Modelling count data with partial differential equation models in biology

Matthew J Simpson[a,*], Ryan J Murphy[b], Oliver J Maclaren[c]

[a] *School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia.*
[b] *School of Mathematics and Statistics, The University of Melbourne, Victoria, Australia.*
[c] *Department of Engineering Science, University of Auckland, Auckland 1142, New Zealand.*

Abstract

Partial differential equation (PDE) models are often used to study biological phenomena involving movement-birth-death processes, including ecological population dynamics and the invasion of populations of biological cells. Count data, by definition, is non-negative, and count data relating to biological populations is often bounded above by some carrying capacity that arises through biological competition for space or nutrients. Parameter estimation, parameter identifiability, and model prediction usually involves working with a measurement error model explicitly relating experimental measurements with the solution of a mathematical model, and in biological applications this is typically achieved by assuming the data are normally distributed about the solution of the mathematical model. Despite the widespread use of the standard additive Gaussian measurement error model, the assumptions inherent in this approach are rarely tested or compared with other options. Here we interpret scratch assay data, involving migration, proliferation and delays in a population of cancer cells using a reaction–diffusion PDE model. Experimental measurements are related to the PDE solution using a standard additive Gaussian measurement error model and we compare results with a biologically motivated binomial measurement error model. While estimates of model parameters are relatively insensitive to the choice of measurement error model, we find that model predictions are very sensitive. The standard additive Gaussian measurement error model leads to biologically inconsistent predictions, such as negative counts as well as counts that exceed the carrying capacity across a relatively large spatial region within the experiment. Furthermore, the standard additive Gaussian measurement error model is more computationally expensive than the binomial measurement error model. In contrast, the binomial measurement error model is both simpler to implement becuase it involves one less parameter to estimate, and the binomial measurement error model leads to biologically plausible predictions. We provide open access Julia software on GitHub to replicate all calculations in this work, and we explain how to generalise our approach to deal with coupled PDE models with several dependent variables through a multinomial measurement error model.

Keywords: Calibration; Parameter estimation; Identifiability; Prediction; Reaction–diffusion; Cell biology.

1. Introduction

Partial differential equation (PDE) models are widely adopted in the study of biological phenomena across a vast range of spatial and temporal scales. Applications include continental-scale spreading and colonisation of human populations [40], tissue-scale migration and colonisation of cells during development and disease progression [41], and molecular-scale motion and degradation of proteins during embryonic development [5]. Reaction–diffusion equations, describing combined diffusive migration and chemical reaction or biological birth/death processes are widely deployed across applications in ecology [14, 25] and cell biology [26, 35]. Some applications of these models involve both qualitative and quantitative matching of the spatial extent occupied by populations over time, such as measuring the area occupied by an invasive species [39] or invading population of cells [26]. Using reaction-diffusion PDE models in this way can be insightful and convenient since matching the spatial extent of spreading populations as a function of time avoids the need for keeping track of individuals within the population. More detailed applications of reaction-diffusion PDEs involves keeping track and counting individuals [20, 36], often expressing these counts in terms of observed densities by estimating the number of individuals per unit area (or per unit volume, depending on the geometry of the application). With such data it is possible to estimate parameters in the reaction-diffusion PDE, interpret biological processes in terms of those parameter values, and then make predictions and forecasts of future scenarios with the parameterised PDE model [38].

Calibrating the solution of a mathematical model to match experimental observations typically involves proposing both a *process model*, such as a differential equation that encodes key mechanisms relevant to the experimental system, together with a *measurement error model* that relates the solution of the process model to the observed data [16, 28]. Measurement error models are sometimes referred to as *noise models* in different parts of the literature and we will use both terminologies. Across various applications in the life sciences, the most commonly encountered measurement error model is a Gaussian additive error model which assumes that the observed data are normally distributed about the solution of the process model with some variance, σ^2 [16, 28], that can be either pre-estimated [16, 36] or determined as part of the model calibration process [28]. This approach is routinely implemented, partly because it is conceptually straightforward, and partly because enables us to write down a likelihood function that can be used for parameter estimation. Parameter estimates obtained by maximising the likelihood function for the Gaussian additive error model are closely related to parameter estimates associated with a best-fit model solution obtained via nonlinear regression and least–squares estimation [33]. In practice, however, the assumption of working with additive

*Corresponding author: matthew.simpson@qut.edu.au

and normally distributed measurement error is often unjustified, and as we will demonstrate, can lead to biologically inconsistent predictions for count data. Developing mathematical modelling techniques that deal with count data is an active area of research, particularly in time series analysis that is relevant to applications such as econometrics [15] and disease transmission [13]. Working with count data in spatiotemporal biological applications, such as ecological or cellular invasion processes, has received less attention.

In this work we consider a set of data describing a standard experiment in cell biology called a scratch assay. These experiments involve combined cell migration and cell proliferation, and they are widely used to understand collective cell behaviour across a range of applications including development and disease, such as cancer [10]. We take a very commonly–used approach and model the scratch assay data with a reaction–diffusion model that is closely related to the well-known Fisher-Kolmogorov model [2, 7, 22, 29]. While the Fisher-Kolmogorov model and generalisations thereof are widely used to interpret cell biology experiments [4, 34], exploring the impact of using different measurement error models is poorly understood for these applications. We consider likelihood-based parameter estimation, parameter identifiability and prediction [30, 38, 45] using two different measurement error models. In the first instance we implement the standard additive Gaussian noise model, and we then repeat the same procedure using an experimentally–motivated binomial measurement error model. While both approaches lead to similar parameter estimates, and both approaches lead to practically identifiable parameter estimates, the two measurement error models perform very differently in terms of making predictions with the calibrated process models. Working with the standard additive Gaussian noise model leads to biologically inconsistent results, whereas the binomial noise model provides biologically plausible predictions and is simpler to implement. While comparisons in this work focus on a single species process model, we also explain how these ideas apply to more complicated process models with coupled systems of PDE models. Open source Julia code is available on GitHub to replicate all results in this work.

2. Methods

2.1. Experimental data

We consider data reported by Jin et al. [17] who conducted a series of scratch assays with a PC-3 prostate cancer cell line [21]. Briefly, in these experiments cells were grown in a tissue culture plate to form a uniform monolayer (Figure 1(a)–(b)), and then scratched using a fine–tipped instrument (Figure 1(c)). The resulting invasion of the population into the initially–scratched region was imaged over 48 hours (Figure 1(d)–(e), (f)–(j)), after which the population appears to be confluent within the initially–scratched region. As we explain later, it is important to note

that Jin et al. [17] do not report any cell death during the experiments. The invasion of the population is thought to be primarily driven by cell migration and carrying capacity–limited cell proliferation [17]. Full experimental details are reported by Jin et al. [17].

Experimental images (Figure 1(f)–(j)) are recorded at 12-hour intervals over a period of 48 hours. While individual cells within the population are free to migrate in any direction, the geometry imposed by the initial condition means that the macroscopic density varies with horizontal position and time, so we aim to collect data to show how count data varies with horizontal location and time using the following simple procedure. Each image is divided into 39 equally–spaced vertical strips of width 50 μm and height 1430 μm (Figure 1(k)). The number of cells within each strip is manually counted at $t = 0, 12, 24, 36, 48$ hours, giving a total of $39 \times 5 = 195$ count measurements. Assuming that each count measurement is associated with the mid-point of each strip $x = 25, 75, 125, \ldots, 1925$ μm we can visualise the spatial and temporal evolution of the experiment by plotting the cell count per strip (Figure 1(l)) where we see the net effect of individual cells moving randomly which acts to increase the number of cells within the initially–vacant region during the first 24 hours of the experiment. The population simultaneously grows towards confluence as the experiment proceeds, which is most notable since the density appears to approach a spatially invariant density 48 hours after scratching.

While results in Figure 1(l) are plotted in terms of cell counts per strip, this data can be converted to cell density by dividing each cell count per strip by the area of the vertical strip $A = 50 \times 1430 = 71500$ μm^2. In this work we will use the variable $u^{\mathrm{o}}_{i,j}$ to denote the observed density data, where $i = 1, 2, 3, \ldots, 39$ is an index corresponding to the spatial location and $j = 1, 2, 3, 4, 5$ indexes time. We will also use the variable $U^{\mathrm{o}}_{i,j}$ to denote the observed count data, noting that $u^{\mathrm{o}}_{i,j} = U^{\mathrm{o}}_{i,j}/A$. Since A is a fixed constant we will use both $U^{\mathrm{o}}_{i,j}$ and $u^{\mathrm{o}}_{i,j}$ as appropriate.

