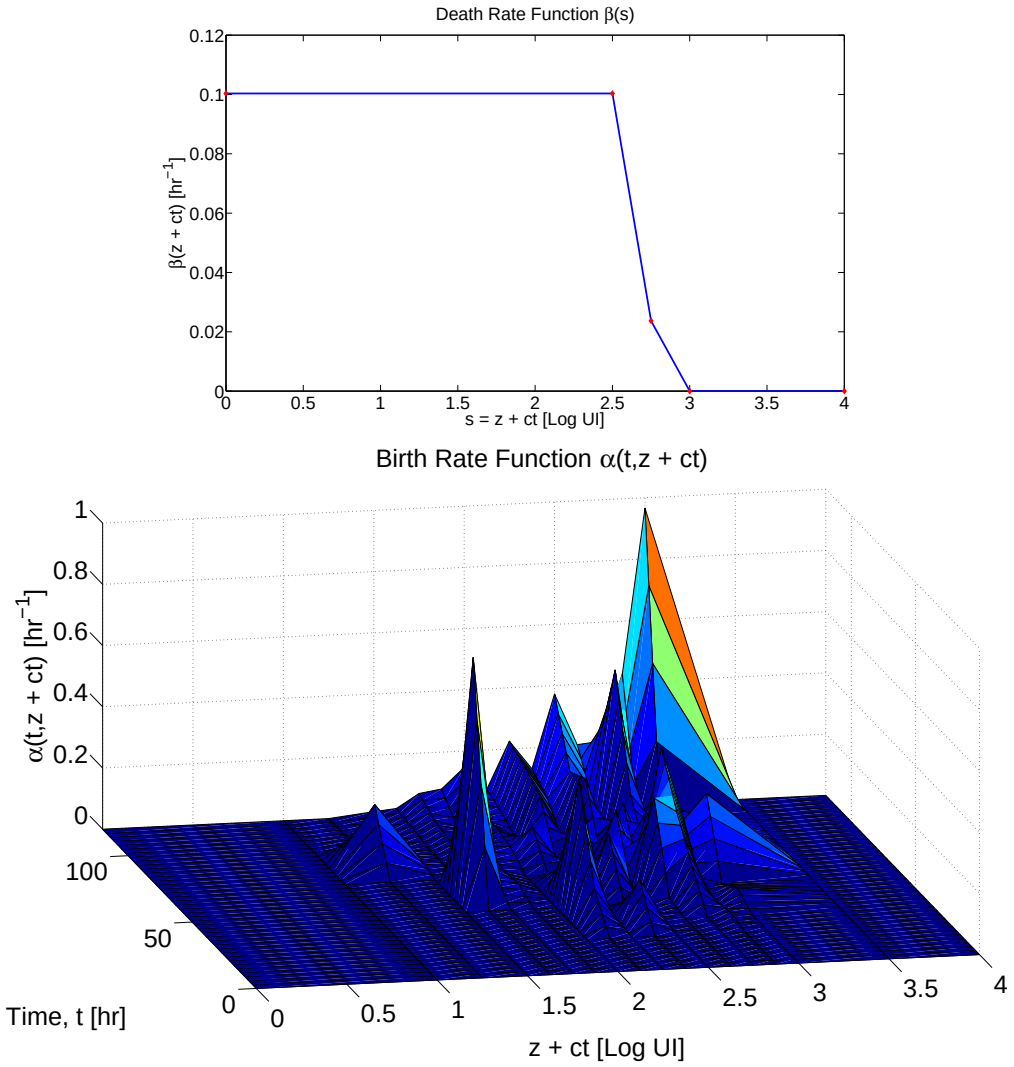

Figure 8: Graphical representation of the best fit (OLS) death rate $\beta(s)$ (top) and proliferation rate $\alpha(t, s)$ (bottom). Numerical values are given in Tables 4 and 5.

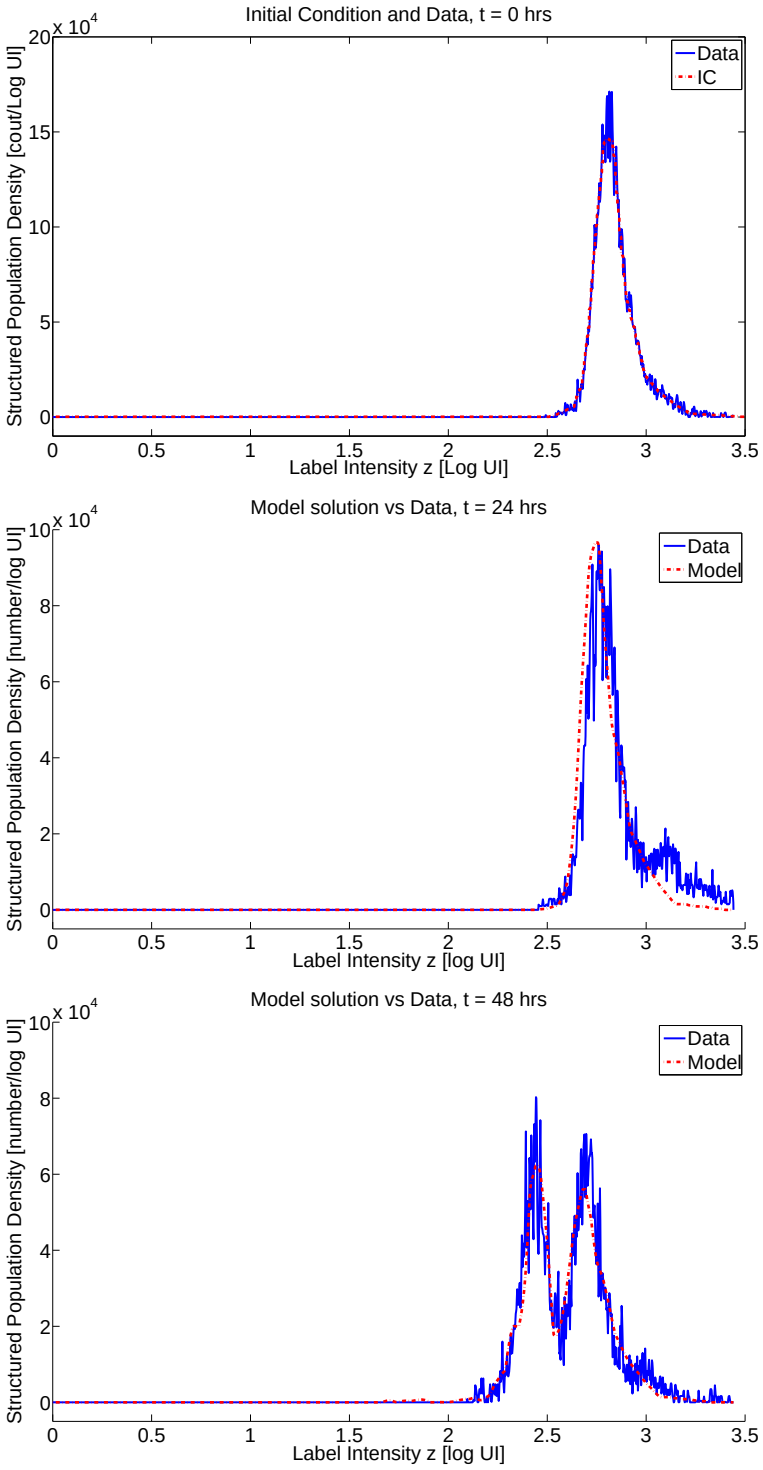

Figure 9: Improved model solution evaluated at the best fit (OLS) parameters in comparison to the original data: t=0, 24, 48 hrs

