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

1. Sigmoid growth models, such as the logistic and Gompertz growth models, are widely used to study various population dynamics ranging from microscopic populations of cancer cells, to continental-scale human populations. Fundamental questions about model selection and precise parameter estimation are critical if these models are to be used to make useful inferences about underlying ecological mechanisms. However, the question of parameter identifiability for these models – whether a data set contains sufficient information to give unique or sufficiently precise parameter estimates for the given model – is often overlooked;
2. We use a profile-likelihood approach to systematically explore practical parameter identifiability using data describing the re-growth of hard coral cover on a coral reef after some ecological disturbance. The relationship between parameter identifiability and checks of model misspecification is also explored. We work with three standard choices of sigmoid growth models, namely the logistic, Gompertz, and Richards' growth models;
3. We find 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, even with relatively-extensive data where we observe the full shape of the sigmoid growth curve. Identifiability issues with the Gompertz model lead us to consider a further model calibration exercise in which we fix the initial density to its observed value, neglecting its uncertainty. This is a common practice, but the results of this exercise suggest that parameter estimates and fundamental statistical assumptions are extremely sensitive under these conditions;
4. Different sigmoid growth models are used within subdisciplines within the biology and ecology literature without necessarily considering whether parameters are identifiable or checking statistical assumptions underlying model family adequacy. Standard practices that do not consider parameter identifiability can lead to unreliable or imprecise parameter estimates and hence potentially misleading interpretations of the underlying mechanisms of interest. While tools in this work focus on three standard sigmoid growth models and one particular data set, our theoretical developments are applicable to any sigmoid growth model and any relevant data set. MATLAB implementations of all software available on GitHub.

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

1. Introduction

Classical sigmoid growth models play a critical role in many areas of ecology, and population biology [18, 30, 38]. 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 [18, 30, 38].

There are many mathematical models of biological growth with a range of resulting sigmoid growth curves [6, 58]. 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 [26]. 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 [6, 58]. 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 [6].

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 [33, 48, 62], as well as various ecological population dynamics across a massive range of scales that include populations of jellyfish polyps [36], field-scale dynamics of forest regeneration [45], as well as continental-scale dynamics of human populations [53]. The Gompertz growth model is routinely used to study *in vivo* tumour dynamics [31, 47] as well as various ecological applications such as reef shark population dynamics [24], dynamics of populations of corals on coral reefs [41, 59, 60], and infectious disease dynamics in human populations [4]. While the logistic and Gompertz models are probably the most commonly-used sigmoid growth models, various extensions of these models have also been proposed and implemented [58], and often these extensions involve working with more complicated choices of $f(C)$ [20, 55, 56, 64], or extending these simple single population models to deal with communities of interacting subpopulations [27].