3. Mathematical and statistical models

Throughout this study we work with dimensional models, where appropriate length units are μm and appropriate time units are hours. We will use a combination of cell count data and cell density data, where all dimensional densities have units of cells/μm^2.

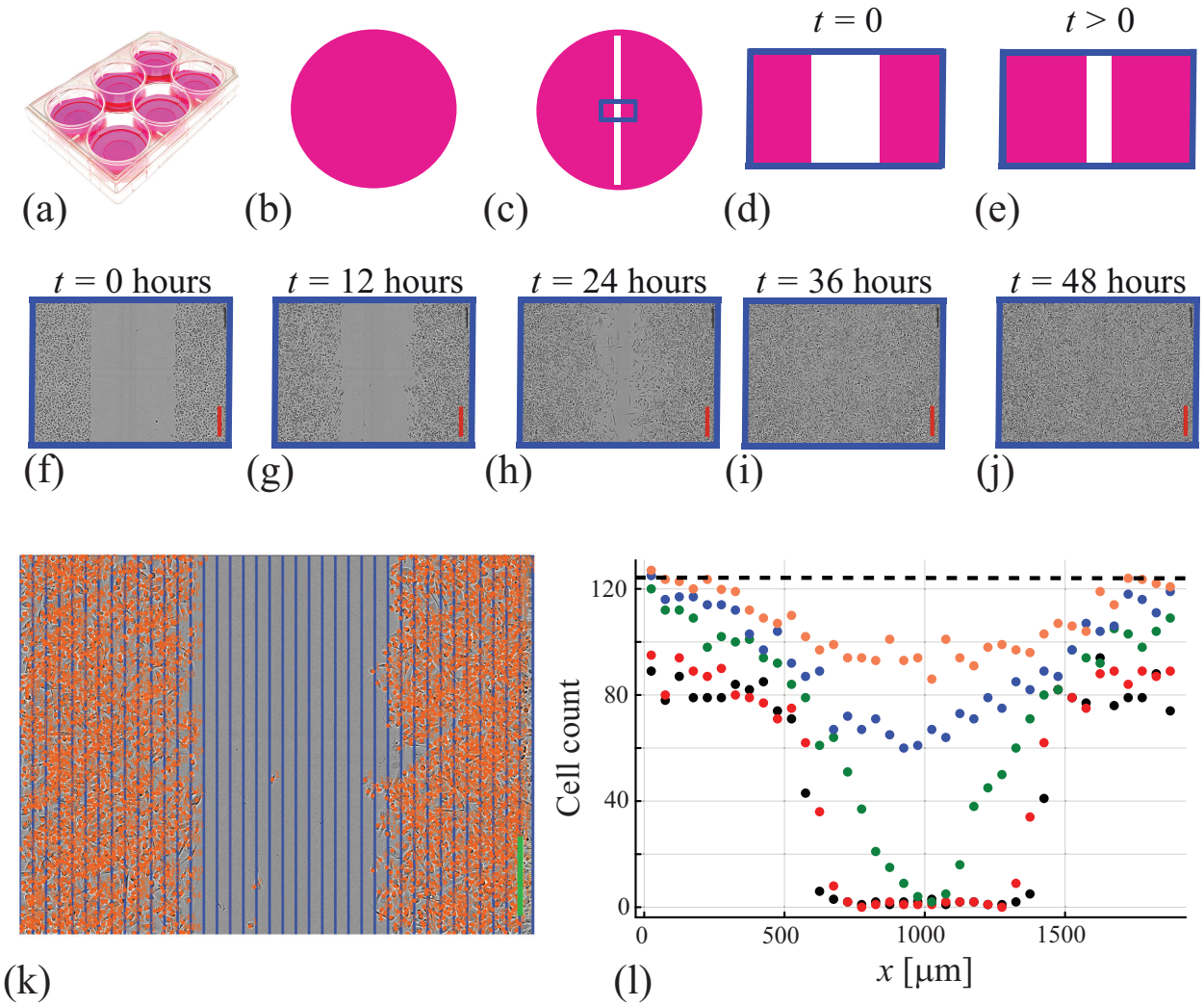

Figure 1: **Experimental protocol and data generation**. (a)–(e) Scratch assay protocol. (a)–(b) Cell monolayers are grown in 6–well tissue culture plates. (c) Scratched monolayer illustrating the field of view (solid blue). (d)–(e) Schematic showing population invasion over time within the field of view (solid blue). (f)–(j) Experimental images from scratch assay using PC-3 prostate cancer cells at $t = 0, 12, 24, 36, 48$ hours. Each image is 1950 μm wide and 1430 μm high, and the scale bar length is 300 μm. (k) Cell counting procedure at $t = 12$ hours, where the image is divided into 39 equally-spaced columns of width 50 μm. Individual cells are identified by placing an orange counter on the nucleus and are associated with each column, giving a total number of cells per column, as plotted in (l) where count data at $t = 0, 12, 24, 36, 48$ hours are reported in black, red, green, blue and orange discs, respectively. (l) includes a dashed horizontal line at $M = 122$ indicating the estimated maximum cell count per column.

3.1. Process model

Jin et al. [17] originally modelled these scratch assay data by assuming the cell populations evolved according to the dimensional Fisher–Kolmogorov model [2, 7, 22, 29] ,

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2} + \lambda u\left[1 - \frac{u}{K}\right], \tag{1}$$

on $0 < x < 1950$ μm and $t > 0$, where $u(x, t)$ [cells/μm^2] is the cell density, D [μm^2/hour] is the cell diffusivity, λ [/hour] is the proliferation rate and K [cells/μm^2] is the carrying capacity density. Jin et al. [17] explain that because the scratch is made within a spatially–uniform population (Figure 1(b)), the appropriate boundary conditions are to set $\partial u/\partial x = 0$ at both $x = 0$ and $x = 1950$ μm. The only parameter in the mathematical model that can be measured directly from these experimental images is the carrying capacity density, which Jin et al. [17] estimated by direct counting to give $K = 1.7 \times 10^{-3}$ cells/μm^2, which we will treat as a known constant. This is a convenient simplification that we will discuss later.

In the first analysis of this data, Jin et al. [17] implemented a least–squares procedure to estimate D and λ in Equation (1) to match the experimental data. Since the data in Figure 1 was first reported by Jin et al. [17], it has been re–examined a number of times [4, 24]. In 2020 Lagergren et al. [24] studied the same data and suggested a simple extension of the Fisher-KPP model by introducing a delay

$$\frac{\partial u}{\partial t} = \mathcal{T}(t)\left(D\frac{\partial^2 u}{\partial x^2} + \lambda u\left[1 - \frac{u}{K}\right]\right), \tag{2}$$

where $\mathcal{T}(t) \in [0, 1]$ is a non-dimensional, non-decreasing *delay* function with the property that $\lim_{t\to\infty} \mathcal{T}(t) = 1^-$. Lagergren et al. [24] showed that incorporating the delay function led to an improved match to the data, and they argued that such a delay is biologically reasonable because the physical act of scratching the monolayer transiently disturbs the population of cells, and some period of time after scratching is required for the cells to recover and resume their normal rate of migration and proliferation. This delay process has been observed in similar experiments [18], and here we take a parsimonious approach by setting $\mathcal{T}(t) = \tanh(\beta t)$, so that the parameter $\beta \geq 0$ [/hour] provides a simple measure of the delay, noting that this formulation recovers the standard Fisher-KPP equation as $\beta \to \infty$. Of course, other forms of $\mathcal{T}(t)$ could be explored, but here we work with $\mathcal{T}(t) = \tanh(\beta t)$ for simplicity since this approach is straightforward to implement, interpret and it works well for this particular application.

