Parameter identifiability and model selection for sigmoid population growth models.

Matthew J Simpson[a,*], Alexander P Browning[a], David J Warne[a,b], Oliver J Maclaren[c], Ruth E Baker[d]

[a]*School of Mathematical Sciences, Queensland University of Technology (QUT), Brisbane, Australia.*
[b]*Centre for Data Science, QUT, Brisbane, Australia.*
[c]*Department of Engineering Science, University of Auckland, Auckland 1142, New Zealand.*
[d]*Mathematical Institute, University of Oxford, Oxford, UK.*

Abstract

Sigmoid growth models, such as the logistic, Gompertz and Richards' models, are widely used to study population dynamics ranging from microscopic populations of cancer cells, to continental-scale human populations. Fundamental questions about model selection and parameter estimation are critical if these models are to be used to make practical inferences. However, the question of parameter identifiability – whether a data set contains sufficient information to give unique or sufficiently precise parameter estimates – is often overlooked. We use a profile-likelihood approach to explore practical parameter identifiability using data describing the re-growth of hard coral. With this approach, we explore the relationship between parameter identifiability and model misspecification, finding that the logistic growth model does not suffer identifiability issues for the type of data we consider whereas the Gompertz and Richards' models encounter practical non-identifiability issues. This analysis of parameter identifiability and model selection is important because different growth models are used within areas of the biological modelling literature without necessarily considering whether parameters are identifiable, or checking statistical assumptions underlying model adequacy. Standard practices that do not consider parameter identifiability can lead to unreliable or imprecise parameter estimates and potentially misleading mechanistic interpretations. While tools developed here focus on three standard sigmoid growth models only, our theoretical developments are applicable to any sigmoid growth model and any relevant data set. MATLAB implementations of all software are available on GitHub.

Keywords: Gompertz growth; identifiability analysis; logistic growth; parameter estimation; Richards growth.

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

1. Introduction

Classical sigmoid growth models play a critical role in many areas of ecology, and population biology [34]. Key features of these models are that they describe: (i) approximately exponential growth at small population densities where competition for resources is relatively weak; and, (ii) saturation effects at larger population densities owing to competition for resources, where the net growth rate decreases to zero as the population density approaches the carrying capacity density [34].

There are many mathematical models of biological growth with a range of resulting sigmoid growth curves [5, 52]. Within a continuum modelling framework, this general class of mathematical models can be written as

$$\frac{\mathrm{d}C(t)}{\mathrm{d}t} = rC(t)f(C), \tag{1}$$

where $C(t) \geq 0$ is the population density, $t \geq 0$ is time, $r > 0$ is the intrinsic, low density growth rate, and $f(C)$ is a crowding function that encodes information about crowding and competition effects that reduce the net growth rate as C increases [24]. There are many potential choices of crowding function, with typical choices of $f(C)$ being decreasing functions, $\mathrm{d}f/\mathrm{d}C < 0$, with $f(K) = 0$, where $K > 0$ is some maximum carrying capacity density. Typical choices include a linear decreasing crowding function, $f(C) = 1 - C/K$, giving rise to the logistic growth model, or a logarithmic decreasing crowding function, $f(C) = \log(K/C)$, giving rise to the Gompertz growth model. Another choice is $f(C) = 1 - (C/K)^{\beta}$, for some constant $\beta > 0$, which gives rise to the Richards' growth model. Key features of the logistic, Gompertz and Richards' growth models are summarised in Figures 1–2. Note that the interpretation of r depends upon the choice of $f(C)$, and this coupling means that we should estimate both r and $f(C)$ simultaneously from data, rather than relying on estimating r from low-density data and then separately estimating $f(C)$ from high-density data [5].

A brief survey of the literature shows that different choices of sigmoid growth functions are routinely used automatically in different areas of application. For example, the logistic growth model is used to study a range of applications, including the dynamics of *in vitro* cell populations [8, 23, 29, 43, 55], as well as various ecological population dynamics across a massive range of scales ranging from jellyfish polyps on oyster shells [32] to continental-scale dynamics of human populations [48]. The Gompertz growth model is routinely used to study *in vivo* tumour dynamics [27, 42]

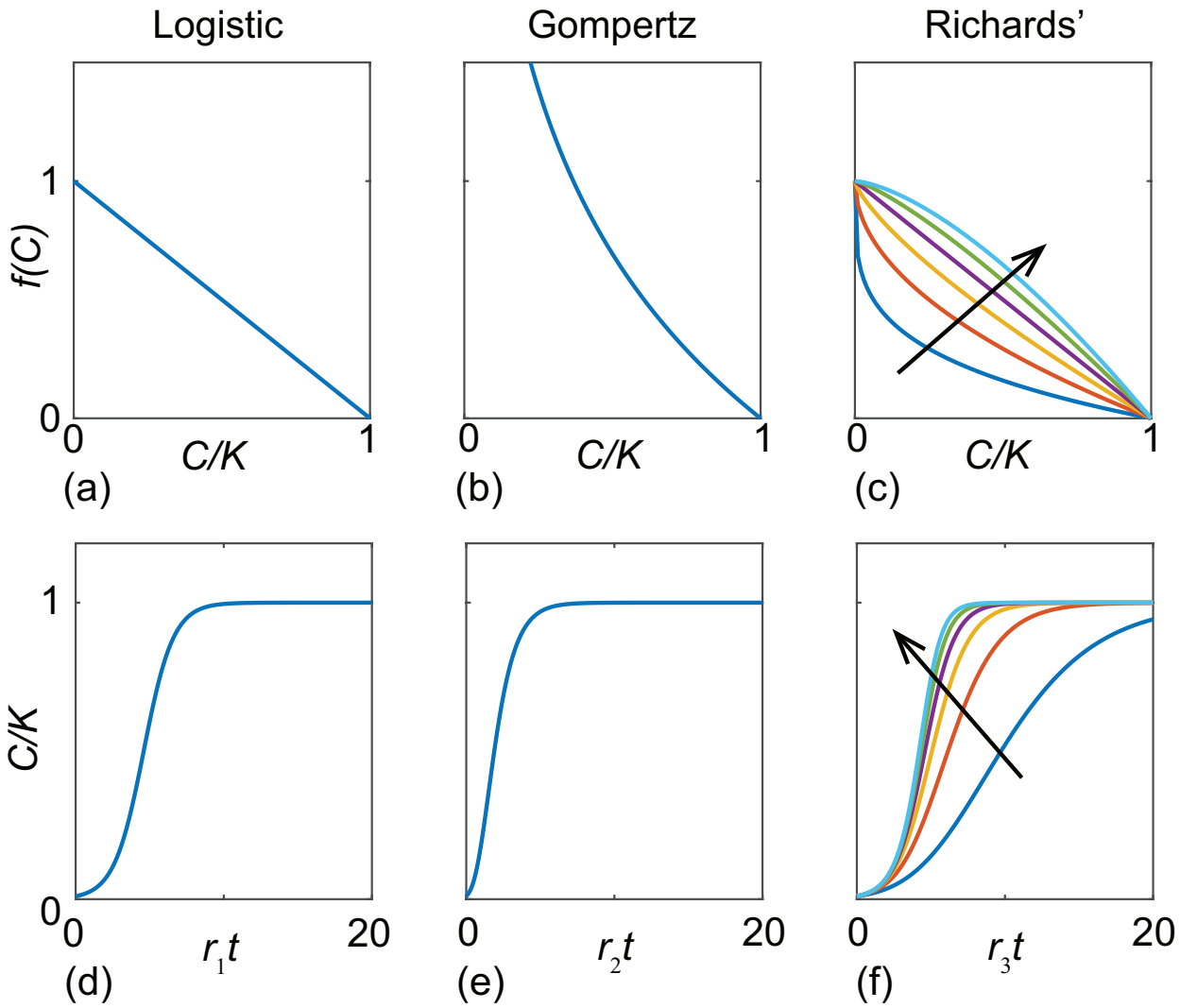

Figure 1: (a)-(c) $f(C)$ as a function of C/K for the logistic, Gomperz and Richards' models, as indicated. (d)-(f) Solutions of the models in (a)-(c), respectively. Curves in (c) and (f) show various profiles for $\beta = 1/4, 1/2, 3/4, 1, 5/4$ and $3/2$, with the direction of increasing β indicated.

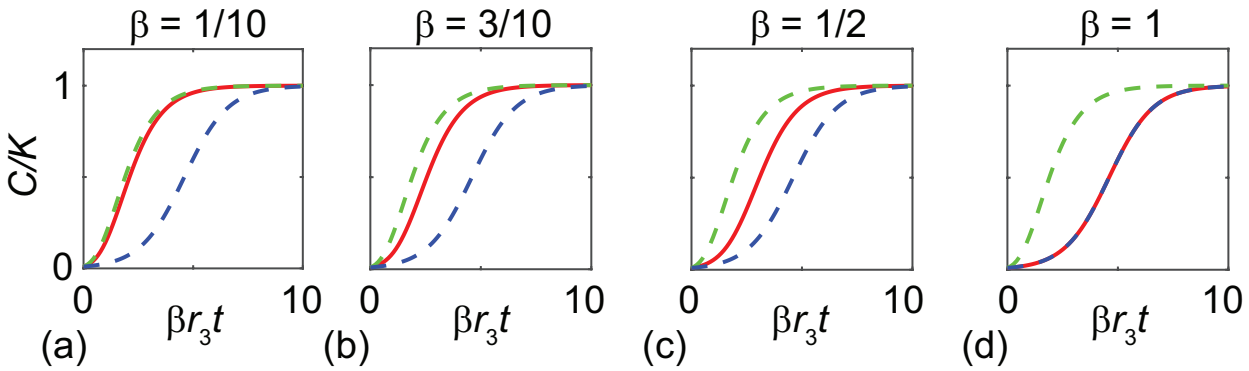

Figure 2: Inter-model comparison. (a)–(d) various solutions of the logistic (blue dashed), Gompertz (green dashed) and Richards' (solid red) models with $\beta = 1/10, 3/10, 1/2$ and 1, as indicated, with $C(0)/K = 1/100$. The Gomperz solution is shown with $r_2 = \beta r_3$ and the logistic solution is shown with $r_1 = r_3$.

as well as various ecological applications such as reef shark population dynamics [20] and dynamics of coral populations [53]. While the logistic and Gompertz models are probably the most commonly-used sigmoid growth models, various extensions have been proposed [17, 50, 52, 58].