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

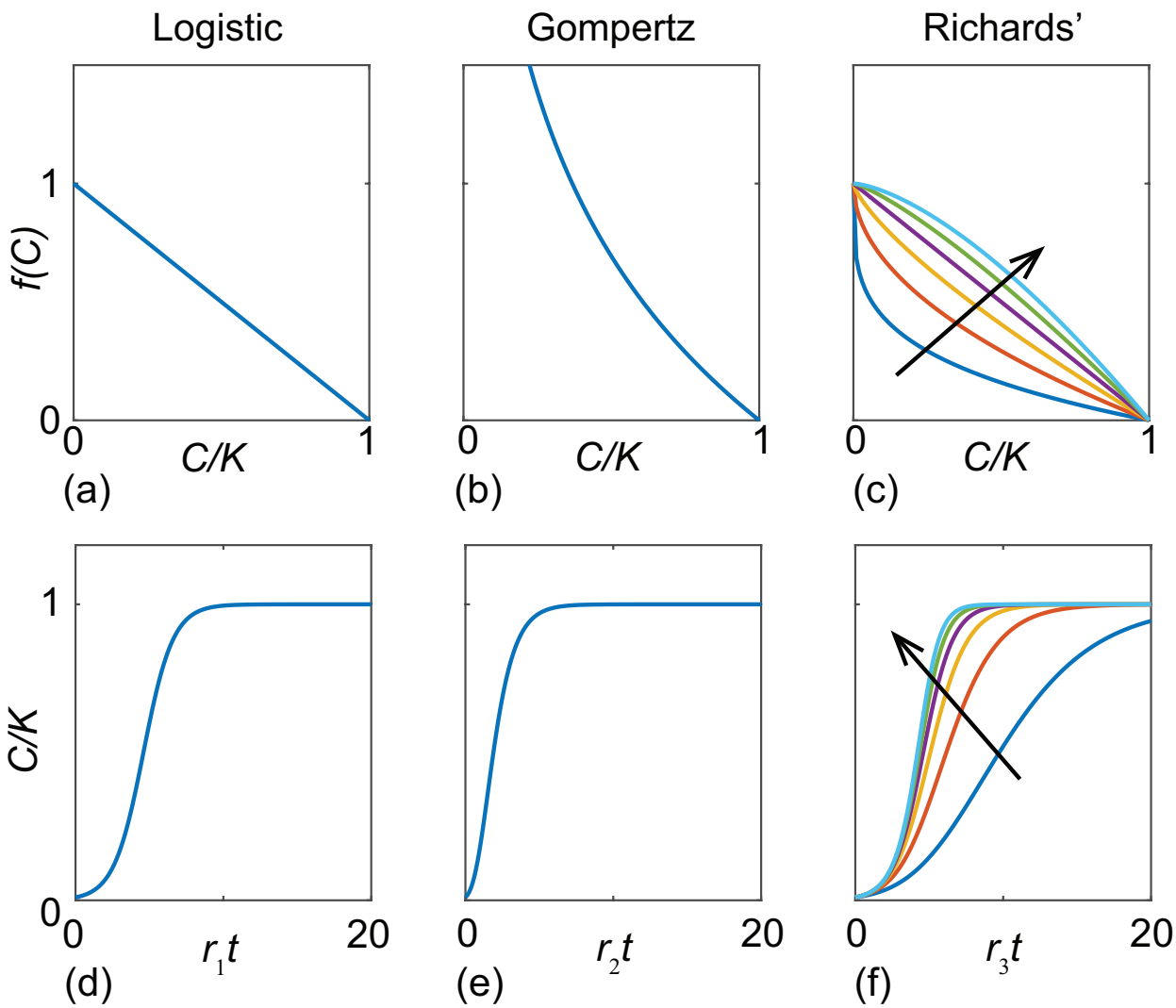

Figure 1: (a)-(c) $f(C)$ as a function of C/K for the logistic, Gomperz and Richards' models, as indicated. (d)-(f) Solutions of the models in (a)-(c), respectively. Curves in (c) and (f) show various profiles for $\beta = 1/4, 1/2, 3/4, 1, 5/4$ and $3/2$, with the direction of increasing β indicated.

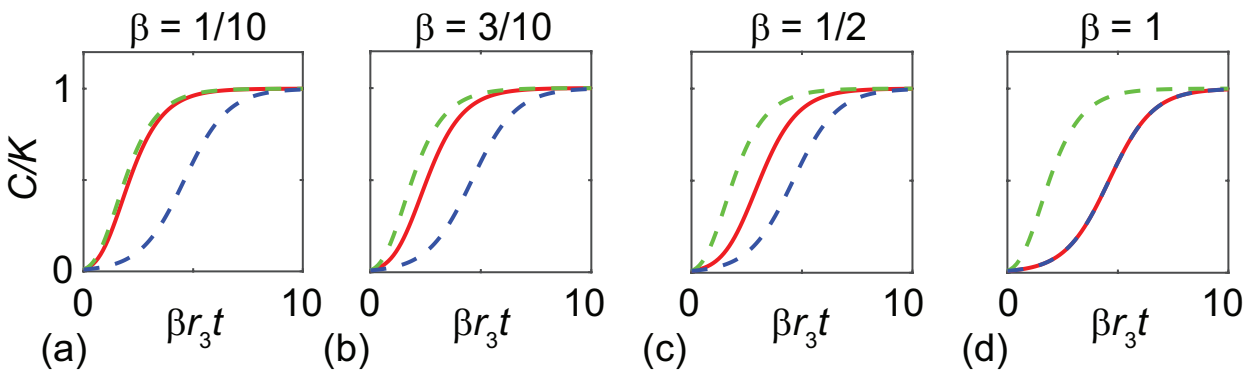

Figure 2: Inter-model comparison. (a)–(d) various solutions of the logistic (blue dashed), Gompertz (green dashed) and Richards' (solid red) models with $\beta = 1/10, 3/10, 1/2$ and 1, as indicated, with $C(0)/K = 1/100$. The Gomperz solution is shown with $r_2 = \beta r_3$ and the logistic solution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