Throughout this study we solve Equation (2) numerically by discretizing the spatial derivatives on a uniformly discretized domain with mesh spacing $h = 5$ μm, leading to a system of coupled ordinary differential equations that we solve using Julia's SimpleDiffEq.jl package where automatic time step selection is implemented to control temporal

truncation error. This choice of discretisation leads to grid-independent results for the parameter ranges considered. The initial condition, $u(x, 0)$, is estimated by taking the 39 measurements of cell density at $t = 0$ and using linear interpolation to define $u(x, 0)$ as a piecewise linear function. Default interpolation options within SimpleDiffEq.jl provides a continuous estimate of $u(x, t)$ for $0 < x < 1950$ μm and $0 < t < 48$ hours. In summary, to solve Equation (2) we need to specify values of D, λ and β, and we will now explore how to relate the solution of Equation (2) to the experimental data using two different approaches.

3.2. Measurement error model 1: Additive Gaussian model

For comparison purposes we follow the usual approach and assume that our experimental measurements of cell density are noisy versions of the solutions of Equation (2) in the form

$$u^{\mathrm{o}}_{i,j} \mid \theta = \mathcal{N}\left(u_{i,j}, \sigma^2\right), \tag{3}$$

where θ is a vector of model parameters so that we treat the observations at index i, j as normally distributed random variables with mean $u(x_i, t_j) = u_{i,j}$ and unknown variance σ^2 that we estimate from the data. Therefore, with this approach we have four parameters to estimate, $\theta = (D, \lambda, \beta, \sigma)^\top$.

We take a likelihood–based approach to estimation, identifiability analysis and prediction [30, 45], and within this framework we have a log-likelihood function of the form

$$\ell(\theta \mid u^{\mathrm{o}}_{i,j}) = \sum_{i=1}^{39} \sum_{j=1}^{5} \log\left[\phi\left(u^{\mathrm{o}}_{i,j}; u_{i,j}(\theta), \sigma^2\right)\right], \tag{4}$$

where $\phi(x; \mu, \sigma^2)$ denotes a Gaussian probability density function with mean μ and variance σ^2.

3.3. Measurement error model 2: Binomial model

Our estimate of the carrying capacity density K implies that the maximum number of cells per column is $\lfloor 1.7 \times 10^{-3} \times 71500 \rceil = 122$. We treat raw cell counts within each column as a fraction of area within each column that is occupied by a population of cells, where each cell has the same size. This amounts to making an exchangeability assumption as we consider the fraction of area in each column occupied by cells to be important and that any rearrangement of individual cells in a particular column would be analysed in the same way. Since the solution of Equation (2) is continuous in space and time whereas our count data is spatially and temporally discrete, we propose a reasonable generative model for going from continuous label fractions to raw counts. First we note that each column can in fact contain two types of *objects*: cells and empty space (that could be putatively occupied by cells). We will

indicate these counts as $U \in \mathbb{Z}_+$ and $E \in \mathbb{Z}_+$, respectively, so that $U + E = M = 122$. Our key assumption is that the combined observed counts at any given spatiotemporal location are obtained by taking a finite number of independent and identically distributed samples that depend only on the underlying true or infinite data label fractions, which we denote by $u(x,t)$ and $e(x,t)$ respectively, such that $u(x,t) + e(x,t) = 1$, with $0 \leq u(x,t) \leq 1$ and $0 \leq e(x,t) \leq 1$.

Due to these constraints, count data are not independent, but have a binomial form likelihood

$$\begin{aligned} \ell(U(x,t), E(x,t); u(x,t), e(x,t)) &\propto u(x,t)^{U(x,t)} e(x,t)^{E(x,t)}, \\ &\propto u(x,t)^{U(x,t)} (1 - u(x,t))^{M - U(x,t)}. \end{aligned} \tag{5}$$

Taking the product over all spatial locations and temporal observation, the log-likelihood can be written as

$$\ell(\theta \mid U^{\mathrm{o}}_{i,j}) \propto \sum_{i=1}^{39} \sum_{j=1}^{5} \log \left[u_{i,j}(\theta)^{U^{\mathrm{o}}_{i,j}} (1 - u_{i,j}(\theta))^{(M - U^{\mathrm{o}}_{i,j})} \right], \tag{6}$$

where $M = 122$ for our application, and the proportionality constant is independent of u and e, and so is not required for the likelihood function. Accordingly, we evaluate this log-likelihood function by simply setting the proportionality constant to unity. Using this approach we have just three parameters to estimate, $\theta = (D, \lambda, \beta)^\top$.

4. Results and Discussion

Given our two likelihood functions, (4) and (6), we now compare likelihood-based parameter estimation, practical identifiability analysis and prediction for the two different choices of measurement noise models.

4.1. Parameter estimation

Applying maximum likelihood estimation (MLE) gives the best fit set of parameters,

$$\hat{\theta} = \sup_{\theta} \left[\ell(\theta \mid u^{\mathrm{o}}_{i,j}) \right]. \tag{7}$$

All numerical optimisation computations in this work are obtained using the Nelder-Mead algorithm with simple bound constraints [19], and we find that all numerical optimisation computations converge using default stopping criteria, giving $\hat{\theta} = (\hat{D}, \hat{\lambda}, \hat{\beta}, \hat{\sigma})^\top = (1466, 0.090, 0.067, 0.054)^\top$ for the additive Gaussian measurement model, and $\hat{\theta} = (\hat{D}, \hat{\lambda}, \hat{\beta})^\top = (1424, 0.095, 0.058)^\top$ for the binomial measurement model, thus indicating that estimates of D, λ and β are not vastly different. Figure 2(a)–(b) compares the experimental data with the solution of Equation (2) evaluated with $\hat{\theta}$ for the standard additive Gaussian noise model and the binomial measurement model, respectively, and we see that, as expected, both MLE solutions compare well with the data. Figure 2(c) compares the two MLE

solutions where we observe just small differences between the two solutions near the centre of the domain at $t = 24$ and 36 hours. Figure 2(d) compares the MLE delay functions, $\tau(t)$, for the two measurement models, where we see that the binomial measurement model leads to a slightly slower delay. Overall, the differences in MLE parameter estimates and the solution of the process model evaluated with $\hat{\theta}$ are relatively small. We note that the two estimates of $\hat{D}$ differ by just 3%, which is very small given that literature estimates of cell diffusivity can vary over an order of magnitude [1, 17, 34]. In summary, the differences in MLE parameter estimates and the solution of the process model evaluated with $\hat{\theta}$ are relatively small. Overall, at this point, the only practical difference we observe in working with the two different measurement error models is that the binomial model involves less parameters which means that the numerical optimisation calculations are simpler than working with the standard additive Gaussian measurement model.