Given the importance of sigmoid growth phenomena in ecology and biology, together with the fact that there are a number of different sigmoid growth models used to interrogate and interpret various forms of data, questions of accurate parameter estimation and model selection are crucial to ensure that appropriate mechanistic models are implemented and analysed. Indeed, the question of parameter estimation for sigmoid growth models has been addressed in the nonlinear regression literature for many years [3, 22, 39, 40], however many analyses do not address the limitations of current approaches where different discipline-specific modelling preferences are often invoked without considering other options. For example the Gompertz growth model is routinely used within the coral reef modelling community without necessarily considering other options [53, 57], while the logistic growth model is widespread within the cancer and cell biology community without necessarily considering other options [29, 43]. While such discipline-specific preference do not imply that mathematical models and their parameterisation within those disciplines are incorrect or invalid, working solely within the confines of discipline-specific choices without considering other modelling options could mean that analysts might not be using the most appropriate mathematical model to interpret data and draw accurate mechanistic conclusions.

In this work, we examine model selection for sigmoid growth models through the lens of *parameter identifia-*

bility [2, 10, 30, 37, 38]. A model, considered as an indexed family of distributions, is formally identifiable when distinct parameter values imply distinct distributions of observations, and hence when it is possible to uniquely determine the model parameters using an infinite amount of ideal data [30, 37]. Practical identifiability involves the ability to estimate parameters to sufficient accuracy given finite, noisy data [10, 30, 37]. Methods of identifiability analysis are often used in the systems biology literature where there are many competing models available to describe similar phenomena [16, 33], and these methods provide insight into the trade-off between model complexity and data availability [7]. We also consider the often-neglected question of whether basic statistical assumptions required for the validity of identifiability analysis hold. In particular, we follow the ideas of likelihood-based frequentist inference outlined in [12, 36, 47] and partition our analysis into two types of question: (i) parameter identifiability based on the likelihood function while assuming a statistically adequate model family, and (ii) misspecification checks of basic statistical assumptions underlying model family adequacy.

There is ongoing controversy over the relative merits of various approaches to model selection in the ecological literature. For example, information-theoretic measures such as AIC [1] are increasingly used as alternatives to null hypothesis testing to select between models or carry out model averaging in ecology [9, 13, 25]. However, the underlying justifications for this trend have been criticised [19, 35, 49]. While we also consider the relationships between model complexity and model fit, we take an alternative approach, avoiding relying solely on simplistic null hypothesis testing or single number summaries such as AIC. The two components of our approach – parameter identifiability and model family adequacy – relate to fundamental statistical tasks (parameter estimation and model checking) and are easy to interpret. Furthermore, while measures such as AIC have a predictive focus and interpretation [1, 13], identifiability analysis explicitly focuses on estimating parameter values and understanding underlying mechanisms. Though prediction and estimation are related, there is often a tension between whether to 'explain or predict' [44].

The data we use describes concerns re-growth of hard coral cover on the Great Barrier Reef, Australia, after some external disturbance [15]. Developing our understanding of whether coral communities can re-grow after disturbances, and how long this re-growth process takes, are important questions that need to be addressed urgently as climate change-related disturbances increase in frequency [21]. If, for example, the time scale of re-growth is faster than the time between disturbances we might conclude that re-growth and recovery is likely, however if the time scale of

re-growth is slower than the time between disturbances we might anticipate that re-growth and recovery is unlikely. In this work we consider a data set describing the temporal re-growth of hard coral cover on a reef near Lady Musgrave Island, Australia. Data in Figure 3 shows a typical sigmoid growth response, after some disturbance, such as a tropical cyclone, where we see apparent exponential growth for low coral cover, and the gradual reduction in net growth rate as the coral cover approaches some maximum density. Unlike small-scale laboratory experiments where it is possible to consider data from several identically prepared experiments (e.g. *in vitro* cell biology [56]), it is not possible to average this kind of reef-scale re-growth across many identically prepared experiments. Therefore, here we take the most fundamental approach and work with a single trace data set. Within the coral reef modelling community, it is typical to model this kind of sigmoid curve with a Gompertz growth model [53, 57], but in this work we objectively explore the suitability of a suite of sigmoid growth models, showing that standard approaches that ignore questions of identifiability can produce misleading results. While the focus of this work is to consider a canonical data set together with three standard choices of sigmoid growth model (logistic, Gompertz and Richards' models), the theoretical tools and algorithms developed here can be applied to any sigmoid data set and to any continuum sigmoid growth model. One of the reasons we focus our work on the coral data is that long-term monitoring data successfully reveal the shape of the sigmoid curve, whereas applications such as cell biology often just focus on the early time data before the point of inflection [55].

2. Results and Discussion

2.1. Mathematical models

We consider three mathematical models of population growth:

– *Logistic model*

$$\frac{\mathrm{d}C(t)}{\mathrm{d}t} = r_1 C(t)\left[1 - \left(\frac{C(t)}{K}\right)\right], \quad \text{with solution} \quad C(t) = \frac{KC(0)}{C(0) + (K - C(0))\exp(-r_1 t)}. \tag{2}$$

The logistic model has three free parameters, $\theta = (r_1, K, C(0))$, and a linear crowding function, $f(C) = 1 - C/K$.

– *Gompertz model*

$$\frac{\mathrm{d}C(t)}{\mathrm{d}t} = r_2 C(t)\log\left(\frac{K}{C(t)}\right), \quad \text{with solution} \quad C(t) = K\exp\left[\log\left(\frac{C(0)}{K}\right)\exp(-r_2 t)\right]. \tag{3}$$

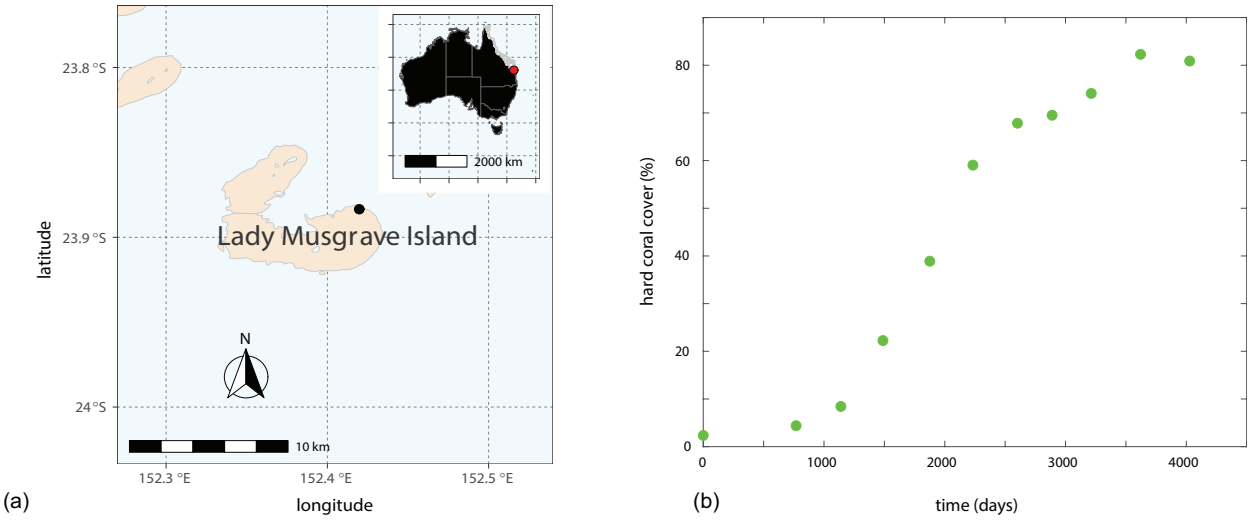

Figure 3: (a) Location of Lady Musgrave Island (black disc) off the East coast of Australia (inset, red disc). (b) Field data showing the % area covered by hard corals (green discs) as a function of time after some external disturbance.

The Gompertz model has three free parameters, $\theta = (r_2, K, C(0))$, and a logarithmic crowding function, $f(C) = \log(K/C)$.

– *Richards' model*

$$\frac{\mathrm{d}C(t)}{\mathrm{d}t} = r_3 C(t)\left[1 - \left(\frac{C(t)}{K}\right)^{\beta}\right], \quad \text{with solution} \quad C(t) = \frac{KC(0)}{\left[C(0)^{\beta} + (K^{\beta} - C(0)^{\beta})\exp(-\beta r_3 t)\right]^{1/\beta}}. \tag{4}$$

The Richards' model has four free parameters, $\theta = (r_3, K, \beta, C(0))$, and a nonlinear crowding function, $f(C) = 1 - (C/K)^{\beta}$, for some constant $\beta > 0$.