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 [3, 10, 13] and model selection [55, 56, 63] are crucial to ensure that appropriate mechanistic models are implemented and analysed. These questions are particularly important given there seems to be different discipline-specific preferences, for example the Gompertz growth model is routinely used within the coral reef modelling community without necessarily considering other options [41, 59, 60], while the logistic growth model is widespread within the cancer and cell biology community [33, 48]. 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 systematically 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 identifiability* [2, 11, 34, 43, 44]. 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 [34, 43]. Practical identifiability involves the ability to estimate parameters to sufficient accuracy given finite, noisy data [11, 34, 43]. Methods of identifiability analysis are often used in the systems biology literature where there are many competing models available to describe similar phenomena [19, 37], and these methods provide insight into the trade-off between model complexity and data availability [8]. 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 [14, 42, 52] 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.

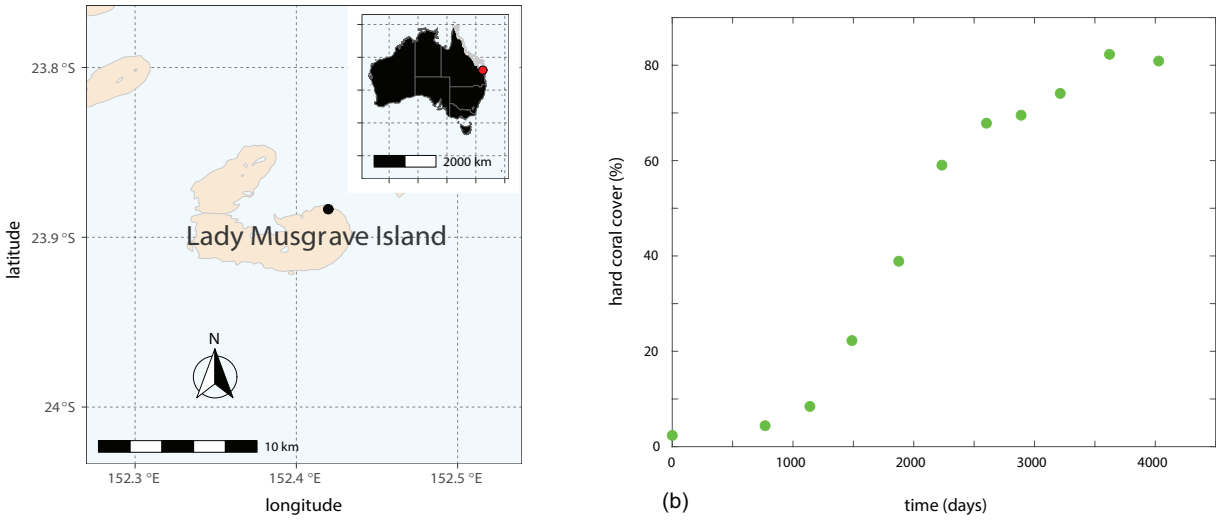

Figure 3: (a) Location of Lady Musgrave Island (black disc) off the East coast of Australia (inset, red disc). (b) Field data showing the % area covered by hard corals (green discs) as a function of time after some external disturbance.

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, 15, 28]. However, the underlying justifications for this trend have been criticised [23, 39, 54]. 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, 15], 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' [49].

The data set we use in this work concerns the re-growth of hard coral cover on the Great Barrier Reef, Australia, after some external disturbance [17]. 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 [25, 40]. 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 (e.g. 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 [63]), it is not possible to average this kind of reef-scale re-growth data 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 [41, 59, 60], 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' growth models), the theoretical tools and algorithms developed in this work 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 other applications such as cell biology often just focus on the early time data before the point of inflection [62]. MATLAB software to implement all algorithms presented in this work are available on GitHub.

2. Materials and methods

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*

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

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 $\beta > 0$.