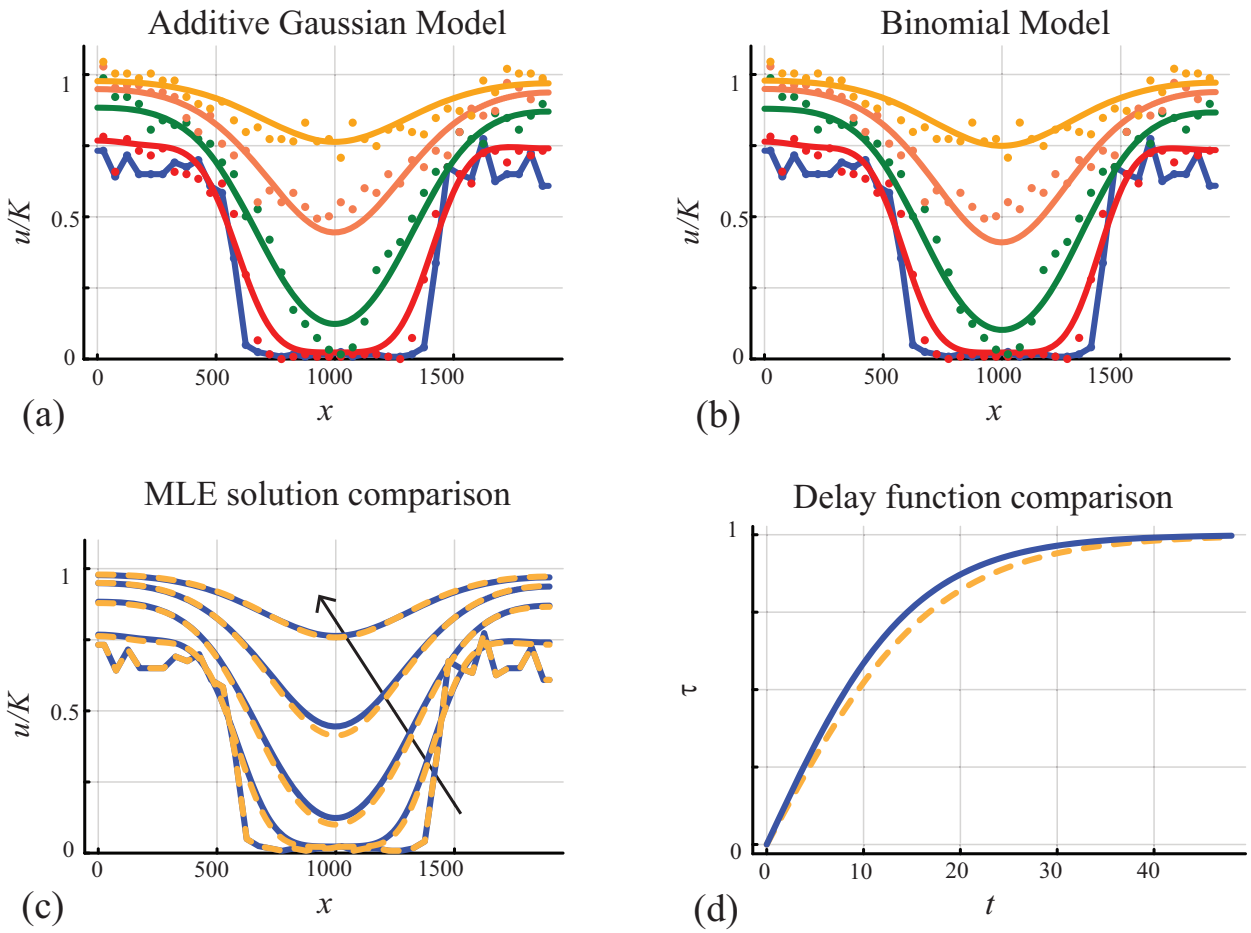

Figure 2: **Parameter estimation**. Comparison of experimental data with the MLE solution with (a) $\hat{\theta} = (\hat{D}, \hat{\lambda}, \hat{\beta}, \hat{\sigma})^\top = (1492, 0.090, 0.068, 0.054)^\top$ for the additive measurement error model, and (b) $\hat{\theta} = (\hat{D}, \hat{\lambda}, \hat{\beta})^\top = (1454, 0.095, 0.057)^\top$ for the binomial measurement error model. (a)–(b) Comparison of u^{o} (discs) and $u(x,t)$ (solid curves) at $t = 0, 12, 24, 36, 48$ hours correspond to black, red, green, blue and orange. (c) Comparison of the additive measurement error model MLE solution (solid blue) with the binomial measurement error model MLE solution (dashed orange) at $t = 0, 12, 24, 36$ and 48 hours, with the arrow showing the direction of increasing time. (d) Comparison of the delay function $\tau(t)$ for the additive noise model (solid blue) and the binomial noise model (dashed orange).

4.2. Practical identifiability analysis

To assess the practical identifiability of our parameter estimates we work with a normalised log-likelihood function $\hat{\ell}(\theta \mid u^{\mathrm{o}}_{i,j}) = \ell(\theta \mid u^{\mathrm{o}}_{i,j}) - \ell(\hat{\theta} \mid u^{\mathrm{o}}_{i,j})$ which we consider as a function of θ for fixed data. Taking a profile likelihood-based approach [30–32, 43, 44], we partition the full parameter θ into interest parameters ψ and nuisance parameters ω, such that $\theta = (\psi, \omega)$. While, in general, an interest parameter can be defined as any function of the full parameter vector $\psi = \psi(\theta)$, here we focus on constructing a series of univariate profiles by taking the interest parameter to be a single parameter. For a set of data $u^{\mathrm{o}}_{i,j}$, the profile log-likelihood for the interest parameter ψ given the partition (ψ, ω) is

$$\hat{\ell}_p(\psi \mid u^{\mathrm{o}}_{i,j}) = \sup_{\omega \mid \psi} \hat{\ell}(\psi, \omega \mid u^{\mathrm{o}}_{i,j}), \tag{8}$$

which implicitly defines a function $\omega^*(\psi)$ of optimal values of ω for each value of ψ, and defines a surface with points $(\psi, \omega^*(\psi))$ in parameter space. For the univariate profiles we consider here $(\psi, \omega^*(\psi))$ is simply a univariate curve. As for calculating the MLE, univariate profile likelihoods are obtained using the Nelder-Mead algorithm to evaluate the necessary numerical optimisation problems [19]

To assess the practical identifiability using the two normalised log-likelihood functions we compute a series of univariate profile likelihoods shown in Figure 3. For the additive measurement error model we have four parameters and four univariate profile likelihood functions obtained by considering: (i) $\psi = D$ and $\omega = (\lambda, \beta, \sigma)$; (ii) $\psi = \lambda$ and $\omega = (D, \beta, \sigma)$; (iii) $\psi = \beta$ and $\omega = (D, \lambda, \sigma)$; and, (iv) $\psi = \sigma$ and $\omega = (D, \lambda, \beta)$. For the binomial measurement error model we have three parameters and three univariate profile likelihood functions: (i) $\psi = D$ and $\omega = (\lambda, \beta)$; (ii) $\psi = \lambda$ and $\omega = (D, \beta)$; and, (iii) $\psi = \beta$ and $\omega = (D, \lambda)$. In each case the univariate profiles are well formed about a single distinct peak indicating that all parameters in both log-likelihood functions are practically identifiable.

The degree of curvature of the log-likelihood function is related to inferential precision [30]. Since, in general, the univariate profiles for the binomial measurement error model exhibit higher curvature near the MLE than the univariate profiles for the standard additive Gaussian measurement error we conclude that inference with the binomial error model is more precise. To see this we can form asymptotic confidence intervals by estimating the interval where $\hat{\ell}_p \geq -\Delta_{0.95,1}/2 \approx -1.92$ which is the asymptotic 95% threshold [30, 38]. For the cell diffusivity D we have $\hat{D} = 1492$ $\mu\mathrm{m}^2$/hour for the additive Gaussian measurement error model and the 95% confidence interval of [1295, 1720], and for the binomial measurement error model we have $\hat{D} = 1454$ $\mu\mathrm{m}^2$/hour and the 95% confidence interval is [1314, 1609], which is narrower and more precise. Overall, since the univariate profiles for D, λ and β are narrower for the binomial noise model for D, λ and β, which means that the binomial measurement error model leads to more precise inference.