For each model the rate parameters have dimensions of $1/\text{day}$, whereas K and $C(0)$ are dimensionless, and given in terms of % of area covered by hard corals. Throughout we carefully chose our nomenclature so that certain variables (e.g. $C(t)$, t) and parameters (e.g. K, $C(0)$) that have the same interpretation across the three models are the same for each model, whereas parameters that do not have a consistent interpretation across the models, such as the growth rates r_m, for $m = 1, 2, 3$, are referred to slightly differently to make this distinction clear.

The crowding functions and solutions of each model are depicted in Figure 1 to make it clear that the question of model selection is subtle, since all three models lead to similar shaped solutions. While the three models are qualitatively similar, they are also quantitatively related as shown in Figure 2 where various solutions with different

β are superimposed. Setting $\beta = 1$ in Figure 2(d) confirms that the Richards' growth model simplifies to the logistic growth model, as expected. Further, comparing solutions in Figure 2(a)–(c) confirms that the solution of the Richards' model approaches the solution of the Gomperz growth model, with $r_2 = \beta r_3$, as $\beta \to 0^+$. These comparisons are instructive since the limiting behaviour of the Richards' model is trivial to establish mathematically [52], but it is not until we visually or numerically compare solutions that we gain a sense of how small β has to be before we can reliably approximate Richards' growth with the simpler Gompertz growth model.

2.2. *Parameter estimation and model checks*

Data in Figure 3(b), denoted y_i^{o}, correspond to measurements at I discrete times, t_i, for $i = 1, 2, 3, \ldots, I$. The data, y_i^{o}, are denoted using a superscript 'o' to distinguish these noisy observations from their modelled counterparts, $y_i = C(t_i \,|\, \theta)$. To estimate θ, we assume the observations are noisy versions of the model solutions, $y_i^{\mathrm{o}} \,|\, \theta \sim \mathcal{N}\left(y_i, \sigma^2\right)$, that is we assume the observation error is additive and normally distributed, with zero mean and a constant variance, σ^2. The constant variance will be estimated along with the other components of θ. To accommodate this we include σ in the vector θ for each model.

We take a likelihood-based approach to parameter inference and uncertainty quantification given a model family. Given a time series of observations, and the above noise assumptions, the log-likelihood function is

$$\ell(\theta \,|\, y_i^{\mathrm{o}}) = \sum_{i=1}^{I} \log\left[\phi\left(y_i^{\mathrm{o}}; y_i(\theta), \sigma^2\right)\right], \tag{5}$$

where $\phi(x; \mu, \sigma^2)$ denotes a Gaussian probability density function with mean μ and variance σ^2. We apply maximum likelihood estimation (MLE) to obtain a best fit set of parameters, $\widehat{\theta}$. The MLE is given by

$$\widehat{\theta} = \sup_{\theta} \left[\ell(\theta \,|\, y_i^{\mathrm{o}})\right], \tag{6}$$

subject to bound constraints: $r_m > 0$ for $m = 1, 2, 3$, $K > 0$, $C(0) > 0$, $\beta > 0$ and $\sigma > 0$. A numerical approximation of $\widehat{\theta}$ is computed using fmincon in MATLAB [31].

Given our estimates of $\widehat{\theta}$, we evaluate each model at the MLE, $\widehat{y_i} = C(t_i \,|\, \widehat{\theta})$ for $i = 1, 2, 3, \ldots, I$, and calculate the scaled-residuals

$$e_i = \frac{y_i^{\mathrm{o}} - \widehat{y_i}}{\widehat{\sigma}}, \quad i = 1, 2, 3, \ldots, I. \tag{7}$$

Figure 4(a)–(c) compares the data with each model evaluated at the MLE. Scaled residuals are given in Figure 4(e)–(f), and Table 1 summarises $\widehat{\theta}$ for each model. Visual checks and the Durbin-Watson test suggest that the scaled residuals are relatively uncorrelated in all cases.

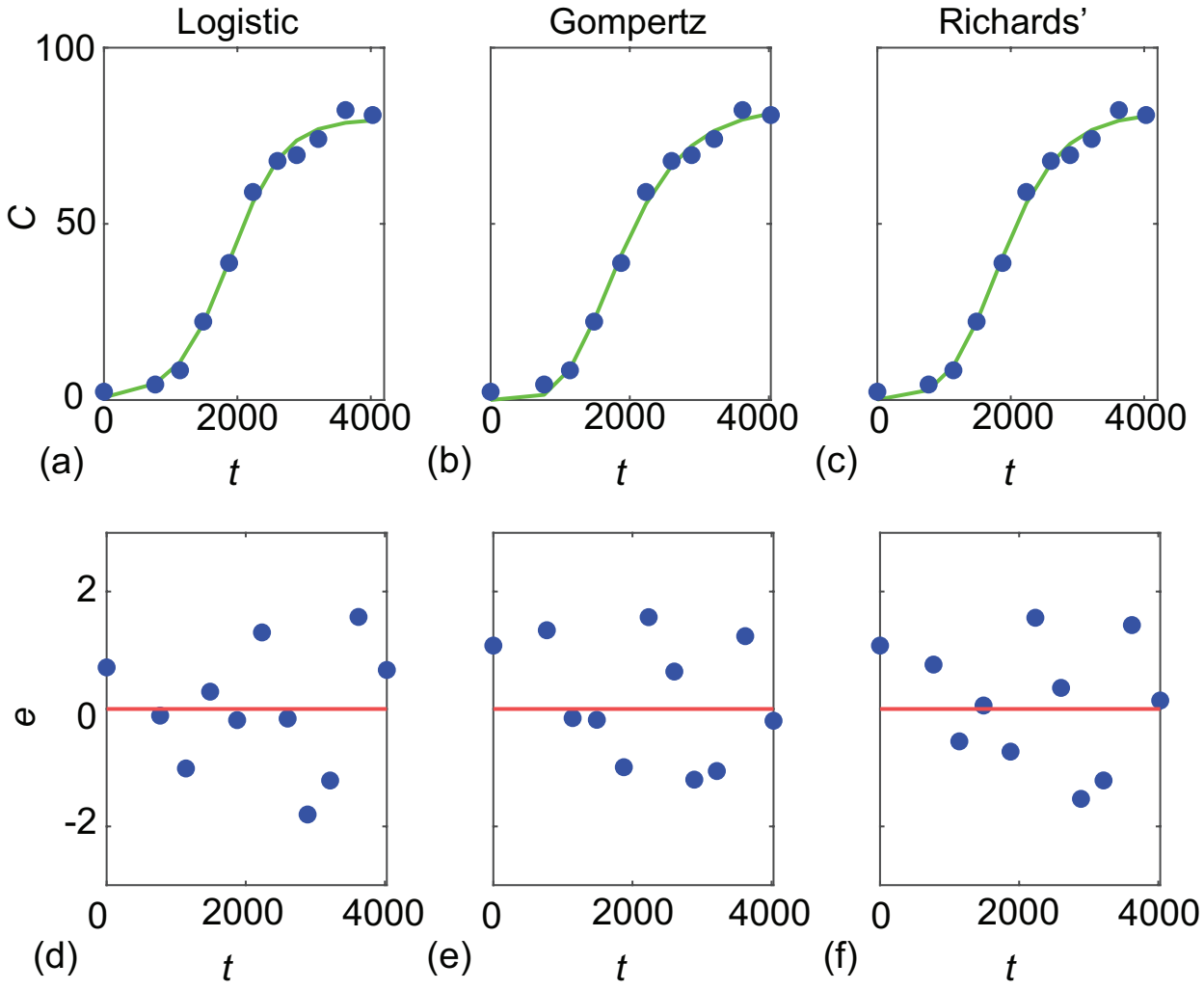

Figure 4: (a)-(c) Observed data (blue discs) superimposed with MLE solutions of the logistic, Gompertz and Richards' models. (d)-(f) scaled residuals for each model. The Durbin-Watson test yields $DW = 1.7624$ ($p = 0.3410$), $DW = 1.9894$ ($p = 0.4927$), and $DW = 2.0073$ ($p = 0.5050$), for the logistic, Gompertz and Richards' models, respectively.

Table 1: MLE parameter estimates with 95% confidence intervals are given in parentheses.

	Logistic ($m = 1$)	Gompertz ($m = 2$)	Richards' ($m = 3$)
$\widehat{r_m}$	2.5×10^{-3} $[2.1 \times 10^{-3}, 3.0 \times 10^{-3}]$	1.6×10^{-3} $[1.3 \times 10^{-3}, 1.9 \times 10^{-3}]$	5.5×10^{-3} $[2.1 \times 10^{-3}, 1.0 \times 10^{-2}]$
$\widehat{K}$	80 [77, 83]	83 [79, 88]	82 [78, 87]
$\widehat{C(0)}$	0.7 [0.3, 1.4]	1.1×10^{-4} $[0, 1.1 \times 10^{-2}]$	9.4×10^{-2} [0, 1.0]
$\widehat{\beta}$	-	-	0.3 [0.1, 1.0]
$\widehat{\sigma}$	2.3 [1.6, 3.7]	2.7 [1.5, 3.5]	2.1 [1.4, 3.4]

The quality of the fit to the data is excellent for each model in Figure 4(a)–(c), and visualising the scaled residuals in Figure 4(d)–(f) is consistent with the standard assumption that the scaled residuals are independent and identically distributed. The apparent uncorrelated structure of these residuals is consistent with the Durbin-Watson test. Here, simply visualising the scaled residuals confirms that standard statistical assumptions inherent in our estimation of $\widehat{\theta}$ are reasonable. However, this insightful step is often overlooked, and we will return to it later.