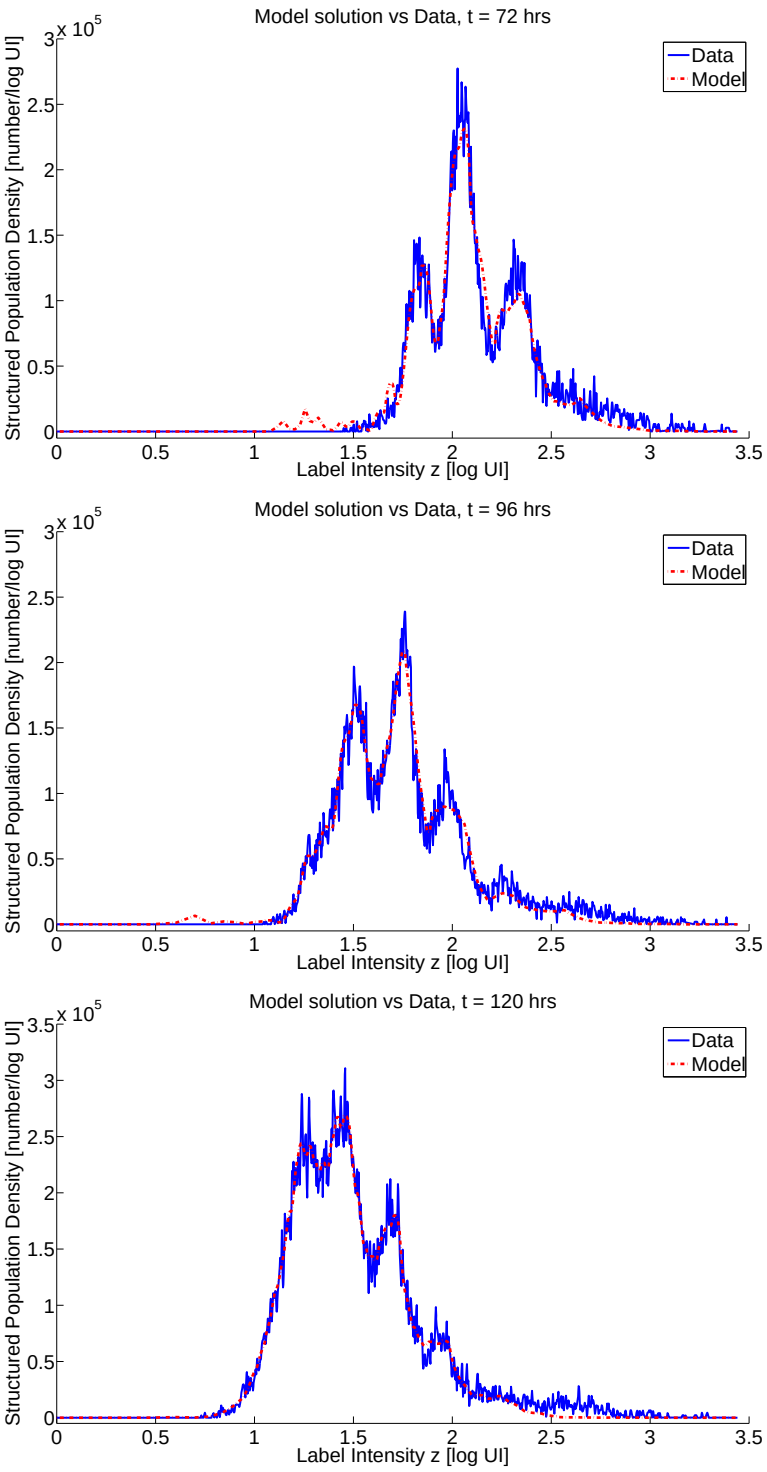

Figure 10: Improved model solution evaluated at the best fit (OLS) parameters in comparison to the original data: t=72, 96, 120 hrs

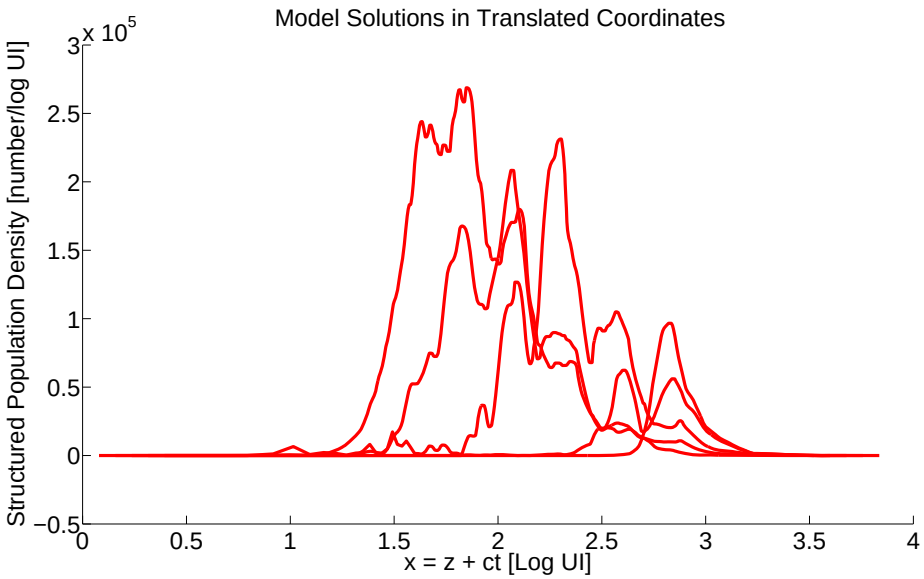

Figure 11: Best fit (OLS) model solution shown in terms of the translated coordinate $s = z+ct$.