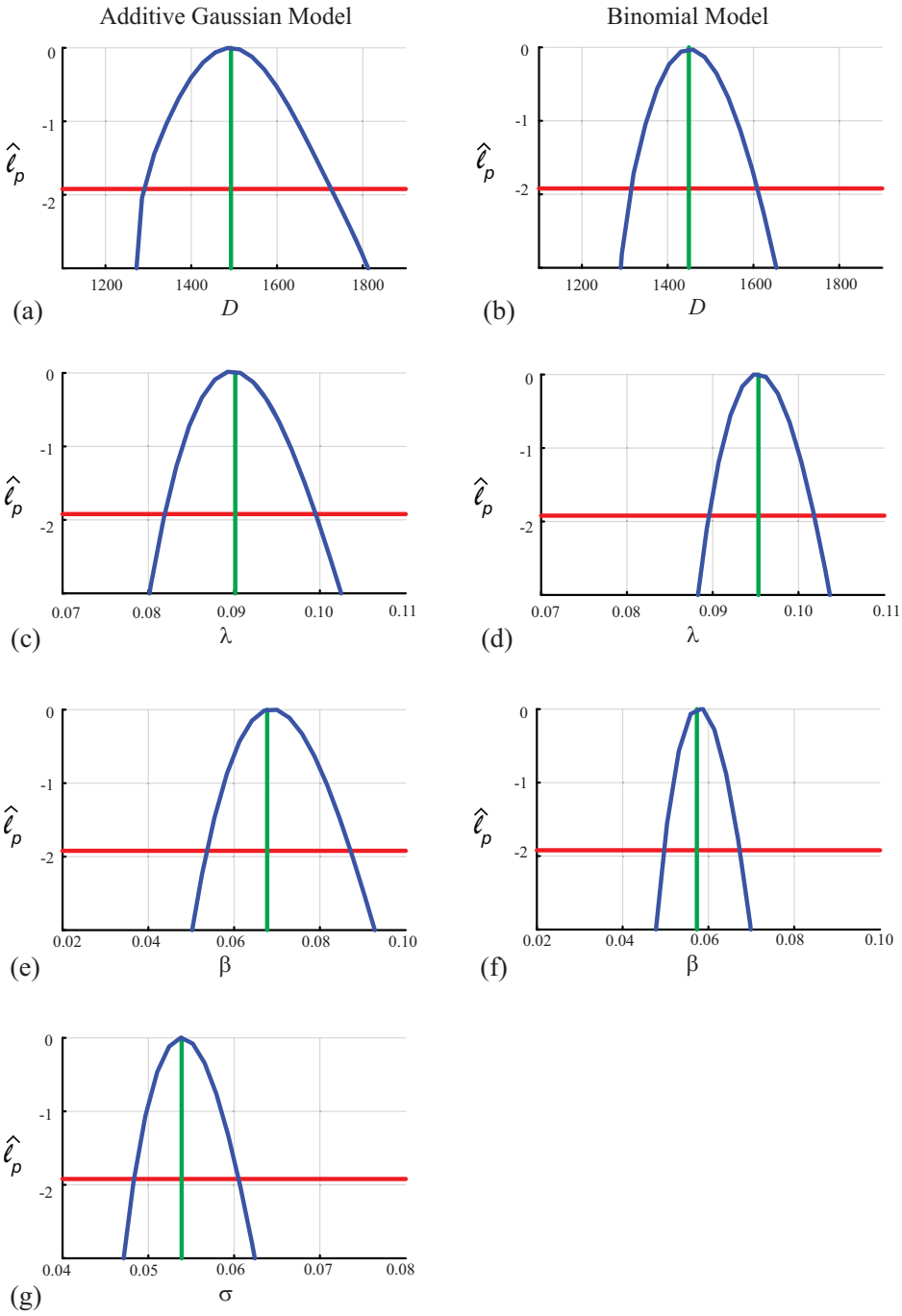

Figure 3: **Parameter identifiability**. Univariate profiles for parameters associated with the additive Gaussian measurement error model are given in the left column, while the right column shows univariate profiles for parameters associated with the binomial measurement error model. Comparisons are given for: (a)–(b) D; (c)–(d) λ and (e)–(f) β. Results for σ for the additive measurement error model are given in (g). Profiles are computed by discretising the interval of interest parameter into 30 equally spaced intervals and evaluating Equation (8) at each point. Continuous profiles are obtained using interpolation. In each profile we superimpose a vertical green line at the MLE value, and the asymptotic 95% threshold is shown as a horizontal red line at $\ell_p^* = -\Delta_{0.95,1}/2 = -1.92$, where $\Delta_{q,d}$ refers to the qth quantile of the χ^2 distribution with d degrees of freedom, here $d = 1$ for univariate profiles. For the additive Gaussian measurement error model the 95% confidence intervals are $\hat{D} = 1492$ [1295, 1720] μm^2/hour; $\hat{\lambda} = 0.090$ [0.082, 0.099] /hour; $\hat{\beta} = 0.068$ [0.054, 0.086] /hour; and $\hat{\sigma} = 0.054$ [0.048, 0.061] cells/μm^2. For the binomial measurement error model the 95% confidence intervals are $\hat{D} = 1454$ [1314, 1609] μm^2/hour; $\hat{\lambda} = 0.095$ [0.090, 0.101] /hour and $\hat{\beta} = 0.057$ [0.050, 0.067] /hour.

Before considering prediction we briefly explore the empirical, finite sample coverage properties for our parameter confidence intervals as they are only expected to hold asymptotically [30]. For each measurement error model we generate 1000 data realisations using the previously–identified MLE $\hat{\theta}$ for both approaches. For each data realisation we re–compute the MLE and count the proportion of those realisations with $\hat{\ell} \geq \ell^* = -\Delta_{0.95,d}/2$, where $\hat{\ell}$ is evaluated at the true parameter values and d is the number of degrees of freedom. This condition means the true parameter vector is within the sample-dependent, likelihood-based confidence set. Note that we have $d = 4$ for the additive Gaussian error model and $d = 3$ for the binomial noise model. This re-sampling procedure gives $957/1000 = 95.7\%$ and $941/1000 = 94.1\%$ for the additive Gaussian and binomial measurement models, respectively, indicating that both cases are close to the expected asymptotic result.

4.3. Likelihood-based predictions for mean trajectories

Having compared the two measurement error models in terms of parameter estimation and practical identifiability, we now consider the implications of these different measurement error models in terms of making likelihood–based predictions to understand how uncertainty in parameter estimates propagates into uncertainty in predictions and observed quantities. For both measurement error models we implement a simple rejection sampling approach to sample parameter combinations from the full parameter space until we have S samples for which $\hat{\ell} \geq -\Delta_{0.95,d}/2$. For each of these parameter samples we solve Equation (2) to generate S solution profiles, $u^{(s)}(x,t)$ for $s = 1, 2, 3, \ldots, S$. Each of these solution profiles is associated with parameter values that lie within the asymptotic 95% confidence region. For each value of $t_j = 12(j-1)$ for $j = 1, 2, 3, 4$, and $x_i = 5(i-1)$ for $i = 1, 2, 3, \ldots 381$, we evaluate $u^{(s)}(x_i, t_j)$ for $s = 1, 2, 3, \ldots, S$. At each spatial location i and time point j, the prediction interval is given by the union of points $u^{(s)}(x_i, t_j)$ for $s = 1, 2, 3, \ldots, S$, which is computed by evaluating the interval $[\min(u^{(s)}(x_i, t_j)), \max(u^{(s)}(x_i, t_j))]$ for each i and j. Figure 4 compare mean prediction intervals for the additive and binomial models, respectively.