At this point, confronted with the data-model comparisons in Figure 4 it is not obvious which model is preferable, but it is worthwhile noting some of the quantitative consequences of working with these three models. Estimates of $\widehat{r_m}$ provide biological insight since the reciprocal of the growth rate, $1/r_m$ for $m = 1, 2, 3$, provides an estimate of the time scale of re-growth to be 400, 625, and 183 days for the logistic, Gompertz and Richards' models, respectively. While these estimates are all of the same order of magnitude, the range of 183–625 days is relatively broad if we wanted to estimate of how long we would have to wait to observe the re-growth process. If we had taken the more standard approach of interpreting this data with just one model, we would not have any insight into this variability and so this simple exercise demonstrates the importance of addressing model selection, which we now explore through the lens of parameter identifiability.

2.3. *Identifiability analysis*

In the first instance we consider the *structural identifiability* of the three models in the hypothetical situation where we have access to and infinite amount of ideal, noise-free data [16, 33]. GenSSI software [11, 28] confirms that all three sigmoid growth models are structurally identifiable. This result is very insightful since it means that identifiability issues relate to the question of *practical identifiability* which deals with the more usual scenario of working with finite, noisy data, as in Figure 3(b) [37, 38, 45].

We use a profile likelihood-based approach to explore practical identifiability. In all cases we work with a normalised log-likelihood function

$$\widehat{\ell}(\theta \,|\, y_i) = \ell(\theta \,|\, y_i) - \sup_{\theta} \ell(\theta \,|\, y_i), \tag{8}$$

which we consider as a function of θ for fixed data, y_i, $i = 1, 2, 3, \ldots, I$.

We assume the full parameter θ can be partitioned into an *interest* parameter ψ and *nuisance* parameter λ so that

we write $\theta = (\psi, \lambda)$. Given a set of data, y_i, the profile log-likelihood for the interest parameter ψ can be written as

$$\widehat{\ell}_{\mathrm{p}}(\psi \,|\, y_i) = \sup_{\lambda} \ell(\theta \,|\, y_i). \tag{9}$$

In Equation (9) λ is optimised out for each value of ψ, and this implicitly defines a function $\lambda^*(\psi)$ of optimal λ values for each value of ψ. For example, in the logistic growth model with $\theta = (r_1, K, C(0), \sigma)$ we may consider the growth rate as the interest parameter, so that $\psi(\theta) = r_1$. The remaining parameters are nuisance parameters, $\lambda(\theta) = (K, C(0), \sigma)$, so that

$$\widehat{\ell}_{\mathrm{p}}(r_1 \,|\, y_i) = \sup_{[K, C(0), \sigma]} \ell(r_1, K, C(0), \sigma \,|\, y_i). \tag{10}$$

In all cases we implement this optimisation using the fmincon function in MATLAB with the same bound constraints used to calculate $\widehat{\theta}$ [31]. For each value of the interest parameter, taken over a sufficiently fine grid, the nuisance parameter is optimised out and the previous optimal value is used as the starting estimate for the next optimisation problem. Uniformly spaced grids of 100 points, defined on problem-specific intervals, are reported throughout.

Assuming an adequate model family, the likelihood function represents the information about θ contained in y_i^{o}, and the relative likelihood for different values of θ indicates the relative evidence for these parameter values [47]. A flat profile indicates non-identifiability and the degree of curvature is related to the inferential precision [36, 47]. We form approximate likelihood-based confidence intervals by choosing a threshold-relative profile log-likelihood value; for univariate profiles thresholds of -1.92 correspond to approximate 95% confidence intervals [36, 41], and the points of intersection are determined using linear interpolation. Following the likelihood-based frequentist ideas of [12, 47], we supplement the likelihood inferences about parameters given a model structure by checks of model family adequacy, here based on residual checks. We use both visual inspection and the classical Durbin-Watson test [14] to assess residual correlation, noting that we have a simple trend model and relatively small sample sizes.

Figure 5 shows various univariate profiles for each mathematical model. Each profile is superimposed with a vertical line at the MLE, and a horizontal line at -1.92 so we can visualise and calculate the width of the confidence intervals reported in Table 1. Univariate profiles in Figure 5 provide information about the practical identifiability of each model with the same data. Results for the logistic growth model indicate that each parameter is identifiable with relatively narrow profiles (Table 1). In contrast, the Gompertz growth model results show that r_2 and K are

identifiable, but $C(0)$ shows signs of practical non-identifiability. In particular, the MLE and upper confidence bound for $C(0)$ are very near the lower edge of feasibility, and the lower confidence bound is only determined by the $C(0) > 0$ constraint (indicating that this may be at best one-sided identifiable). Profiles for the Richards' growth model indicate that while K is identifiable, the remaining parameters are not since r_3 is one-sided identifiable, and profiles for $C(0)$ and β are relatively flat, with confidence intervals determined by the *a priori* bounds rather than data.

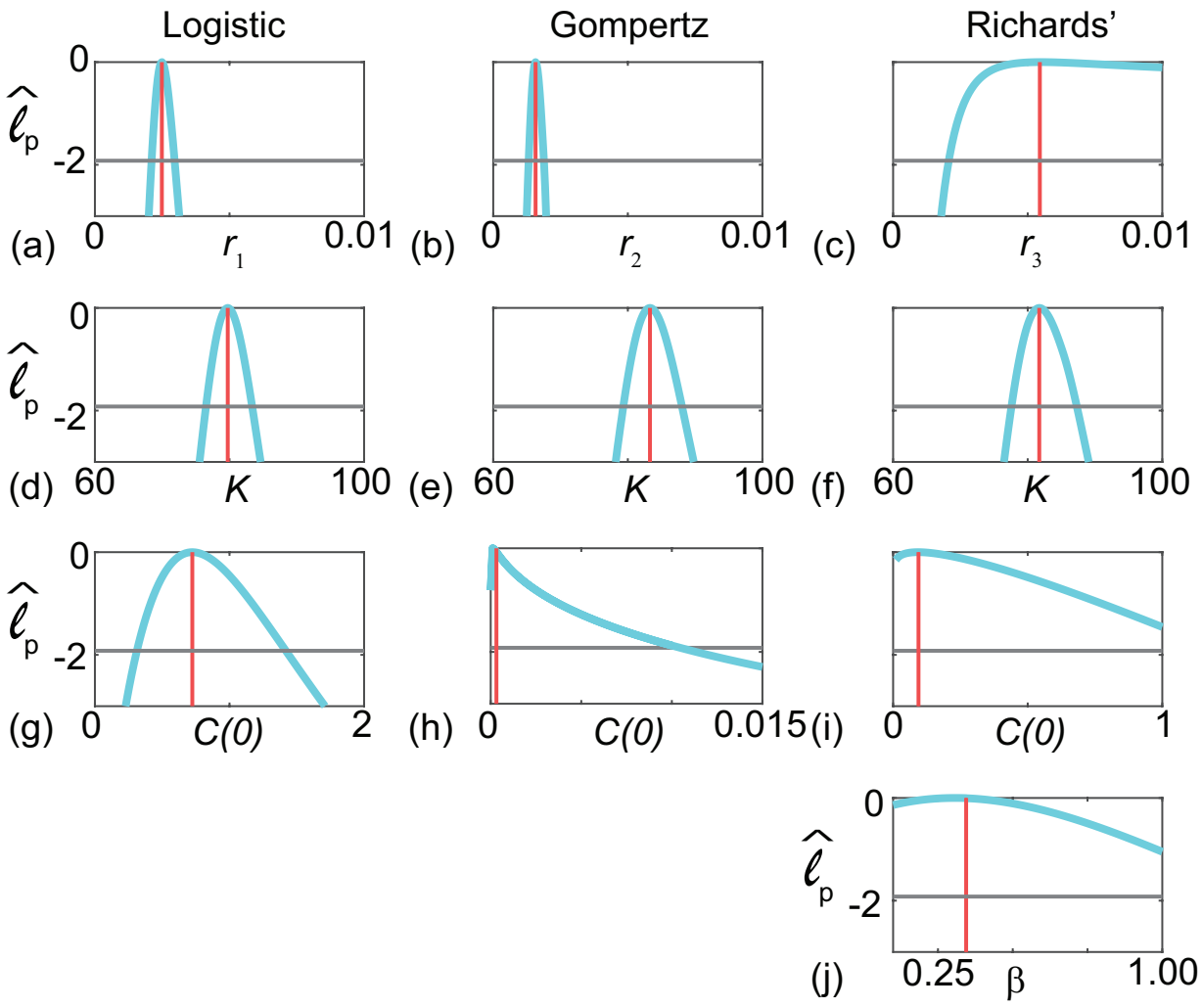

Figure 5: Univariate profile likelihoods for parameters in the logistic, Gompertz and Richards' model, as indicated. In each case the profile is shown (solid blue) together with the MLE (vertical red) and a horizontal line at -1.92.

Bivariate profiles can be used to provide further insight into the identifiability issues with the Richards' model. In particular, we consider bivariate profile likelihoods by taking the full parameter vector, $\theta = (r_3, K, \beta, C(0), \sigma)$, and

treating $\psi(\theta) = (r_3, \beta)$ and $\lambda(\theta) = (K, C(0), \sigma)$, allowing us to evaluate

$$\widehat{\ell}_{\mathrm{p}}(r_3, \beta \,|\, y_i) = \sup_{[K, C(0), \sigma]} \ell(r_3, K, \beta, C(0), \sigma \,|\, y_i), \tag{11}$$

which again we compute using fmincon [31] on a uniform mesh of pairs of the interest parameter, and optimise out the nuisance parameters at each mesh point. Using uniformly-spaced grids of 100×100 across the interval of interest, and a threshold of -3.00 to give the approximate 95% confidence regions [36, 41] we generate the bivariate profile in Figure 6(a) where the solid grey lines define the boundary of the 95% confidence region, which, suggests there are many combinations of β and r_3 that match the data. Finally, we now consider the product βr_3 as an interest parameter, and univariate profiles for this product in Figure 6(b) suggests that our data is sufficient to obtain precise estimates of the product βr_3, but does not identify β and r_3 individually.