There is some evidence [39, 40] that the death rate function might be treated as a constant, $\beta(s) = \beta$, thereby reducing the total number of parameters while still accurately fitting the data. Adding this restriction to the present formulation (14) and running the OLS estimation procedure again, we obtained a best fit cost of $F_{\text{OLS}}(\hat{\theta}_{\text{OLS}}) = 8.2045 \times 10^{11}$. The corresponding test statistic is $U_N = 2905$, for which it may be concluded from the $\chi^2(3)$ distribution with very high confidence ($> 99.999\%$) that $\beta(z)$ cannot be treated as a constant for the current data set.

It has been assumed that label loss is strictly proportional to CFSE concentration ($v(x) = -cx$; $v(z) = -c$), but it may in fact vary in a more complex way based upon the nature of the catabolic activity within the cell. Thus the possibility is considered that the label loss function might have the form $v(z) = v_0 z - c$. Returning to the OLS minimization, we obtained the result $\hat{v}_0 = 7.8921 \times 10^{-5}$ with a corresponding cost $F_{\text{OLS}}(\hat{\theta}_{\text{OLS}}) = 5.3152 \times 10^{11}$. The other parameters in the model remained largely unchanged. The resulting test statistic (for $v_0 = 0$) is $U_N = 2.92$, and the additional term in the form of the label loss rate would only be supported with 91.25% confidence. The fit of this model to the data is not noticeably different from that depicted in Figures 9 and 10.

Considering other possibilities, we note that there is also a large body of work on so-called Growth Rate Distribution (GRD) (or Class Rate Distribution (CRD)) models [2, 4, 7]. Adapted to the present application, these models assume that the population is divided into small groups of individuals which share a common label loss rate (the 'class' rate) within the group. This is in place of assuming the affine term in the above paragraph. The dynamics of the total population are then defined by the probability distribution of label loss rates within the population. Such variability within the overall population may readily be biologically justified by the variety of catabolic mechanisms underlying the label loss. With a slight change of notation, let $n(t, z; c_k)$ be the structured population density of a cohort of cells all of which have label loss rate c_k. We consider only the parameter c as being distributed as the affine term has already been shown to offer no statistically significant improvement. For simplicity,

assume there are a finite number of c_k's and that there is a discrete probability measure P defining their distribution within the total population. Then the total population is given by

$$N(t, z; P) = \sum_k n(t, z; c_k) p_k.$$

The estimation of the probabilities p_k can be reduced to a quadratic programming problem (as described in [4]) which can be quickly and easily solved via MATLAB's `quadprog` routine provided the other parameters of the model are fixed. For the current problem, 21 evenly spaced label loss rates in the region $c \in [0.0015, 0.0055]$ were considered. In the interest of computational efficiency, `fmincon` and `quadprog` were used in an alternating fashion in a hybrid algorithm. First, `fmincon` was used to estimate the other parameters leaving the p_k fixed, and then `quadprog` was used to estimate the p_k holding the other parameters fixed. Doing so results in a final cost of $F_{\text{OLS}}(\hat{\theta}_{\text{OLS}}) = 5.3137 \times 10^{11}$, which gives a model comparison statistic of $U_N = 4.43$. This is not sufficient to warrant the inclusion of this additional complexity when U_N is compared to the critical values for a $\chi^2(20)$ distribution. The estimated distribution of label loss rates within the population is depicted in Figure 12. The modifications made from the model in Equation (14) and the corresponding statistics are summarized in Table 6.

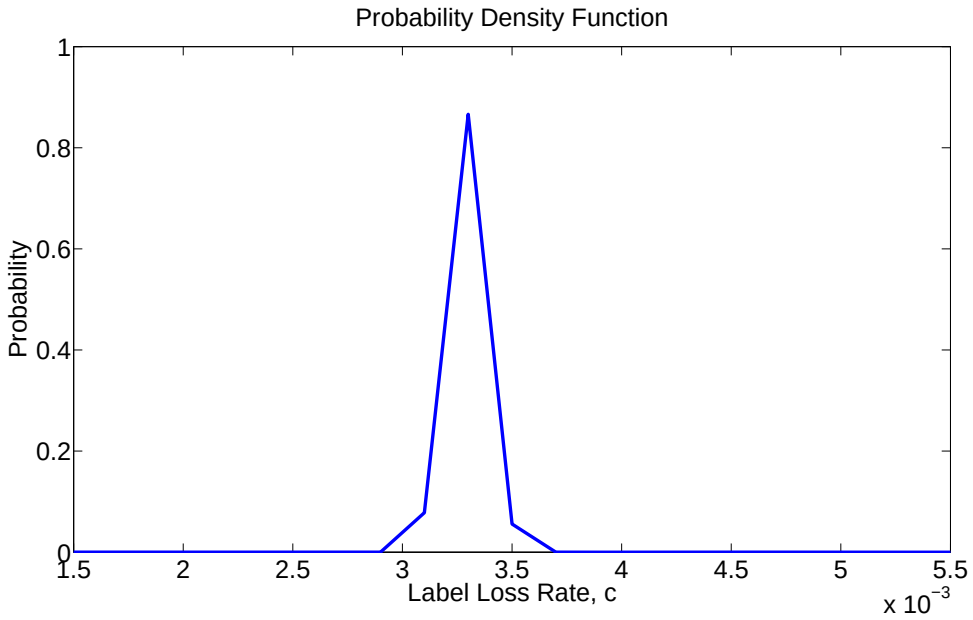

Figure 12: Estimated probability distribution of label loss rates c within the population.

Together, these improvements result in a final model of the form

$$\begin{aligned} \frac{\partial n(t,z)}{\partial t} &+ \frac{\partial[-cn(t,z)]}{\partial z} = -(\alpha(t, z+ct) + \beta(z+ct))n(t,z) \\ &+ \chi_{[z_{\min}, z_{\max} - \log\gamma]} 2\gamma\alpha(t, z+ct+\log\gamma) n(t, z+\log\gamma). \end{aligned} \tag{17}$$

It is worth noting that this particular model, at least at the present, has only been modified to fit the particular data set shown. It is certainly possible that there are biomechanisms which are not manifested in the available data set and hence are not incorporated into this model. As this mathematical model is applied to different cell types and proliferation assays, it is expected

	β	$v(z) = v_0 z - c$	$v(z) \sim P$
$F_{\text{OLS}}(\hat{\theta}_{OLS})$	8.2045×10^{11}	5.3152×10^{11}	5.3137×10^{11}
U_N	2905	2.92	4.43
r	3	1	20
improvement	yes	no	no

Table 6: Model comparison statistic for the modifications in the table: constant death rate β, affine label loss $v(z) = v_0 z - c$, and distributed label loss $v(z) = c \sim P$. These are all compared to the version of the model in Equation (14) with $\alpha(t, s)$, $\beta(s)$, $v(z) = -c$, with a cost functional of $F_{\text{OLS}}(\hat{\theta}_{OLS}) = 5.3181 \times 10^{11}$. Note the row labeled r contains the degrees of freedom for the χ^2 distribution that the model comparison statistic U_N is to be compared. The bottom row depicts whether or not the more complex model is supported by the statistic.

that the model may take on slightly different forms. Still, the techniques used in obtaining this improved model as well as the overall form of the model lay a solid foundation for future work. Given the accuracy of the current model in replicating the experimental data, it is now reasonable to turn to a discussion of the validity of the statistical assumptions underlying the OLS minimization procedure.