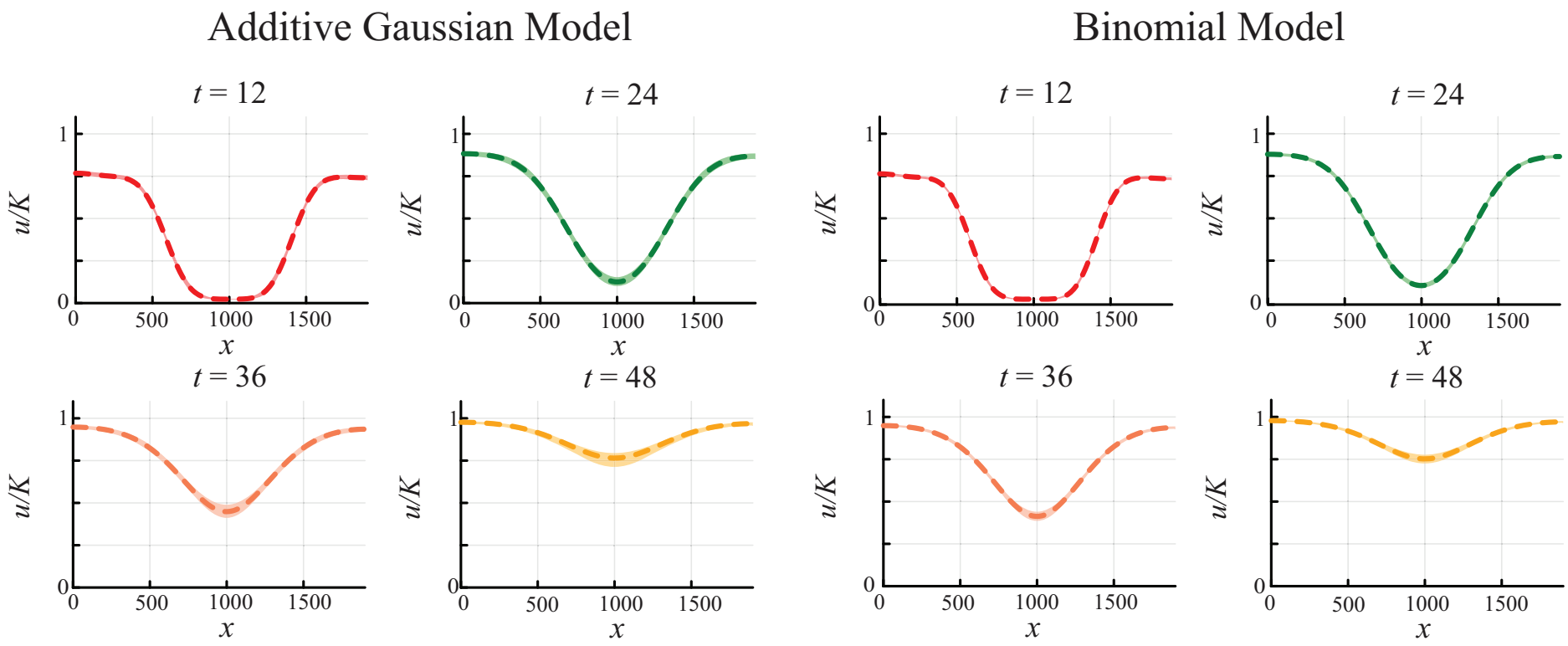

Figure 4: **Mean trajectory prediction intervals**. Comparison of mean trajectory prediction intervals for the additive Gaussian measurement error model (left four panels) with the binomial measurement error model (right four panels) constructed using $S = 10^3$ parameter samples. Each panel shows the MLE solution at $t = 0, 12, 24, 36, 48$ hours plotted in red, green, orange and yellow dashed lines, respectively. Each panel shows a shaded interval about the MLE solution corresponding to the prediction interval for mean model predictions.

Comparing the widths of the prediction intervals for mean trajectories in Figure 4 indicates that the binomial measurement error model leads to slightly narrower prediction intervals than the standard additive Gaussian measurement error model. This is consistent with the results in Figure 3 where we saw that the univariate profiles for the binomial measurement error model were slightly narrower than the respective profiles obtained using with the additive Gaussian measurement error model. Overall, however, if we visually compare the magnitude of the fluctuations present in the individual data realisations in Figure 1(l) with the width of the mean prediction intervals in Figure 4 it is clear that these mean prediction intervals do not capture the observed variability in a single experiment for either of the measurement error models . This observation motivates us to consider prediction intervals for data realisations [28].

4.4. Likelihood-based predictions for data realisations

Using the same S parameter samples identified in Section 4.3, we compute $u^{(s)}(x_i, t_j)$ for $i = 1, 2, 3, \ldots 381$, $j = 1, 2, 3, 4$ and $s = 1, 2, 3, \ldots, S$ as before. Again, we note that each of these S solution profiles is associated with parameter values that lie within the asymptotic 95% confidence region. For each value of $u^{(s)}(x_i, t_j)$ for $i = 1, 2, 3, \ldots 381$, $j = 1, 2, 3, 4$ and $s = 1, 2, 3, \ldots, S$, we evaluate the 5% and 95% quantile of the associated measurement error distribution, which we denote $u^{(s)}_{0.05}(x_i, t_j)$ and $u^{(s)}_{0.95}(x_i, t_j)$, respectively. One way of interpreting this is that instead of considering a single *mean* prediction for each θ like we did in Figure 4, we now compute an individual prediction interval for each parameter value, and we refer to this as a prediction for *data realisations* [38]. For each x_i and t_j we take the union over the individual 5% and 95% quantiles, $[\min(u^{(s)}_{0.05}(x_i, t_j)), \max(u^{(s)}_{0.95}(x_i, t_j))]$ which gives the prediction intervals for data realisations in Figure 5. Each panel in Figure 5 is superimposed with dashed horizontal lines at $u = 0$ and $u = K$ which represent lower and upper bounds on the biologically–relevant values of density.

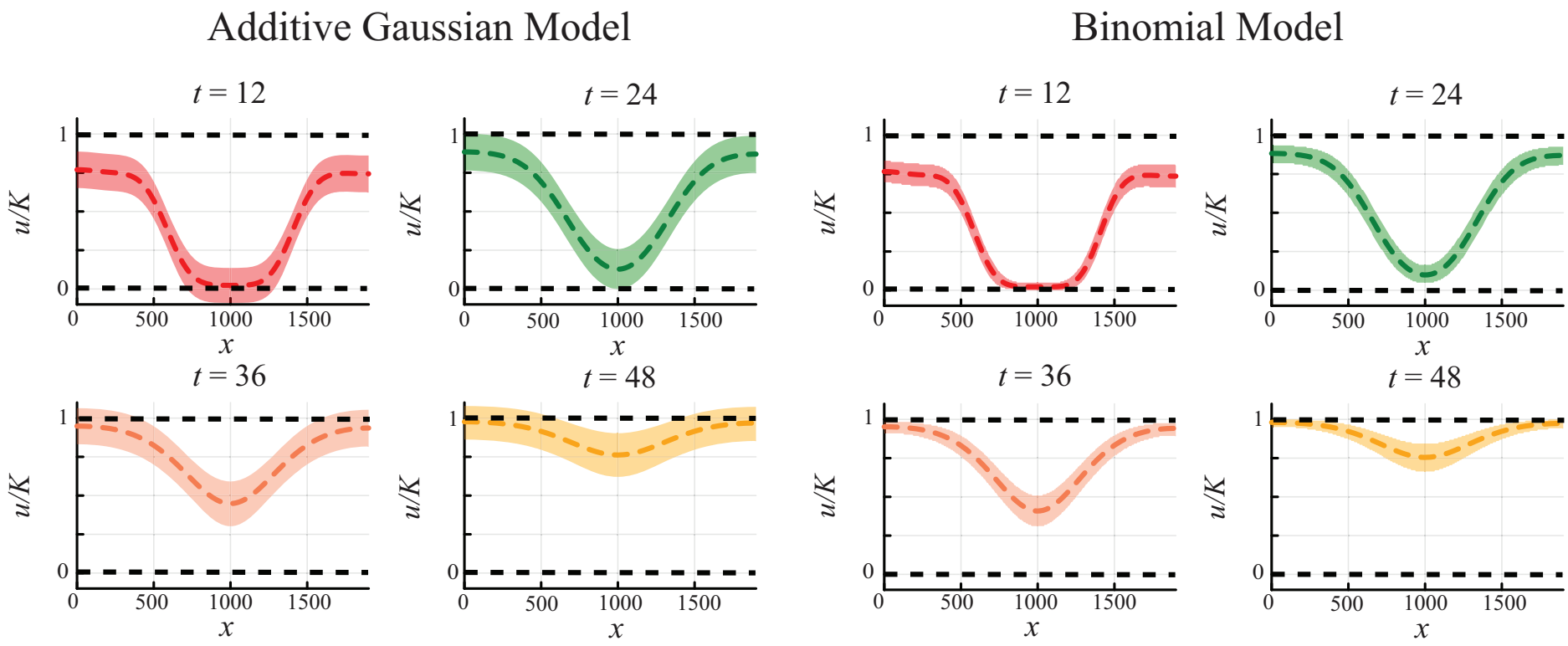

Figure 5: **Data realisation prediction intervals**. Comparison of data realisation prediction intervals for the additive Gaussian measurement error model (left four panels) with the binomial measurement error model (right four panels) constructed using $S = 10^3$ parameter samples. Each panel shows the MLE solution at $t = 0, 12, 24, 36, 48$ hours plotted in red, green, orange and yellow dashed lines, respectively. Each panel shows a shaded interval about the MLE solution corresponding to the prediction interval for data realisations generated by considering the 5% quantile and 95% quantile for each mean trajectory. Each panel shows horizontal dashed lines at $u = 0$ and $u = K$ which indicate biologically relevant lower and upper bounds for the density.