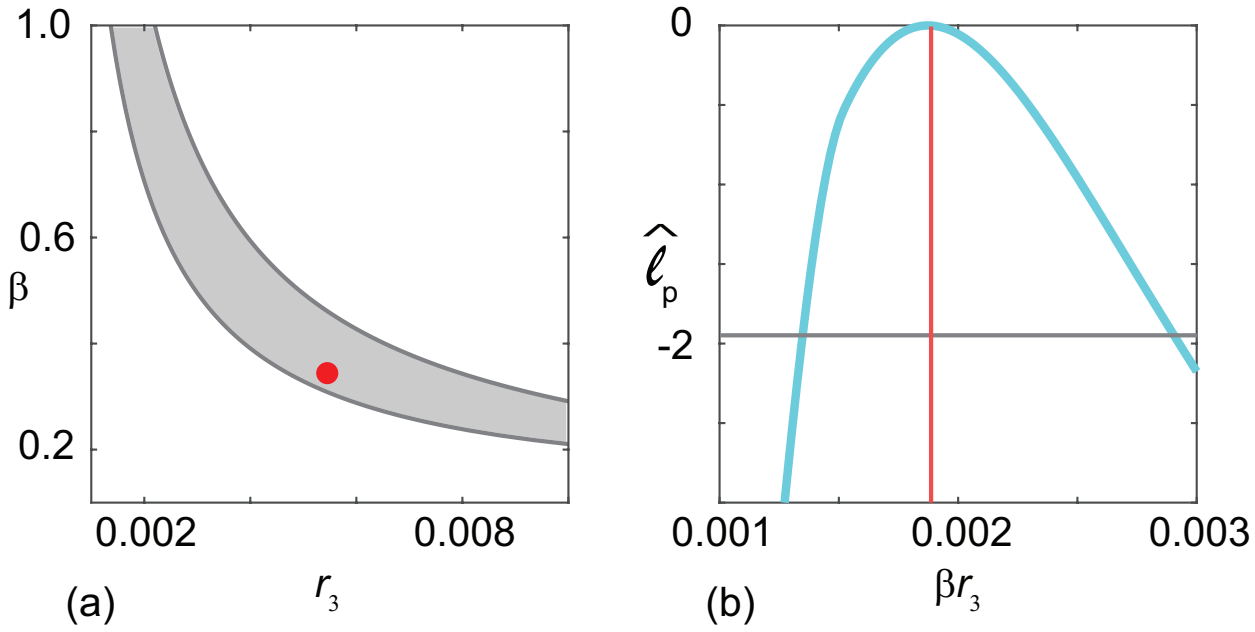

Figure 6: (a) Bivariate profiles for β and λ_3. (b) Univariate profile for the product βr_3. The MLE is $\widehat{\beta r_3} = 0.0019\,[0.0014, 0.0029]$.

In summary, comparing the profiles between different models in Figure 5 suggests that this data leads to identifiable parameter estimates for the logistic growth model, but that the two most straightforward extensions of that model, namely the Gompertz and Richards' models, suffer from practical identifiability issues. Therefore, we interpret these results to mean that the most reliable way to interpret this data is with the logistic growth model. While this approach to model selection using parameter identifiability is not routinely considered, it provides a straightforward,

computationally-efficient means of informing quantitative distinctions between model suitability. We note, however, that difficulties with identifiability do not mean that these models are wrong, simply that they are difficult to reliably estimate parameters for, given the type of data available. This motivates our next investigation.

2.4. *Parameter estimation and model checks with fixed $C(0)$*

Our observation that $C(0)$ for the Gompertz model (Figure 5(h)) is not well-identified by the data prompts us to consider a further set of results where we adopt the standard practice of estimating θ under the assumption that variability in $C(0)$ is neglected, and is treated as a known quantity given by the first observation, $C(0) = y_1^{\mathrm{o}}$ [53, 57]. Figure 7(a)–(c) compares the data with the model evaluated at the MLE with $C(0) = y_1^{\mathrm{o}}$, and the scaled residuals are shown in Figure 7(e)–(f). A summary of the MLE estimates for each model are given in Table 2. The Durbin-Watson test indicates that the scaled residuals are more correlated when $C(0)$ is fixed.

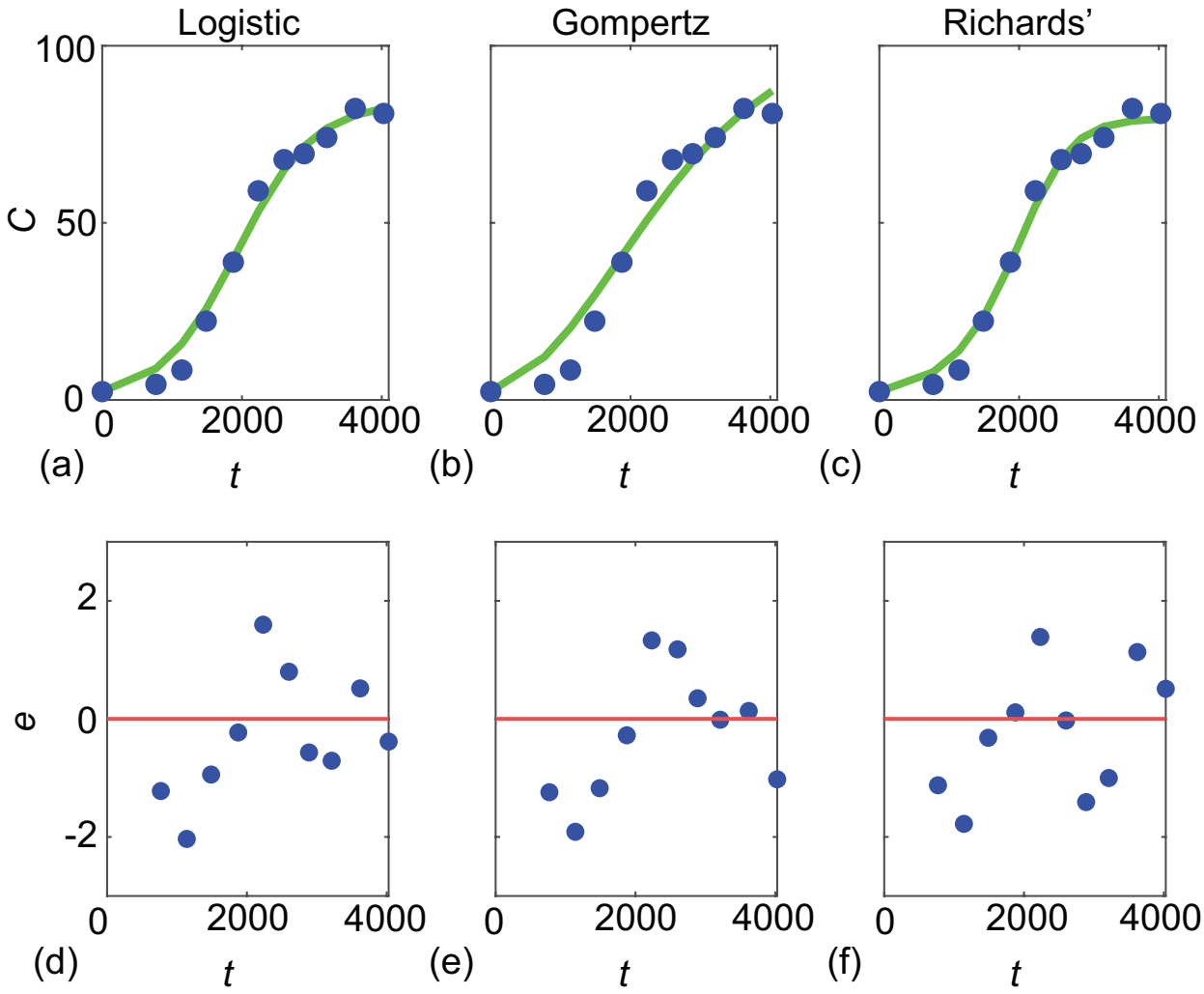

Figure 7: (a)-(c) Observed data (blue discs) superimposed with MLE solutions of the logistic, Gompertz and Richards' models with $C(0) = y_1^o$. (d)-(f) scaled residuals for each model. The Durbin-Watson test yields $DW = 1.0544$ ($p = 0.04841$), $DW = 0.6397$ ($p = 0.0045$), and $DW = 1.2919$ ($p = 0.1138$), for the logistic, Gompertz and Richards' models, respectively.

Table 2: Parameter estimates with fixed $C(0) = y_1^o$, 95% confidence intervals are given in parentheses.

	Logistic ($m = 1$)	Gompertz ($m = 2$)	Richards' ($m = 3$)
$\widehat{r_m}$	1.8×10^{-3} $[1.7 \times 10^{-3}, 2.0 \times 10^{-3}]$	7.0×10^{-4} $[5.9 \times 10^{-4}, 9.0 \times 10^{-4}]$	1.6×10^{-3} $[1.5 \times 10^{-3}, 1.9 \times 10^{-3}]$
$\widehat{K}$	84 [79, 90]	106 [88,138]	80 [75, 86]
$\widehat{\beta}$	-	-	1.8 [1.0, 3.9]
$\widehat{\sigma}$	3.7 [2.5, 5.9]	6.3 [4.3, 10.2]	3.1 [2.1, 5.0]

The quality of the model-data match in Figure 7(a)–(c) is clearly poorer than when we simultaneously estimate $C(0)$ (Figure 4), and in particular we see that the MLE solutions systematically overestimate the data at early times, and then underestimate the data at later times. These trends are clear in the scaled residuals in Figure 7(d)–(f) that are visually correlated. The Durbin-Watson test indicates that the scaled residuals are more correlated when $C(0)$ is fixed. This is particularly true for the Gompertz model where we see the potential for drawing erroneous conclusions about the underlying biological mechanisms since $\widehat{r_2}$ differs by an order of magnitude depending on whether we treat $C(0)$ as a fixed or variable quantity. The estimate of the time scale of re-growth is 625 days in the case where we estimate $C(0)$ from the data whereas it is 1429 days variability in $C(0)$ is neglected. This exercise clearly illustrates the danger in adopting discipline-specific procedures without carefully considering alternative approaches.