4 Estimation Improvements

To this point, our focus has remained on the OLS formulation of the inverse problem. Given the best fit model (17) and the best fit OLS estimate $\hat{\theta}_{\text{OLS}}$, the resulting residuals r_{ij} versus the model values are plotted in Figure 13.

On one hand, these residuals certainly do not exhibit the fan-like structure characteristic of data sets in which noise is proportional to the magnitude of the observation (see Figure 2). However, they do not appear to be random either, slowly growing in variation as the model values increase. Thus, it appears that the assumption of constant variance may not be particularly accurate, and perhaps another assumed error structure should be investigated, resulting in a more general least squares estimation approach, distinct from the two somewhat standard formulations used here.

The results (e.g., model fits, estimated parameter values) of the GLS procedure under the assumption of *relative error* (i.e., statistical model (7)) that we obtained are sufficiently similar to the results presented in the previous section so that separate graphics and tables are not included here. What is of interest for the current discussion are the residuals $\tilde{r}_{ij}$, shown in Figure 14. It is clear that the GLS residuals are not random but slowly decay in time. The conclusion then, is that the underlying statistical model for the variation in our data may have neither CV noise nor noise which is proportional to the magnitude of the observation. This is not surprising given the complexity of the error in the observation or measurement process (see [52] for a discussion of the general analysis of error in data collection procedures). We remark that we do not use log likelihood estimation or error quantification methodology here because such an approach explicitly (by the form of the likelihood function employed) implies that we *know* a priori the distribution for measurement error in the underlying statistical model (e.g., see the discussions in [3]).

In order to accurately and correctly compute standard errors and confidence intervals, the assumptions of the underlying statistical model must be reasonably correct (for the asymptotic

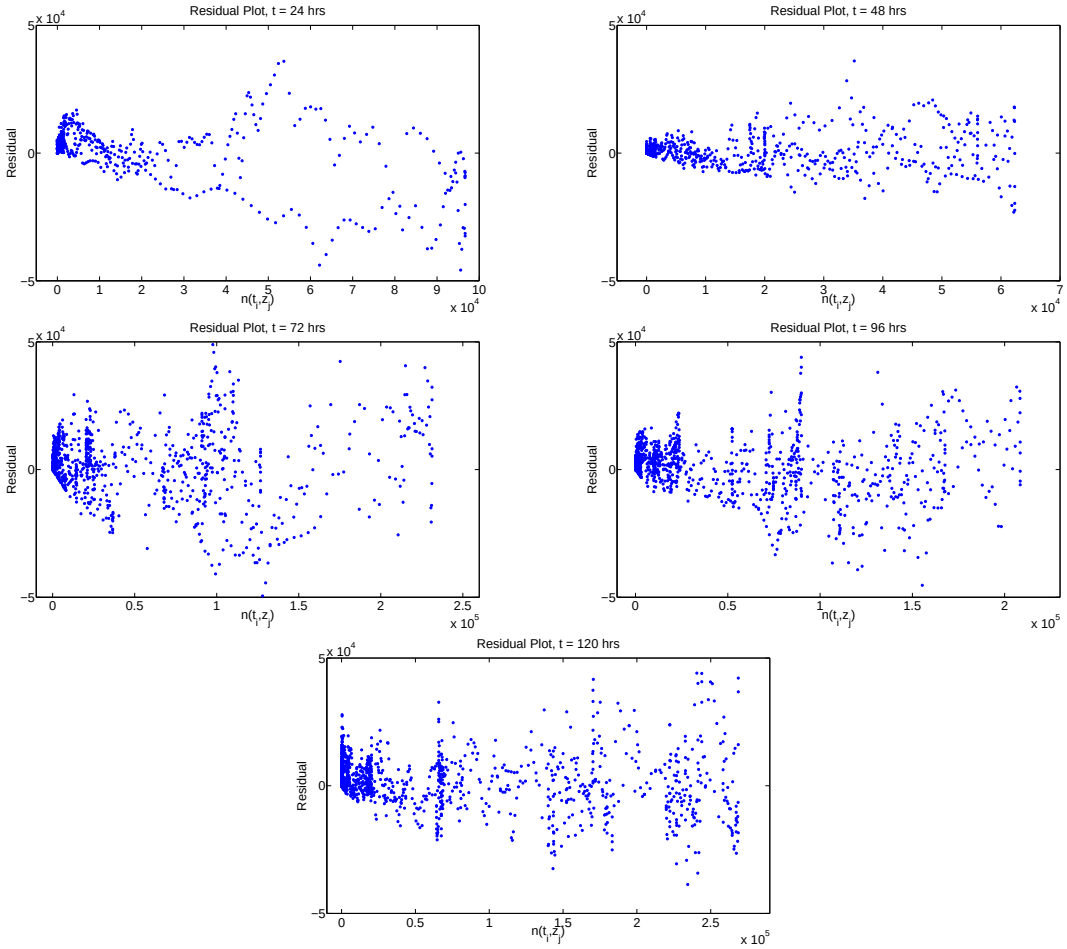

Figure 13: OLS residuals as a function of model value for each time measurement.