Our nomenclature has been carefully chosen 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. In contrast, 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.

2.2. Model calibration

Data in Figure 3(b) are denoted y_i^{o}, and 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 given by

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

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 [35]. In all cases we first estimate $\widehat{\theta}$ using a local search option, and then double check those results using a global search option [35] to ensure we estimate $\widehat{\theta}$ accurately. In most cases there is no difference between the local and global search results.

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

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

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 [19, 37]. GenSSI software [12, 32] confirms that all three sigmoid growth models are structurally identifiable. This result is very insightful since it means that any identifiability issues are related to the question of *practical identifiability* which deals with the more usual scenario of working with finite, noisy data, as in Figure 3(b) [43, 44, 50].

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 $\widehat{\theta}$ [35]. 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 [52]. A flat profile indicates non-identifiability and the degree of curvature is related to the inferential precision [42, 52]. 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 [42, 46], and the points of intersection are determined using linear interpolation. Following the likelihood-based frequentist ideas of [14, 52], 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 [16] to assess residual correlation, noting that we have a simple trend model and relatively small sample sizes.

Most of this work considers univariate profiles, but we find that bivariate profiles provide additional information in cases where we have identifiability issues. For instance, we also consider bivariate profile likelihoods for the Richards' model 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 with bound constraints [35] on a uniform mesh of pairs of the interest parameter, and we optimise out the nuisance parameters at each mesh point. We use uniformly-spaced grids of 100×100 across the interval of interest, and thresholds of -3.00 to give the approximate 95% confidence regions [42, 46] that are determined by MATLAB contouring software.

3. Results

3.1. Comparison of mathematical models

Key features of the logistic, Gompertz and Richards' models are compared in Figure 1, and relationships between the models are highlighted in Figure 2.

3.2. Parameter estimation and model checks

Figure 4(a)–(c) compares the data with each model evaluated at the MLE, with the scaled residuals in Figure 4(e)–(f), and a summary of $\widehat{\theta}$ in Table 1. Visual checks and the Durbin-Watson test suggest that the scaled residuals are relatively uncorrelated in all cases. Note that the rate parameters have dimensions of 1/day, whereas K and $C(0)$ are in terms of % of area covered by hard corals.

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

	Logistic ($m = 1$)	Gompertz ($m = 2$)	Richards' ($m = 3$)
$\widehat{r_m}$	0.0025 [0.0021, 0.0030]	0.0016 [0.0013, 0.0019]	0.00545 [0.0021, 0.01]
$\widehat{K}$	79.74 [76.51, 83.37]	83.29 [79.33, 88.05]	81.72 [77.59, 87.30]
$\widehat{C(0)}$	0.7237 [0.3121, 1.4184]	1.1167×10^{-4} [0, 0.0105]	0.0943 [0, 1.0]
$\widehat{\beta}$	-	-	0.3442 [0.1, 1.0]
$\widehat{\sigma}$	2.301 [1.5945, 3.7370]	2.7187 [1.5102, 3.5389]	2.0916 [1.4534, 3.3975]

3.3. Identifiability analysis

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 profile likelihoods in Figure 5(c) and (j) indicate that r_3 and β are not well-identified by the data for the Richards' growth model. Figure 6(a) further explores this by showing the bivariate profile, Equation (11), where the solid grey lines indicate the $\ell_p = -3.0$ contours that define the 95% confidence region. We also re-parameterise the Richards' model and calculate the univariate profile for the product βr_3 in Figure 6(b).

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

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.