The prediction intervals for data realisations in Figure 5 for the additive Gaussian measurement error model are biologically unrealistic since they predict extensive regions where the density is negative across the central region of the experiment at $t = 12$ and 24 hours. In particular, these predictions include $u/K < 0$ in the region $500 < x < 1200$ μm at $t = 12$ hours. Furthermore, the additive Gaussian measurement model predicts the possibility of extensive overcrowded regions with $u/K > 1$ at $t = 36$ and 48 hours. Figure 5 shows that the prediction intervals at $t = 48$ hours include $u/K > 1$ for $x < 600$ μm and for $x > 1400$ μm. We refer to these predictions as biologically *unrealistic* because cell density is non-negative by definition, and that setting $u/K > 1$ in Equation (2) implies that the source term is negative and that would be associated with localised cell death that was never observed experimentally by Jin et al. [17]. While the experimental data in Figure 1(l) does include a few isolated columns for which $u/K > 1$, these measurements are relatively infrequent (i.e. approximately 2% of measurements) and are not observed across extensive regions of the experiments. Instead, we expect that these experimental observations are likely associated with our simplifying assumption that all cells have the same size, and that this size does not change over the course of the cell cycle. In contrast to the prediction intervals from the additive Gaussian measurement error model, the prediction intervals generated by using the binomial measurement model are naturally bounded, meaning that all predictions are within the biologically–plausible interval $0 < u/K < 1$ at each time point at all spatial locations, which is far more consistent with our experimental observations. In addition to making biologically realistic predictions, the width of the data realisation prediction intervals in Figure 5 for the binomial measurement error model is comparable to magnitude of the fluctuations in the experimental data observed in Figure 1(l).

5. Conclusions and Outlook

In this work we have explicitly compared different measurement error models that relate a reaction–diffusion PDE model with count data describing the invasion of a population of motile and proliferative cancer cells in a scratch assay [10]. In particular we compare the standard additive Gaussian measurement error model with a biologically–motivated binomial measurement error model that explicitly avoids negative count data and count data exceeding the carrying capacity density. Implementing both measurement error models together with a reaction–diffusion process model based on the Fisher–Kolmogorov PDE provides insight into the relative advantages of both approaches. Both measurement error models provide similar parameter estimates, and that the parameter estimates are practically identifiable in both cases. This comparison is not necessarily surprising, especially if we interpret the binomial model by the usual normal approximation. The usual central limit theorem states that $U/M \to \mathcal{N}(u, u(1-u)/M)$ for any value

of the cell counts at a fixed spatiotemporal location. In this approximation the maximum variance is $\sigma^2 = 0.0041$ or $\sigma = 0.064$, which compares very well with our estimated value of $\sigma = 0.054$ obtained under the constant variance assumption. This provides a simple explanation of why the two measurement error models lead to similar parameter estimates and similar profile likelihoods. However, as we demonstrate, likelihood–based predictions are very different between the two approaches, with the standard additive model leading to biologically unrealistic predictions, whereas the binomial measurement model leads to naturally bounded, biologically reasonable predictions. We feel that the predictive differences between the two measurement error models is important because understanding the variability in model predictions and outputs is often more of greater interest to practitioners than understanding the variability in model parameters. In addition to these practical considerations, from a computational point of view the binomial measurement error model can be implemented at a reduced computational cost since it involves dealing with one less unknown parameter.

There are many options for extending the work presented in this study. Here we deliberately work with a relatively straightforward process model, Equation (2), involving linear diffusion, logistic growth, and delay. In principle the same procedures outlined here can be applied to more complicated process models that might describe other biological processes of interest, such as generalising the linear diffusion term in Equation (2) to a nonlinear diffusion term to capture sharp–fronted profiles [26, 34, 35]. In essence this involves generalising the linear diffusion flux in Equation (2), given by $J = -D\partial u/\partial x$, to a nonlinear diffusion flux, given by $J = -\mathcal{D}(u)\partial u/\partial x$. In practice this involves replacing the constant linear diffusivity D with a nonlinear diffusion function $\mathcal{D}(u)$ [26, 34, 35]. In many applications the choice of $\mathcal{D}(u)$ is an open question [27], but in practice it is often taken to be a power law, $\mathcal{D}(u) = D(u/K)^{\alpha}$, with $\alpha > 0$ [17]. Therefore, calibrating the extended model with a nonlinear diffusion term involves estimating both the coefficient D and exponent α. A similar option is to generalise the logistic source term in Equation (2) by replacing $\lambda u(1-[u/K])$ with some other sigmoid model [9], such as the Gompertz model $\lambda u \log([u/K])$ [23], the Richards model $\lambda u(1 - [u/K]^{\omega})$ [37], or a more general sigmoid growth model $\lambda u(1 - [u/K]^{\omega})^{\delta}$, with $\omega > 0$ and $\delta > 0$ [42]. Again, this typically involves introducing additional unknown exponents into the process model that need to be estimated. In practice, however, we caution against working with unnecessarily complicated, potentially over parameterised models without gathering additional data otherwise we may encounter practical identifiability issues where there is insufficient information in the data to constrain the parameter values, and this would manifest in flat or insufficiently peaked univariate profiles [37]. We could also take a similar approach to modelling the delay $\tau(t)$. In this work we assume that both cell migration and cell proliferation are modelled using a very simple delay function $\tau(t) = \tanh(\beta t)$

which appears to capture the experimental data reasonably well. Of course, we could take many different approaches by specifying a more complicated form for $\tau(t)$, or we could specify different function forms for the delay to migration $\tau_{\mathrm{m}}(t)$ and proliferation $\tau_{\mathrm{p}}(t)$. Both of these different approaches involve working with more unknown parameters, and again we do not take this approach here to avoid issues of parameter nonidentifiability.

The proposed binomial measurement error model is one member of a family of measurement error models that can be implemented to interpret biological or ecological data. Consider, for example, a more detailed and practical scenario where the population of individuals undergoing migration and proliferation is composed of more than one type of subpopulation. In ecology this situation is often referred to as a population composed of several communities, whereas in cell biology this kind of situation arises in a co-culture experiment where two or more different types of cells are present in the same experiment. Performing scratch assays and other types of cell migration assays under co-culture conditions is fairly common, such as the work of Haridas et al. [11, 12] who studied the migration, proliferation and invasion of a population of cells composed of a mixture of melanoma and fibroblast cells to mimic the *in vivo* situation. Similarly, Simpson et al. [36] analysed scratch assay data involving a population of melanoma cells that were labelled with a fluorescent marker so that the subpopulation of cells in G1 phase of the cell cycle fluoresced red and the subpopulation of cells in the G2/S/M phase of the cell cycle fluoresced green. In both these cases, instead of having a single reaction–diffusion PDE describing the evolution of the total cell density $u(x,t) \in [0,K]$, it is natural to generalise the modelling framework to deal with a system of N coupled PDE models describing the evolution of subpopulations with density $u_n(x,t)$, typically constrained so that $\sum_{n=1}^{N} u_n(x,t) \in [0,K]$. For example, if we consider a co-culture experiment with $N = 2$ subpopulations [11, 36], it is straightforward to extend the binomial measurement error model in the following way. Treating raw counts of each cell type at each spatial location as a fraction of space occupied by cells amounts to invoking an exchangeability assumption In this case, each column can contain three types of objects: cells from subpopulation 1, cells from subpopulation 2, and empty space, and we indicate these counts as $U_1 \in \mathbb{Z}_+$, $U_2 \in \mathbb{Z}_+$ and $E \in \mathbb{Z}_+$, respectively, so that $U_1 + U_2 + E = M$. Similarly, we denote the associated infinite data label fractions as $u_1(x,t)$, $u_2(x,t)$ and $e(x,t)$, with $u_1(x,t) + u_2(x,t) + e(x,t) = 1$, leading to a multinomial likelihood

$$\begin{aligned}\ell(U_1(x,t), U_2(x,t), E(x,t); u_1(x,t), u_2(x,t), e(x,t)) &\propto u_1(x,t)^{U_1(x,t)} u_2(x,t)^{U_2(x,t)} e(x,t)^{E(x,t)}, \\ &\propto u_1(x,t)^{U_1(x,t)} u_2(x,t)^{U_2(x,t)} (1 - u_1(x,t) - u_2(x,t))^{M - U_1(x,t) - U_2(x,t)}.\end{aligned} \tag{9}$$

This process can be extended to $I > 2$ subpopulations in an obvious way.