2.5. *Profiling misspecified models*

To conclude we present some further synthetic data exploration to mimic the commonly-encountered scenario where we have sparse, noisy data generated by a relatively complicated process, which we then interpret with a simpler mathematical model. Data in Figure 8(a)–(d) show solutions of the Richards' growth model over a range of $\beta \in [1/10, 3/2]$, where we consider the time interval $t \in [0, T]$ where T is defined implicitly by $C(T) = 0.999K$. Noisy data are generated by taking 20 equally-spaced temporal observations within this interval and corrupting the data by adding normally distributed noise with zero mean and constant variance, σ^2. Data in Figure 8(a)–(d) are generated by setting $r_3 = 0.0055$, $K = 80$, $C(0) = 1$ and $\sigma = 2$, which are consistent with the MLE in Table 1.

To explore how well we can interpret noisy data from the relatively complicated Richards' growth model with the simpler logistic and Gompertz growth models we estimate $\widehat{\theta}$ and calculate univariate profiles for r_1 and r_2. Scaled residuals for the logistic and Gompertz models in Figure 8(e)–(h) and Figure8(i)–(l), respectively, where we see no obvious visual evidence of correlation. Taking the noisy data from the Richards' growth model and constructing univariate profiles for r_1 and r_2 leads to well-shaped, relatively narrow profiles in Figure 8(m)–(p). These profiles again highlight certain relationships between the Richards', logistic and Gompertz growth models established in Figure 4. In particular, for sufficiently small β in Figure 8(m) we see the expected relationship between the Richards' growth model and the Gompertz growth model since the MLE for r_2 is very close to the expected limiting behaviour, $r_2 = \beta r_3$ as $\beta \to 0^+$. Furthermore, not only does the profile match the expected point estimate, but we see that the

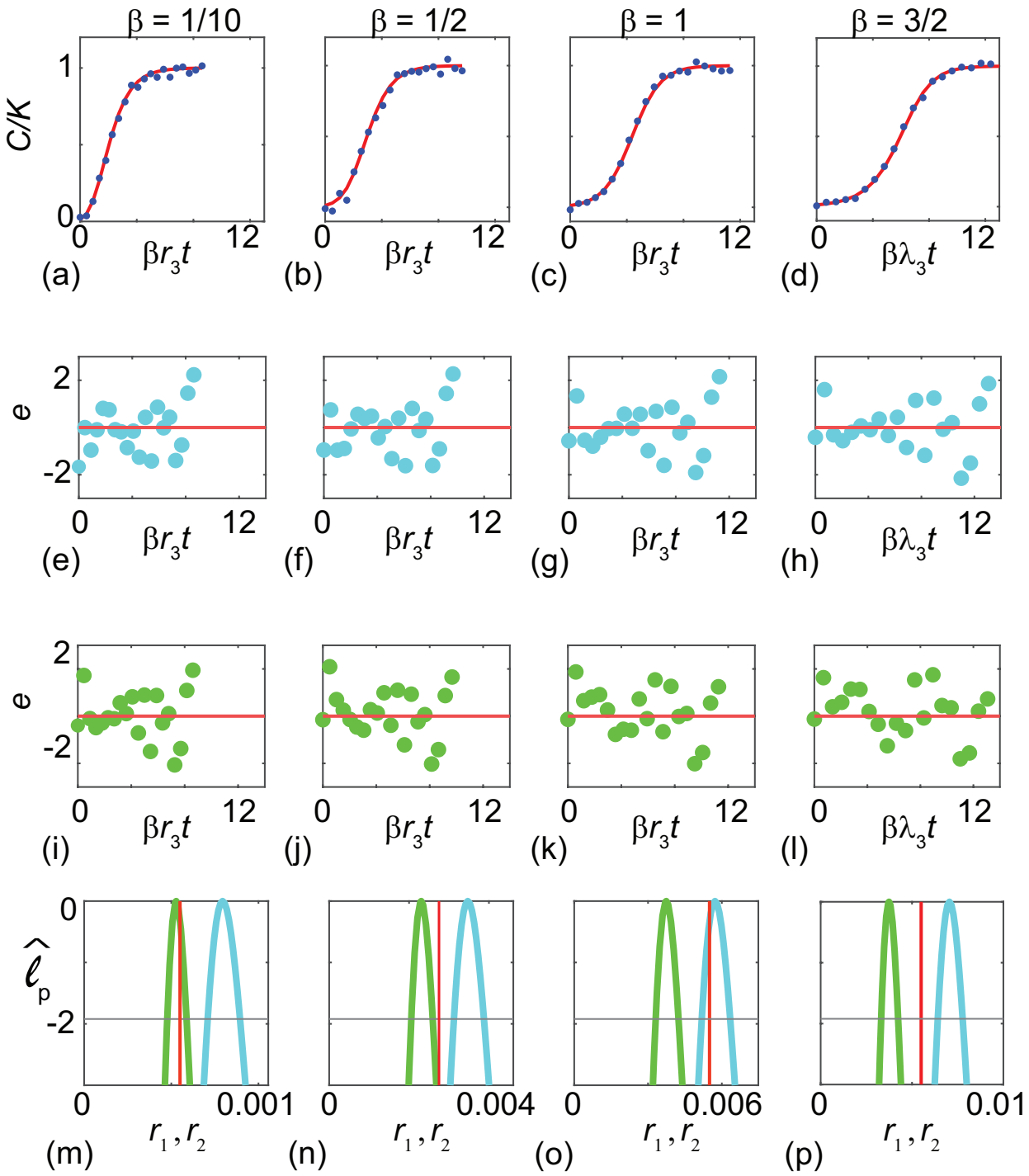

Figure 8: (a)–(d) Solution of the Richards' growth curves (solid red) for the values of β indicated with $C(0) = 1$, $r_3 = 0.0055$ and $K = 80$. Noisy data (blue dots) are obtained by sampling 20 equally-spaced observations in the interval $0 < t < T$ and adding normally distributed noise with zero mean and $\sigma = 2$. Here T is defined implicitly by $C(T) = 0.999K$. For each data we estimate $\widehat{\theta}$ using the methods in the main paper (Figure 4, Table 1). Scaled residuals for the MLE using the logistic growth model are given in (e)–(h) and for the Gompertz growth model are given in (i)–(l). Univariate profiles for r_1 (blue) and r_2 (green) are given in (m)–(p). Subfigures (m)–(p) contain a red vertical line at $\beta r_3 = 0.0055\beta$.

width of the profile is relatively narrow, implying that it is possible to recover relatively precise and interpretable parameter estimates working with a misspecified model. Similarly, when working with the special case $\beta = 1$ in Figure 8(o) we see the profile for r_1 for the logistic growth model is peaked close to the expected result, $r_1 = r_3$. Again, this profile relatively narrow, implying that it is possible to obtain meaningful and interpretable parameter estimates when working with a misspecified model. The remaining profiles in Figure 8(m)–(p) lead to well-shaped profiles with relatively narrow confidence intervals. Generating data from either the logistic or Gompertz models at the MLE for r_1 and r_2, respectively, leads to relatively uncorrelated residuals and an excellent match to the data in Figure 8(a)–(d) (not shown).

We now further explore the extent to which we are able to estimate parameters under model misspecification by repeating the calculation of $\widehat{\theta}$ and univariate profiles for r_1 and r_2 using data from the Richards' growth model by considering 30 equally-spaced values of $\beta \in [1/10, 3/2]$. We repeat this exercise using both relatively noisy data with $\sigma = 2$, and with noise-free data, $\sigma = 0$. For each set of synthetic data set we generate $\widehat{\theta}$ and compute the quantities $\beta\widehat{r_3} - \widehat{r_1}$ and $\beta\widehat{r_3} - \widehat{r_2}$, that are plotted in Figure 9(a) for the logistic and Gompertz models. As before in Figure 8, the quantity $\beta\widehat{r_3} - \widehat{r_2}$ approaches zero as β becomes sufficiently small, while the quantity $\beta\widehat{r_3} - \widehat{r_1}$ is approximately zero when $\beta = 1$, as expected. Plotting these quantities with noise-free data in Figure 9(a) illustrates how the MLE for $\widehat{r_1}$ and $\widehat{r_2}$ varies with β without the influence of external noise. Analogous data in Figure 9(b) shows the influence of corrupting the data with some noise and we see that the same underlying trends in the quantities $\beta\widehat{r_3} - \widehat{r_2}$ and $\beta\widehat{r_3} - \widehat{r_1}$ as functions of β are modestly impacted by incorporating noise in to the data.

We now quantify the precision of the estimates of r_1 and r_2 in terms of the width of the 95% confidence interval, w. Data in Figure 9(c) shows how the width of the profiles varies with β for noise-free data, $\sigma = 0$, where we see the width of the profile for the logistic model approaches zero when $\beta = 1$, while the width of the profile for the Gompertz growth model approaches zero as β becomes sufficiently small, as expected. Additional results in Figure 9(d) show how the addition of noise to the data affects the width of the profiles, where we see that the clear trends in Figure 9(c) are completely obscured by the incorporation of noise.