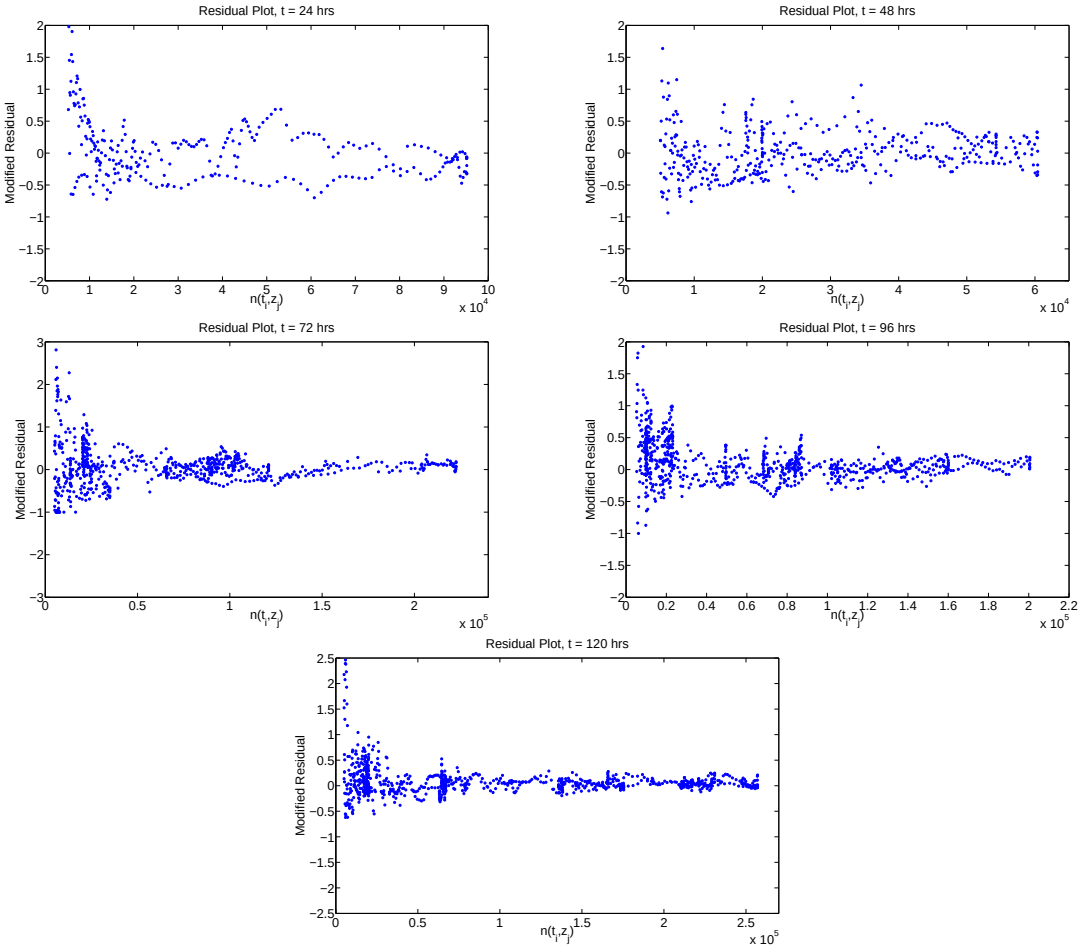

Figure 14: GLS residuals as a function of model value for each time measurement.

formulae to be meaningful–see [3]). However, it appears that the error is not represented well by the statistical models in either the OLS (with constant variance error) or GLS (with proportional error) formulations. Therefore, estimations of the reliability of the parameter estimates (e.g., standard errors and confidence intervals) would be invalid and are not pursued further in this work. One could, of course, use bootstrapping to compute standard errors, but that is extremely computationally expensive for our problem and also involves some underlying assumptions [8], perhaps unsatisfied for our problem. Indeed, the error structure of the observed data may take several different forms. One possibility is that the noise may be proportional to some power λ of the observations

$$N_{ij}^d = n(t_i, z_j; \vec{\theta}_0) + n^\lambda(t_i, z_j; \vec{\theta}_0)\mathcal{E}_{ij},$$

where λ is now an additional parameter to be estimated (see [21, 23] for discussions). More generally, the noise could depend on some general function g of $n(t_i, z_j; \vec{\theta}_0)$:

$$N_{ij}^d = n(t_i, z_j; \vec{\theta}_0) + g(t_i, z_j; \vec{\theta}_0, n(t_i, z_j; \vec{\theta}_0))\mathcal{E}_{ij}.$$

The determination of the parameter λ or the function g represents a difficult computational challenge and is a topic for future work that will require more careful analysis of the data collection process in the context of multiple data sets. Once ascertained and verified by the appropriate residual plots, the statistical model could then be used to quantify uncertainty in the estimates of the parameters of the model.

It is worth one further remark to note that the inaccuracies in the underlying statistical models do not invalidate the entire parameter estimation procedures we have pursued. The various methods of parameter estimation (OLS, GLS, etc.) all give very similar results and the figures of the previous sections demonstrate the accuracy of those results. The determination and validation of the statistical model simply provides an improvement to the estimation procedure that would permit the accurate quantification of uncertainty in the resulting estimated parameters.

5 Concluding Remarks

The use of CFSE-based proliferation assays has been and will continue to be a powerful technique for monitoring a dividing cell population. Coupled with the high-throughput capacity of FACS, there are near limitless possibilities for the use of this technology to track cell divisions and division dependent changes. In this report, a mathematical model governing the population density of a population of proliferating CD4+ lymphocytes is proposed and its remarkable agreement with a FACS data set is demonstrated. This approach provides an alternative to current modeling efforts and does not require prescribed distributions on CFSE or other birth and death events. It also allows for direct comparison of solutions to data without preprocessing.

The current model demonstrates that the effective rate of proliferation within a population of cells depends not only on the number of divisions undergone, but also on the elapsed time since stimulation, as supported by a model comparison statistic. In addition to quantifying this effect, the estimation of the proliferation and death rate functions defined on a translated domain (which correlates more closely with division number than CFSE intensity alone at the current time) permits a better quantitative understanding of how these rates change with division number.

The effective rates of proliferation $\alpha(t, z+ct)$ and death $\beta(z+ct)$ introduced in our models here can be correctly viewed as rates relative to the moving coordinate system $s = z+ct$ (which corresponds more closely with division number) as compared to rates relative to the more obvious fixed coordinate z of log intensity. As in other fields of science and engineering, it can also be valuable in biological rate estimation to use such quantities to compare, characterize or delineate cell populations with respect to their normality or lack thereof (as in diseased, infected, etc., cell populations). Specifically, if one can effectively use inverse problem techniques to reliably estimate these effective or relative rates from CFSE data on cell populations, this could lead to a significant new disease identification procedure.

Analysis of the statistical models underlying the noise in the data reveals that the noise in the data may have neither constant variance nor variance proportional to the magnitude of the observation itself. While this does not invalidate the results presented, it does suggest a possible direction for future computational work which will aid in the accurate quantification of the uncertainty associated with the estimated parameters in such problems.

While the current report focuses only on cell proliferation and death rates in terms of the number of divisions undergone, other division dependent properties (cell surface marker exhibition, cytokine content, etc.) can also be measured simultaneously by FACS during a proliferation assay [47]. This, along with the applicability of the technique to a wide variety of cell types, has potentially profound applications in oncology (cancer metastasis and differentiation from normal cells), virology (latent viruses, HIV), and immunology (allergens, tissue grafting), either in the context of an interpretive framework, as a diagnostic tool, or even as part of a control mechanism (see, e.g. [17, 30, 32, 35, 36]).