We conclude by commenting that all work here has been presented in a frequentist, likelihood–based framework for parameter estimation, parameter identifiability analysis and prediction. We have chosen to work within a frequentist framework for the sake of expositional simplicity and computational efficiency. However, all results presented here could have been considered within a Bayesian setting by treating the two likelihood functions within the context of sampling rather than optimisation [8, 45]. Our previous work that directly compared parameter estimation, identifiability analysis and prediction suggests that taking a Bayesian approach leads to comparable results to a frequentist likelihood–based approach [36], albeit incurring an increased computational overhead since sampling is, in general, more computationally expensive than optimisation.

Acknowledgements MJS is supported by the Australian Research Council (DP230100025).

References

[1] Cai AQ, Landman KA, Hughes BD. 2002. Multi-scale modelling of a wound-healing migration assay. Journal of Theoretical Biology. 245, 576–594.

[2] Canosa J. 1973. On a nonlinear diffusion equation describing population growth. IBM Journal of Research and Development. 17, 307–313.

[3] Casella G, Berger R. 2001. Statistical Inference. Belmont, CA: Duxbury.

[4] Chen Z, Liu Y, Sun H. 2021. Physics-informed learning of governing equations from scarce data. Nature Communications. 12, 6136.

[5] Ciocanel M-V, Ding L, Mastromatteo L, Reichheld S, Cabral S, Mowry K, Sandstede B. 2023. Parameter identifiability in PDE models of fluorescence recovery after photobleaching. Preprint.

[6] Cole D. 2020. Parameter redundancy and identifiability. CRC Press.

[7] Fisher RA. 1937. The wave of advance of advantageous genes. Annals of Eugenics. 7, 355–369.

[8] Gelman A, Vehtari A, Simpson D, Margossian CC, Carpenter B, Yao Y, Kennedy L, Gabry J, Bürkner PC, Modrák M. 2020. Bayesian workflow. arXiv preprint. (Preprint).

[9] Gerlee P. 2013. The model muddle: In search of tumor growth laws. Cancer Research. 73, 2407–2411.

[10] Grada A, Otero-Vinas M, Prieto-Castrillo F, Obagi Z, Falanga V. 2016. Research techniques made simple: analysis of collective cell migration using the wound healing assay. Journal of Investigative Dermatology. 137, e11-e16.

[11] Haridas P, Penington CJ, McGovern JA, McElwain DLS, Simpson MJ. 2017a. Quantifying rates of cell migration and cell proliferation in co-culture barrier assays reveals how skin and melanoma cells interact during melanoma spreading and invasion. Journal of Theoretical Biology. 423, 13–25.

[12] Haridas P, McGovern JA, McElwain DLS, Simpson MJ. 2017b. Quantitative comparison of the spreading and invasion of radial growth phase and metastatic melanoma cells in a three-dimensional human skin equivalent model. Peer J. 5, e3754.

[13] He D, Ionides EL, King AA. 2010. Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. Journal of the Royal Society Interface. 7, 271–283

[14] Holmes EE, Lewis MA, Banks JE, Veit RR. 1994. Partial differential equations in ecology: Spatial interactions and population dynamics. Ecology. 75, 17–29.

[15] Hilbe JM. 2014. Modeling count data. Cambridge University Press.

[16] Hines KE, Middendorf TR, Aldrich RW. 2014. Determination of parameter identifiability in nonlinear biophysical models: A Bayesian approach. Journal of General Physiology. 143, 401.

[17] Jin W, Shah ET, Penington CJ, McCue SW, Chopin LK, Simpson MJ. 2016. Reproducibility of scratch assays is affected by the initial degree of confluence: experiments, modelling and model selection. Journal of Theoretical Biology. 390, 136–145.

[18] Jin W, Shah ET, Penington CJ, McCue SW, Maini PK, Simpson MJ. 2017. Logistic proliferation of cells in scratch assays is delayed. Bulletin of Mathematical Biology. 79, 1028–1050.

[19] Johnson SG. 2022. The NLopt module for Julia. Retrieved September 2023 NLopt.

[20] Johnston ST, Ross JV, Binder BJ, McElwain DLS, Haridas P, Simpson MJ. 201). Quantifying the effect of experimental design choices for *in vitro* scratch assays. Journal of Theoretical Biology. 400, 19–31.

[21] Kaighn ME, Narayan KS, Ohnuki Y, Lechner JF, Jones LW. 1979. Establishment and characterization of a human prostatic carcinoma cell line (PC-3). Investagative and Clinical Urology. 17, 16–23.

[22] Kolmogorov AN, Petrovskii PG, Piskunov NS. 1937. A study of the diffusion equation with increase in the amount of substance, and its application to a biological problem. Moscow University Mathematics Bulletin. 1, 1–26.

[23] Laird AK. 1964. Dynamics of tumour growth. British Journal of Cancer. 18, 490–502.

[24] Lagergren JH, Nardini JT, Baker RE, Simpson MJ, Flores KB. 2020. Biologically-informed neural networks guide mechanistic modeling from sparse experimental data. PLoS Computational Biology. 16, e1008462.

[25] Lewis MA, Kareiva P. 1993. Allee dynamics and the spread of invading organisms. Theoretical Population Biology. 43, 141–158.

[26] Maini PK, McElwain DLS, Leavesley DI. 2004. Traveling wave model to interpret a wound-healing cell migration assay for human peritoneal mesothelial cells. Tissue Engineering. 10, 475–482.

[27] McCue SW, Jin W, Moroney TJ, Lo K-Y, Chou S-E, Simpson MJ. 2019. Hole-closing model reveals exponents for nonlinear degenerate diffusivity functions in cell biology. Physica D: Nonlinear Phenomena. 398, 130–140.

[28] Murphy RJ, Maclaren OJ, Simpson M. 2023. Implementing measurement error models for estimation and prediction in the life sciences. Preprint.

[29] Murray JD. 2002. Mathematical biology I: An introduction. Heidelberg: Springer.

[30] Pawitan Y. 2001. In all likelihood: statistical modelling and inference using likelihood. Oxford University Press.

[31] Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller U, Timmer J. 2009. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics. 25, 1923–1929.

[32] Raue A, Kreutz C, Theis FJ, Timmer J. 2013. Joining forces of Bayesian and frequentist methodology: a study for inference in the presence of non-identifiability. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 371, 20110544.

[33] Seber GAF, Wild CJ. 2003. Nonlinear regression. Wiley-Interscience, New Jersey.

[34] Sengers BG, Please CP, Oreffo ROC. 2007. Experimental characterization and computational modelling of two-dimensional cell spreading for skeletal regeneration. Journal of the Royal Society Interface. 4, 1107–1117.

[35] Sherratt JA, Murray JD. 1990. Models of epidermal wound healing. Proceedings of the Royal Society of London Series B. 241, 29–36.

[36] Simpson MJ, Baker RE, Vittadello ST, Maclaren OJ. 2020. Parameter identifiability analysis for spatiotemporal models of cell invasion. Journal of the Royal Society Interface. 17, 20200055.

[37] Simpson MJ, Browning AP, Warne DJ, Maclaren OJ, Baker RE. 2022). Parameter identifiability and model selection for sigmoid population growth models. Journal of Theoretical Biology. 535, 110998.

[38] Simpson MJ, Maclaren OJ. 2023. Profile-Wise Analysis: A profile likelihood-based workflow for identifiability analysis, estimation, and prediction with mechanistic mathematical models. Preprint.

[39] Skellam JG. 1951. Random dispersal in theoretical populations. Biometrika. 38, 196–218.

[40] Steele J, Adams J, Sluckin T. 1998. Modelling paleoindian dispersals. World Archaeology. 30, 286–305.

[41] Swanson KR, Bridge C, Murray JD, Alvord Jr EC. 2003. Virtual and real brain tumors: using mathematical modeling to quantify glioma growth and invasion. Journal of Neurological Sciences. 216, 1–10.

[42] Tsoularis A, Wallace J. 2002. Analysis of logistic growth models. Mathematical Biosciences. 179, 21–55.

[43] Villaverde AF, Tsiantis N, Banga JR. 2019. Full observability and estimation of unknown inputs, states and parameters of nonlinear biological models. Journal of the Royal Society Interface. 16, 20190043.

[44] Villaverde AF, Pathirana D, Fröhlich F, Hasenauer J, Banga JR. 2022. A protocol for dynamic model calibration. Briefings in Bioinformatics. 23, bbab387.

[45] Wasserman L. 2004. All of statistics: a concise course in statistical inference. Springer.