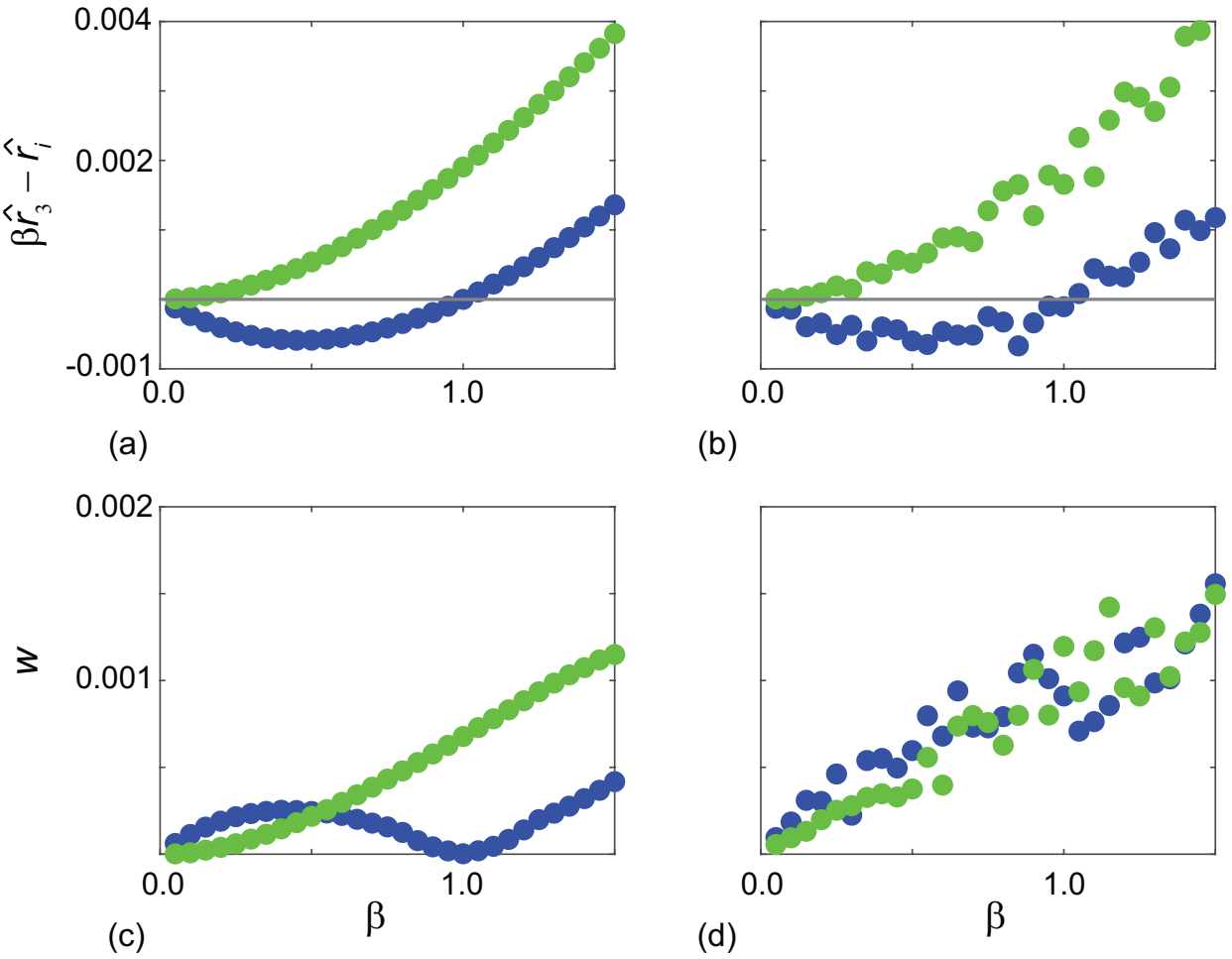

Figure 9: (a)–(b) show quantities $\beta\widehat{r_3} - \widehat{r_m}$ for $m = 1, 2$ as functions of β for noise free data ($\sigma = 0$) and noisy data ($\sigma = 2$), respectively. (c)–(d) show the width of the profiles, w, plotted as functions of β for noise free data ($\sigma = 0$) and noisy data ($\sigma = 2$), respectively. In all cases results for the logistic growth model are shown in blue while results for the Gompertz growth model are shown in green.

3. Conclusion and Future Work

In this work we present a general framework that can aid model selection by comparing different models through the lens of practical parameter identifiability and checks of statistical assumptions underlying valid identifiability analysis. Population biology is one of many fields where questions of model selection are extremely important, with the potential to be very challenging because it is possible to interpret data with more than one candidate model [18]. In practice it is often unclear which model is most appropriate to work with and so developing methodologies that can assist in supporting these decisions is valuable.

Various approaches are adopted to deal with model selection in the literature, from the very simplest and surprisingly common approach of neglecting the question completely, to using more developed ideas such as information criteria [1, 4, 9, 25, 44, 56] and Bayesian approaches using Bayes factors [6, 26, 51]. Many of these established approaches, however, are difficult to interpret inferentially [13] and do not explicitly consider the fundamental question of practical identifiability where we assess whether the data contains sufficient information to form precise parameter estimates, and whether this varies between different candidate models.

For our particular data, we show that a simple sigmoid-growth predicted by several candidate mathematical models without any obvious challenges at first. Only when we ask the question whether the parameters are identifiable does it becomes clear that some of these models are very attractive while others are not. In this case we show that typical data relating to sigmoid re-growth of hard corals can be modelled using logistic, Gompertz and Richards' models. However, both the Gompertz and Richards' models encounter identifiability issues, whereas there is no such issue with the logistic model. While this finding is consistent with our intuitive expectation that increasingly complicated models ought to be used only when there is sufficient data available to estimate the unknown parameters, the difference in identifiability is somewhat surprising. For example, the logistic growth model with four unknown quantities $\theta = (r_1, K, C(0), \sigma)$ does not lead to any identifiability issues, whereas the more commonly-invoked Gompertz growth model with the same number of unknowns, $\theta = (r_2, K, C(0), \sigma)$, encounters practical identifiability issues since estimates of $C(0)$ are not identifiable. This difference is surprising given that these two models involve the same number of unknowns, and even more surprising in this context since the Gompertz growth model is routinely used in the study of coral reef re-growth, but often this model is used without any explicit consideration of parameter

identifiability [53, 57].

As we demonstrated, attempts to simplify parameter estimation by the common but ad-hoc practice fixing of initial conditions introduces separate issues with the statistical assumptions underlying valid identifiability analysis, as indicated by residual checks. This suggests that great care ought to be taken when we interpret experimental data with a mathematical model since the temptation that more complex models lead to better insight is not necessarily true. We suggest that the adoption of practical identifiability analysis along with model checks (e.g. inspection of residuals) is a very straightforward way to help guide model selection, and thereby aid in understanding underlying biological mechanisms. While our work here focuses on three simple models and one sigmoid data set, our approach is broad and applies to any continuum sigmoid model and any appropriate sigmoid data set. Further, our approach of selecting a model based upon the degree to which parameters are identifiable, provided it also passes statistical misspecification checks, has a strong implication for the question of experimental design, since we can use the tools presented in this work to explore what additional data is required to improve the identifiability of the parameters [55]. We leave the question of experimental design for future consideration.

Acknowledgements MJS is supported by the Australian Research Council (DP200100177). OJM received support from the University of Auckland, Faculty of Engineering James and Hazel D. Lord Emerging Faculty Fellowship. REB acknowledges the Royal Society for a Wolfson Research Merit Award, and the BBSRC (BB/R000816/1).

Author Contributions All authors conceived ideas the designed methodology, MJS developed software and led the writing of the manuscript. All authors analysed the data and contributed critically to the drafts and gave final approval for publication.

Data Availability Data and software are available on GitHub, and the complete LTMP data set is available at eAtlas.

References

[1] Akaike H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 19, 716-723. (doi:10.1109/TAC.1974.1100705).

[2] Audoly S, Bellu G, D'Angiò L, Saccomani MP, Cobelli C. 2001. Global identifiability of nonlinear models of biological systems. IEEE Transactions on Biomedical Engineering. 48. 55-65. (doi:10.1109/10.900248).

[3] Bates DM, Watts DG. 1988. Nonlinear regression analysis and its applications. John Wiley & Sons.

[4] Bortz DM, Nelson PW. 2006. Model selection and mixed-effects modeling of HIV infection dynamics. Bulletin of Mathematical Biology. 68, 2005–2025. (doi: 10.1007/s11538-006-9084-x).

[5] Browning AP, McCue SW, Simpson MJ. 2017. A Bayesian computational approach to explore the optimal duration of a cell proliferation assay. Bulletin of Mathematical Biology. 79. 1888-1906. (doi: 10.1007/s11538-017-0311-4).

[6] Browning AP, Jin W, Plank MJ, Simpson MJ. 2020. Identifying density-dependent interactions in cell populations. Journal of the Royal Society Interface. 17. 20200143. (doi: 10.1098/rsif.2020.0143).

[7] Browning AP, Warne DJ, Burrage K, Baker RE, Simpson MJ (2020). Identifiability analysis for stochastic differential equations models in systems biology. Journal of the Royal Society Interface. 17, 20200652. (doi:10.1098/rsif.2020.0652).