Appendix: Derivation of Model

The original PDE model proposed in [40] is a case of the Bell-Anderson model [16] for populations which divide by fission. It was published, like the Sinko-Streifer model [50], in 1967 and arrives at the same general equation form. Let $n(t, x)$ be the population density of a labeled population with structure variable x, where x represents the fluorescence intensity (FI) of a cell. Then

$$N(t) = \int_{x_0}^{x_1} n(t, x)dx. \tag{18}$$

represents the total number of cells with fluorescence intensity in (x_0, x_1). Here x_0 and x_1 are arbitrary with the exception that $[x_0, x_1] \subset [x_{\min}, x_{\max/\gamma}]$ or $[x_0, x_1] \subset (x_{\max/\gamma}, x_{\max}]$. Let $\Delta x(t, x, \Delta t)$ be the average increase of FI of cells with initial intensity x during the interval $(t, t + \Delta t)$ and assume that Δt is chosen such that $|\Delta x| << x_1 - x_0$, so that the number of cells which move into the region via division and subsequently divide, die, or drift out of the region is negligible. It should be noted that Δx will be non-positive as cells cannot increase in FI. Thus subtraction by Δx actually results in a larger value. While counterintuitive, this definition is maintained in order to harmonize with other structured population models.

Consider the change in $N(t)$ during the time interval $(t, t + \Delta t)$, i.e., the quantity $N(t + \Delta t) - N(t)$. Five possible contributions will be considered:

(i.) Cells of intensity in the interval $[x_1, x_1 - \Delta x(t, x_1, \Delta t)]$, losing FI according to Δx:

$$\int_{x_1}^{x_1 - \Delta x(t, x_1, \Delta t)} n(t, x)dx$$

.

(ii.) Cells of intensity in the interval $[x_0, x_0 - \Delta x(t, x_0, \Delta t)]$, losing FI according to Δx:

$$\int_{x_0}^{x_0-\Delta x(t,x_0,\Delta t)} n(t,x)dx$$

.

(iii.) Cells which would have contributed to $N(t + \Delta t)$ had they not died:

$$\int_t^{t+\Delta t} \int_{x_0-\Delta x(t,x_0,\tau-t)}^{x_1-\Delta x(t,x_1,\tau-t)} \beta(x)n(\tau,x)dxd\tau.$$

(iv.) The disappearance of cells from the region due to proliferation:

$$\int_t^{t+\Delta t} \int_{x_0-\Delta x(t,x_0,\tau-t)}^{x_1-\Delta x(t,x_1,\tau-t)} \alpha(x)n(\tau,x)dxd\tau.$$

(v.) The gain of daughter cells (two of them) in the region as a result of proliferation in the parent region:

$$\chi_{[x_{\min},x_{\max}/\gamma]} 2 \int_t^{t+\Delta t} \int_{\gamma x_0}^{\gamma x_1} \alpha(x)n(\tau,x)dxd\tau.$$

Then the difference $\Delta N(t)$ can be computed by summing the quantities in items (i.) through (iv.) in the following way: $N(t+\Delta t)-N(t) = (i.)-(ii.)-(iii.)-(iv.)+(v.)$. Following the standard procedure of dividing by Δt and letting $\Delta t \to 0$, this gives $\frac{dN}{dt}$ on the left side. For the first term on the right side, if $n(t, x)$ is continuous in t and x (a reasonable assumption), the Mean Value Theorem (MVT) implies that there exists a $\theta \in [x_1, x_1 - \Delta x(t, x_1, \Delta t)]$ such that

$$\int_{x_1}^{x_1-\Delta x(t,x_1,\Delta t)} n(t,x)dx = -\Delta x(t,x_1,\Delta t)n(t,\theta). \tag{19}$$

Assuming Δx is continuous in Δt (that is, there is no instantaneous label loss) and varies smoothly, we have

$$\lim_{\Delta t\to 0} \frac{-\Delta x(t,x_1,\Delta t)}{\Delta t} n(t,\theta)d\theta = -v(x_1)n(t,x_1). \tag{20}$$

where $\frac{dx}{dt} = v(x)$, the rate of FI change of cells of intensity x. Applying the same argument for the second term, we find

$$\int_{x_0}^{x_0-\Delta x(t,x_0,\Delta t)} n(t,x)dx = -v(x_0)n(t,x_0). \tag{21}$$

In the consideration of the third term, define

$$u_\beta(\tau) = \int_{x_0-\Delta x(t,x_0,\tau-t)}^{x_1-\Delta x(t,x_1,\tau-t)} \beta(x)n(\tau,x)dx. \tag{22}$$

Then if $\Delta x(t, x, \Delta t)$ and $\beta(x)n(t, x)$ are continuous functions of their variables, so is $u_\beta(\tau)$ and by the MVT, there exists a $\theta' \in [t, t + \Delta t]$ such that

$$\frac{1}{\Delta t} \int_t^{t+\Delta t} u_\beta(\tau)d\tau = u_\beta(\theta'). \tag{23}$$

Thus it follows that

$$\lim_{\Delta t \to 0} u_\beta(\theta') = u(t) = \int_{x_0}^{x_1} \beta(x) n(t,x) dx, \tag{24}$$

assuming $\Delta x(t,x,0) = 0$ for all t, x (which follows from the previous assertion regarding the smoothness of Δx in Δt). Using a similar argument for the fourth term, we have

$$\lim_{\Delta t \to 0} u_\alpha(\theta') = u_\alpha(t) = \int_{x_0}^{x_1} \alpha(x) n(t,x) dx, \tag{25}$$

where $u_\alpha(\tau)$ has the obvious definition.