[8] Browning AP, Maclaren OJ, Burnzli PR, Lanaro M, Allenby MC, Woodruff MA, Simpson MJ (2021). Model-based data analysis of tissue growth in thin 3D printed scaffolds. Joutnal of Theoretical Biology. 528, 110852. (doi:10.1016/j.jtbi.2021.110852).

[9] Burnham KP, Anderson DR, Huyvaert KP. 2011. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behavioral Ecology and Sociobiology. 65(1):23-35. (doi: 10.1007/s00265-010-1029-6).

[10] Chiş O, Banga JR, Balsa-Canto E. 2011. Structural identifiability of systems biology models: a critical comparison of methods. PLoS One. 6. e27755. (doi: 10.1371/journal.pone.0027755).

[11] Chiş O, Banga JR, Balsa-Canto E. 2011. GenSSI: a software toolbox for structural identifiabilty analysis of biological models. Bioinformatics. 18. 2610–2611. (doi: 10.1093/bioinformatics/btr431).

[12] Cox DR. 2006. Principles of statistical inference. Cambridge University Press.

[13] Dormann CF, Calabrese JM, Guillera-Arroita G, Matechou E, Bahn V, Bartoń K, Beale CM, Ciuti S, Elith J, Gerstner K, Guelat J. 2018. Model averaging in ecology: A review of Bayesian, information-theoretic, and tactical approaches for predictive inference. Ecological Monographs. 88(4):485-504. (doi: 10.1002/ecm.1309).

[14] Durbin J, Watson GS. 1950. Testing for serial correlation in least squares regression: I. Biometrika. 37(3/4):409-28.

[15] eAtlas: Largest GBR coral reef survey data repository. Retrieved October 2021 eAtlas.

[16] Eisenberg MC, Hayashi MAL. 2014. Determining identifiable parameter combinations using subset profiling. Mathematical Biosciences. 256. 115-126. (doi: 10.1016/j.mbs.2014.08.008).

[17] Gerlee P. 2013. The model muddle: In search of tumor growth laws. Cancer Research. 73. 2407-2411. (doi: 10.1158/0008-5472.CAN-12-4355).

[18] Gutenkunst R, Waterfall J, Casey F, Brown K, Myers C, Sethna J. 2007. Universally sloppy parameter sensitivities in systems biology models. PLoS Computational Biology. 3. e189. (doi: 10.1371/journal.pcbi.0030189).

[19] Guthery FS, Brennan LA, Peterson MJ, Lusk JJ. 2005. Information theory in wildlife science: critique and viewpoint. The Journal of Wildlife Management. 69(2):457-65. (doi:).

[20] Hisano M, Connolly SR, Robbins WD. 2011. Population growth rates of reef sharks with and without fishing on the Great Barrier Reef: Robust estimation with multiple models. PLoS One. 6. e25028. (doi: 10.1371/journal.pone.0025028).

[21] Hughes TP, Kerry JT, Baird AH, Connolly SR, Chase TJ, Dietzel A, Hill T, Hoey AS, Hoogenboom MO, Jacobson M, Kerswell A, Madin JS, Mieog A, Paley AS, Pratchett MS, Torda G, Woods RM. 2019. Global warming impairs stock-recruitment dynamics of corals. Nature. 568. 387-390. (doi: 10.1038/s41586-019-1081-y).

[22] Huet S, Bouvier A, Poursat M-A, Jolivet S. 2004. Statistical tools for nonlinear regression. Second edition. Springer, New York.

[23] 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. (doi:10.1016/j.jtbi.2015.10.040).

[24] Jin W, Penington CJ, McCue SW, Simpson MJ. 2016. Stochastic simulation tools and continuum models for describing two-dimensional collective cell spreading with universal growth functions. Physical Biology. 13, 056003. (doi: 10.1088/1478-3975/13/5/056003).

[25] Johnson, JB, Omland, KS. 2004. Model selection in ecology and evolution. Trends in Ecology & Evolution. 19(2), 101–108. (doi: 10.1016/j.tree.2003.10.013).

[26] Kass R, Raftery A. 1995. Bayes factors. Journal of the American Statistical Association. 90, 773-795. (doi: 10.1080/01621459.1995.10476572).

[27] Laird AK. 1964. Dynamics of tumour growth. British Journal of Cancer. 18. 490–502. (doi: 10.1038/bjc.1964.55).

[28] Ligon TS, Frölich F, Chiş O, Banga JR, Balsa-Canto E, Hasenauer J. 2018. GenSSI 2.0: multi-experimental structural identifiability analysis of SBML models. Bioinformatics. 34. 1421–1423. (doi: 10.1093/bioinformatics/btx735).

[29] 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. (doi:10.1089/107632704323061834).

[30] Maclaren OJ, Nicholson R. 2020. What can be estimated? Identifiability, estimability, causal inference and ill-posed inverse problems. arXiv preprint. (arXiv:1904.02826).

[31] fmincon: Find minimum of constrained nonlinear multivariable function. Retrieved October 2021 fmincon.

[32] Melica V, Invernizzi S, Caristi G. 2014. Logistic density-dependent growth of *Aurelia aurita* polyps population. Ecological Modelling. 10. 1-5. (doi: 10.1016/j.ecolmodel.2014.07.009).

[33] Meshkat N, Sullivant S, Eisenberg M. 2018. Identifiability results for several classes of linear compartment models. Bulletin of Mathematical Biology. 8. 1620-1651. (doi: 10.1007/s11538-015-0098-0).

[34] Murray JD. 2002. Mathematical biology I: An introduction. Heidelberg: Springer. (doi: 10.1007/978-3-662-08539-4).

[35] Murtaugh PA. In defense of P values. Ecology. 2014. 95(3):611-7.

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

[37] 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(15). 1923-1929. (doi: 10.1093/bioinformatics/btp358).

[38] 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. (doi: 10.1098/rsta.2011.0544).

[39] Ritz C, Streibig JC. 2008. Nonlinear regression with R. Springer, New York.

[40] Ross GJS. 1990. Nonlinear estimation. Springer, New York.

[41] Royston P. 2007. Profile likelihood for estimation and confidence intervals. The Stata Journal. 7, 376-387. (doi: 10.1177/1536867X0700700305).

[42] Sarapata EA, de Pillis LG. 2014. A comparison and cataolg of intrinsic tumor growth models. Bulletin of Mathematical Biology. 76. 2010-2024. (doi: 10.1007/s11538-014-9986-y).

[43] Sherratt JA, Murray JD. 1990. Models of epidermal wound healing. Proceedings of the Royal Society B: Biological Sciences. 241. 29–36. (doi: 10.1098/rspb.1990.0061).

[44] Shmueli G. 2010. To explain or to predict? Staistical Science. 25. 289–310. (doi: 10.1214/10-STS330).

[45] 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. (doi: 10.1098/rsif.2020.0055).

[46] Simpson MJ, Browning AP, Drovandi C, Carr EJ, Maclaren OJ, Baker RE. 2021. Profile likelihood analysis for a stochastic model of diffusion in heterogeneous media. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences. 477. 20210214. (doi: 10.1098/rspa.2021.0214).

[47] Sprott, DA. 2008. Statistical inference in science. Springer Science & Business Media.

[48] Steele J, Adams J, Slukin T. 1998. Modelling paleoindian dispersals. World Archaeology. 30. 286-305. (doi: 10.1080/00438243.1998.9980411).

[49] Stephens PA, Buskirk SW, Hayward GD, Martinez Del Rio C. Information theory and hypothesis testing: a call for pluralism. Journal of Applied Ecology. 42(1):4-12. (doi: 10.1111/j.1365-2664.2005.01002.x)

[50] Tjørve E, Tjørve KMC. 2010. A unified approach to the Richards-model family for use in growth analyses: why we need only two model forms. Journal of Theoretical Biology. 3. 417-425. (doi: 10.1016/j.jtbi.2010.09.008).

[51] Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH. 2009. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of the Royal Society Interface. 6, 187-120. (doi: 10.1098/rsif.2008.0172).

[52] Tsoularis A, Wallace J. 2002. Analysis of logistic growth models. Mathematical Biosciences. 179. 21-55. (doi: 10.1016/S0025-5564(02)00096-2).

[53] Thompson A, Martin K, Logan M. 2020. Development of the coral index, a summary of coral reef resilience as a guide for management. Journal of Environmental Management. 271. 111038. (doi: 10.1016/j.jenvman.2020.111038).

[54] 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(156). 20190043. (doi: 10.1098/rsif.2019.0043).

[55] Warne DJ, Baker RE, Simpson MJ. 2017. Optimal quantification of contact inhibition in cell populations. Biophysical Journal. 113. 1920-1924. (doi: 10.1016/j.bpj.2017.09.016).

[56] Warne DJ, Baker RE, Simpson MJ. 2018. Using experimental data and information criteria to guide model selection for reaction-diffusion problems in mathematical biology. Bulletin of Mathematical Biology. 81. 1760-1804. (doi: 10.1007/s11538-019-00589-x).

[57] Warne DJ, Crossman KA, Jin W, Mengersen K, Osborne K, Simpson MJ, Thompson AA, Wu P, Ortiz J-C. (2021). Identification of two-phase recovery for interpretation of coral reef monitoring data. Journal of Applied Ecology. To appear, 28-October-2021. (doi: 10.1111/1365-2664.14039).

[58] West GB, Brown JH, Enquist BJ. 2001. A general model for ontogenetic growth. Nature. 413. 628-631. (doi: 10.1038/35098076).

[59] Wieland F-G, Hauber AL, Rosenblatt M, Tönsing C, Timmer J. 2021. On structural and practical identifiability. Current Opinion in Systems Biology. 25. 60–69. (doi: doi.org/10.1016/j.coisb.2021.03.005).