For the final term, the same argument along with the change of variables $\xi = x/\gamma$ results in

$$\chi_{[x_{\min}, x_{\max}/\gamma]} 2 \lim_{\Delta t \to 0} u_{\tilde{\alpha}}(\theta') = \chi_{[x_{\min}, x_{\max}/\gamma]} 2\gamma \int_{x_0}^{x_1} \alpha(\gamma x) n(t, \gamma x) dx. \tag{26}$$

Altogether,

$$\begin{aligned} \frac{dN}{dt} &= -v(x_1) n(t,x_1) + v(x_0) n(t,x_0) - \int_{x_0}^{x_1} \beta(x) n(t,x) dx \\ &\quad - \int_{x_0}^{x_1} \alpha(x) n(t,x) dx + \chi_{[x_{min}, x_{max/\gamma}]} 2\gamma \int_{x_0}^{x_1} \alpha(\gamma x) n(t, \gamma x) dx. \end{aligned} \tag{27}$$

On the left side, differentiating $N(t) = \int_{x_0}^{x_1} n(t,x) dx$ with respect to t we find

$$\frac{dN}{dt} = \int_{x_0}^{x_1} \frac{\partial n(t,x)}{\partial t} dx. \tag{28}$$

Finally, by applying the Fundamental Theorem of Calculus to the first two terms of (27), simplifying and rearranging, we have

$$\begin{aligned} &\int_{x_0}^{x_1} \frac{\partial n(t,x)}{\partial t} + \int_{x_0}^{x_1} \frac{\partial (v(x) n(t,x)}{\partial x} = \\ &- \int_{x_0}^{x_1} (\alpha(x) + \beta(x)) n(t,x) + \chi_{[x_{\min}, x_{\max}/\gamma]} 2\gamma \int_{x_0}^{x_1} \alpha(\gamma x) n(t, \gamma x). \end{aligned} \tag{29}$$

Equivalently (because x_0 and x_1 were arbitrary),

$$\begin{aligned} \frac{\partial n(t,x)}{\partial t} &+ \frac{\partial [v(x) n(t,x)]}{\partial x} = \\ &- (\alpha(x) + \beta(x)) n(t,x) + \chi_{[x_{\min}, x_{\max}/\gamma]} 2\gamma \alpha(\gamma x) n(t, \gamma x). \end{aligned} \tag{30}$$

Acknowledgements

This research was supported in part by Grant Number R01AI071915-07 from the National Institute of Allergy and Infectious Diseases and in part by the Air Force Office of Scientific Research under grant number FA9550-09-1-0226. The authors are grateful to Dr. T. Luzyanina and two anonymous referees for constructive comments on these efforts.

References

[1] H. T. Banks, D. M. Bortz and S. E. Holte, Incorporation of variability into the modeling of viral delays in HIV infection dynamics , *Math Biosci.*, **183** (2003), 63–91.

[2] H. T. Banks, L. W. Botsford, F. Kappel and C. Wang, Modeling and estimation in size structured population models, LCDC-CSS Report 87-13, Brown University; *Proceedings 2nd Course on Mathematical Ecology,* (Trieste, December 8-12, 1986) World Press, Singapore, 521-541, 1988.

[3] H. T. Banks, M. Davidian, J. R. Samuels, Jr. and Karyn L. Sutton, An Inverse Problem Statistical Methodology Summary, CRSC-TR08-01, January, 2008; Chapter 11 in *Statistical Estimation Approaches in Epidemiology*, (edited by Gerardo Chowell, Mac Hyman, Nick Hengartner, Luis M.A Bettencourt and Carlos Castillo-Chavez), Springer, Berlin Heidelberg New York, 2009, pp. 249–302.

[4] H. T. Banks and J. L. Davis, A comparison of approximation methods for the estimation of probability distributions on parameters, *Appl. Num. Math.,* **57** (2007), 753–777.

[5] H. T. Banks and B. G. Fitzpatrick, Inverse problems for distributed systems: statistical tests and ANOVA, LCDS/CCS Rep. 88-16, July, 1988, Brown University; *Proc. International Symposium on Math. Approaches to Envir. and Ecol. Problems*, Springer Lecture Note in Biomath., **81** (1989), 262–273.

[6] H. T. Banks and B. G. Fitzpatrick, Statistical methods for model comparison in parameter estimation problems for distributed systems, CAMS Tech. Rep. 89-4, September, 1989, University of Southern California; *J. Math. Biol.*, **28** (1990), 501–527.

[7] H. T. Banks and B. G. Fitzpatrick, Estimation of growth rate distributions in size-structured population models, CAMS Tech. Rep. 90-2, January, 1990, University of Southern California; *Quart. Appl. Math.*, **49** (1991), 215-235.

[8] H. T. Banks, K. Holm and D. Robbins, Standard error computations for uncertainty quantification in inverse problems: Asymptotic theory vs. bootstrapping, CRSC-TR09-13, June, 2009; Revised August, 2009; Arabian Journal for Science and Engineering: Mathematics (AJSE-Mathematics), submitted.

[9] H. T. Banks and D. W. Iles, On compactness of admissible parameter sets: convergence and stability in inverse problems for distributed parameter systems, ICASE Report #86-38, NASA Langley Res. Ctr., Hampton VA 1986; *Proc. Conf. on Control Systems Governed by PDE's*, February, 1986, Gainesville, FL, Springer Lecture Notes in Control & Inf. Science, **97** (1987), 130–142.

[10] H. T. Banks and K. Kunsich, *Estimation Techniques for Distributed Parameter Systems*, Birkhauser, Boston, 1989.

[11] H. T. Banks and N. Lybeck, Modeling methodology for elastomer dynamics, CRSC-TR96-29, NCSU, September, 1996; in *Systems and Control in the 21st Century*, Birkhauser, Boston, 1996, 37–50.

[12] H. T. Banks and M. Pedersen, Well-posedness of inverse problems for systems with time dependent parameters, CRSC-TR08-10, August, 2008; *Arabian Journal of Science and Engineering: Mathematics,* **1** (2009), 39–58.

[13] H. T. Banks and J. R. Samuels, Detection of cardiac occlusions using viscoelastic wave propagation, CRSC-TR08-23, December, 2008; *Advances in Applied Mathematics and Mechanics,* **1** (2009), 1–28.

[14] H. T. Banks, R. C. Smith and Y. Wang, *Smart Material Structures: Modeling, Estimation and Control*, Masson Series on Research in Applied Math, Masson/J. Wiley, 1996.

[15] H. T. Banks and H. T. Tran, *Mathematical and Experimental Modeling of Physical and Biological Processes*, CRC Press, Boca Raton London New York, 2009.

[16] G. Bell and E. Anderson, Cell growth and division I. A mathematical model with applications to cell volume distributions in mammalian suspension cultures, *Biophysical Journal,* **7** (1967), 329–351.

[17] N. Bellomo and L. Preziosi, Modelling and mathematical problems related to tumor evolution and its interation with the immune system, *Math. and Comp. Modeling,* **32** (2000), 413–452.

[18] S. Bernard, L. Pujo-Menjouet and M. C. Mackey, Analysis of cell kinetics using a cell division marker: Mathematical modeling of experimental data, *Biophysical Journal*, **84** (2003), 3414–3424.

[19] J. J. Bird, D. R. Brown, A. C. Mullen, N. H. Moskowitz, M. A. Mahowald, J. R. Sider, T. F. Ajewski, C. Wang and S. L. Reiner, Helper T cell differentiation is controlled by the cell cycle, *Immunity,* **9** (1998), 229–237.

[20] S. Bonhoeffer, H. Mohri, D. Ho and A. S. Perleson, Quantification of cell turnover kinetics using 5-Bromo-2'-deoxyuridine, *J. of Immunology,* **164** (2000), 5049–5054.

[21] R. J. Carroll and D. Ruppert, *Transformation and Weighting in Regression*, Chapman & Hall, London, 1988.

[22] D. L. Chao, M. P. Davenport, S. Forrest and A. S. Perleson, Stochastic stage-structured modeling of the adaptive immune system, *Bioinformatics Conference CSB (2003): Proceedings 2003 IEEE* (Albuquerque, August 11-14, 2003) 124–131.

[23] M. Davidian and D. M. Giltinan, *Nonlinear Models for Repeated Measurement Data,* Chapman & Hall, London, 1995.

[24] R. J. de Boer, V. V. Ganusov, D. Milutinovic, P. Hodgkin and A. S. Perelson, Estimating lymphocyte division and death rates from CFSE data, *Bulletin of Mathematical Biology,* **68** (2006), 1011–1031.

[25] Y. C. Fung, *A First Course in Continuum Mechanics*, Prentice Hall, Englewood Cliffs, NJ, 1994.

[26] Y. C. Fung, *Biomechanics: Mechanical Properties of Living Tissue*, Springer-Verlag, Berlin, 1993.

[27] V. V. Ganusov, S. S. Pilyugin. R. J. de Boer, K. Murali-Krishna, R. Ahmed and R. Anti, Quantifying cell turnover using CFSE data, *J. Immunol. Methods,* **298** (2005), 183–200.

[28] A. V. Gett and P. D. Hodgkin, Cell division regulates the T cell cytokine repertoire, revealing a mechanism uderlying immune class regulation, *Proc. Natl. Acad. Sci. USA.,* **95** (1998), 9488–9493.

[29] A. V. Gett and P. D. Hodgkin, A cellular calculus for signal integration by T cells, *Nat. Immunol.,* **1** (2000), 239–244.

[30] M. Gyllenberg and G. F. Webb, A nonlinear structured population model of tumor growth with quiescence, *J. Math. Biol.,* **28** (1990), 671–694.

[31] E. D. Hawkins, Mirja Hommel, M. L Turner, Francis Battye, J. Markham and P. D Hodgkin, Measuring lymphocyte proliferation, survival and differentiation using CFSE time-series data, *Nature Protocols,* **2** (2007), 2057–2067.

[32] E. D. Hawkins, M. L. Turner, M. R. Dowling, C. van Gend and P. D. Hodgkin, A model of immune regulation as a consequence of randomized lymphocyte division and death times, *Proc. Natl. Acad. Sci.,* **104**:12, (2007). 5032–5037.

[33] P. D. Hodgkin, N. F. Go, J. E. Cupp and M. Howard, Interleukin-4 enhances anti-IgM stimulation of B cells by improving cell viaility and by increasing the sensitivity of B cells to the anti-IgM signal, *Cell. Immunol.,* **134** (1996), 14–30.

[34] H. Y. Lee, E. D. Hawkins, M. S. Zand, T. Mosmann, H. Wu, P. D. Hodgkin and A. S. Perelson, Interpreting CFSE obtained division histories of B cells in vitro with Smith-Martin and cyton type models, *Bull. Math Biol.,* **71** (2009), 1649–1670.

[35] N. L. Komarova, Stochastic modeling of drug resistance in cancer, *J. Theoretical Biol.,* **239** (2006), 351–366.

[36] N. L. Komarova and D. Wodarz, Effect of cellular quiescence on the success of targeted CML therapy, *PloS ONE* **2**:10, (2007), e990.

[37] K. León, J. Faro and J. Carneiro, A general mathematical framework to model generation structure in a population of asynchronously dividing cells, *J. Theor. Biol.*, **229** (2004), 455–476.

[38] T. Luzyanina, S. Mrusek, J. T. Edwards, D. Roose, S. Ehl and G. Bocharov, Computational analysis of CFSE proliferation assay, *J. Math. Biol.,* **54** (2007), 57–89.

[39] T. Luzyanina, D. Roose and G. Bocharov, Distributed parameter identification for a label-structured cell population dynamics model using CFSE histogram time-series data, *J. Math. Biol.,* to appear.

[40] T. Luzyanina, D. Roose, T. Schenkel, M. Sester, S. Ehl, A. Meyerhans and G. Bocharov, Numerical modelling of label-structured cell population growth using CFSE distribution data, *Theoretical Biology and Medical Modelling,* **4** (2007), 1–26.

[41] A. B. Lyons, Divided we stand: tracking cell proliferation with carboxyfluorescein diacetate succinimidyl ester, *Immunology and Cell Biology,* **77** (1999), 509–515.

[42] A. B. Lyons and K. V. Doherty, Flow cytometric analysis of cell division by dye dilution, *Current Protocols in Cytometry*, (2004), 9.11.1-9.11.10.

[43] A. B. Lyons and C. R. Parish, Determination of lymphocyte division by flow cytometry, *J. Immunol. Methods,* **171** (1994), 131–137.

[44] J. E. Marsden and T. J. R. Hughes, *Mathematical Foundations of Elasticity*, Dover Publications, Inc., Mineola, NY, 1994.

[45] G. Matera, M. Lupi and P. Ubezio, Heterogeneous cell response to topotecan in a CFSE-based proliferative test, *Cytometry A,* **62** (2004), 118–128.

[46] R.W. Ogden, *Non-Linear Elastic Deformations*, Dover Publications, Inc., Mineola, NY, 1984.

[47] B. Quah, H. Warren and C. Parish, Monitoring lymphocyte proliferation in vitro and in vivo with the intracellular fluorescent dye carboxyfluorescein diacetate succinimidyl ester, *Nature Protocols,* **2** (2007), 2049–2056.

[48] G. A. F. Seber and C. J. Wild, *Nonlinear Regression*, J. Wiley & Sons, Hoboken, NJ, 2003.

[49] L. F. Shampine, Solving hyperbolic PDEs in MATLAB, *Applied Numerical Analysis and Computational Mathematics,* **2** (2005), 346–358.

[50] J. Sinko and W. Streifer, A new model for age-size structure of a population, *Ecology,* **48** (1967), 910–918.

[51] J. A. Smith and L. Martin, Do cells cycle?, *Proc. Nat. Acad. Sci. USA,* **70** (1973), 1263–1267.

[52] Wikipedia, http://en.wikipedia.org/wiki/Fluorescence_spectroscopy.