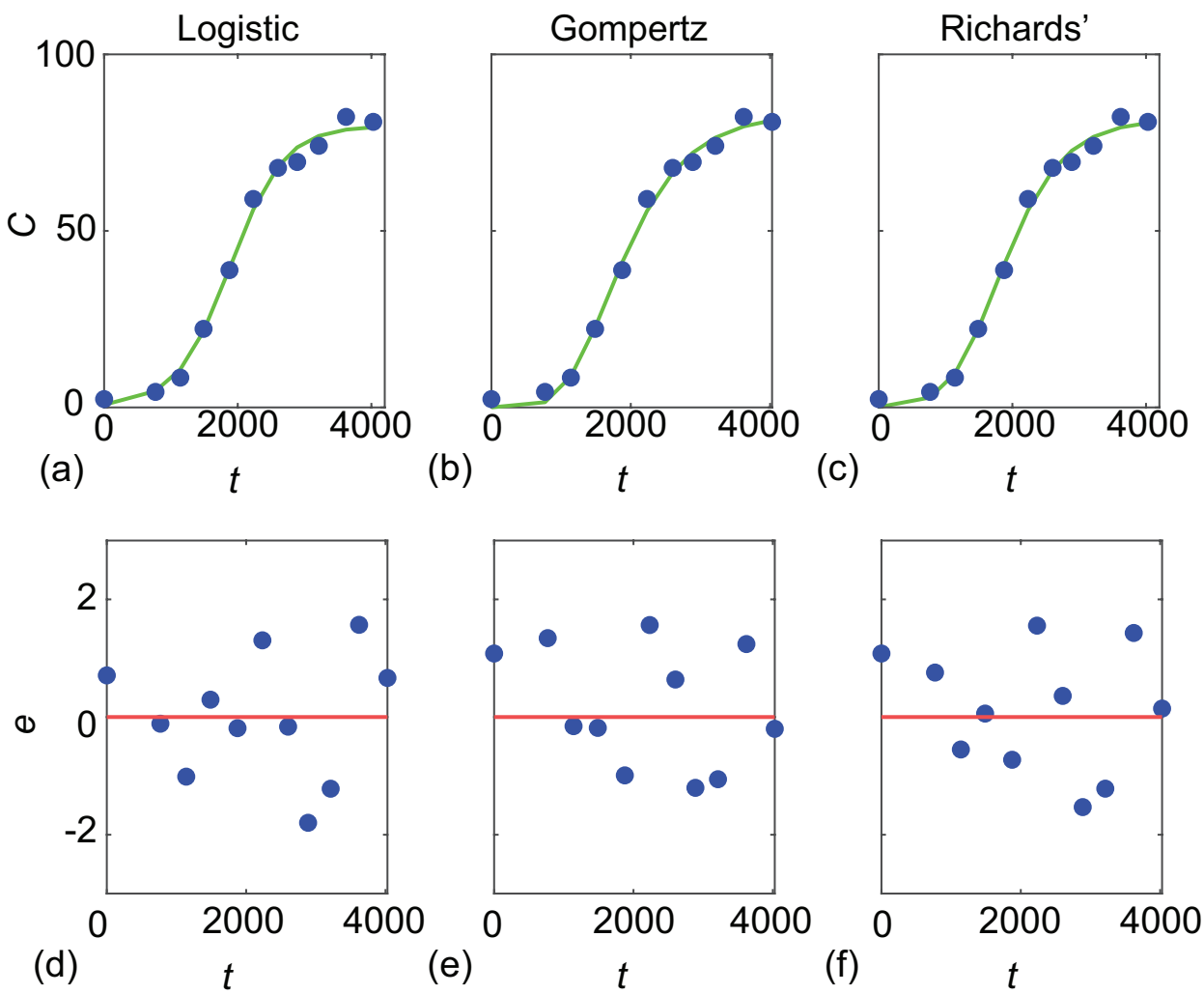

Figure 4: (a)-(c) Observed data (blue discs) superimposed with MLE solutions of the logistic, Gompertz and Richards' models. (d)-(f) scaled residuals for each model. The Durbin-Watson test yields $DW = 1.7624$ ($p = 0.3410$), $DW = 1.9894$ ($p = 0.4927$), and $DW = 2.0073$ ($p = 0.5050$), for the logistic,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

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

	Logistic ($m = 1$)	Gompertz ($m = 2$)	Richards' ($m = 3$)
$\widehat{r_m}$	0.0018 [0.0017,0.0020]	0.0007 [0.0005882,0.0008997]	0.0016 [0.0015,0.0019]
$\widehat{K}$	84.07 [78.8140, 89.8558]	106.4 [87.7981,137.8404]	79.5373 [74.7953, 85.7590]
$\widehat{\beta}$	-	-	1.8476 [0.9886, 3.8755]
$\widehat{\sigma}$	3.6545 [2.5330, 5.9359]	6.2547 [4.3364, 10.1598]	3.0840 [2.1392, 5.0096]

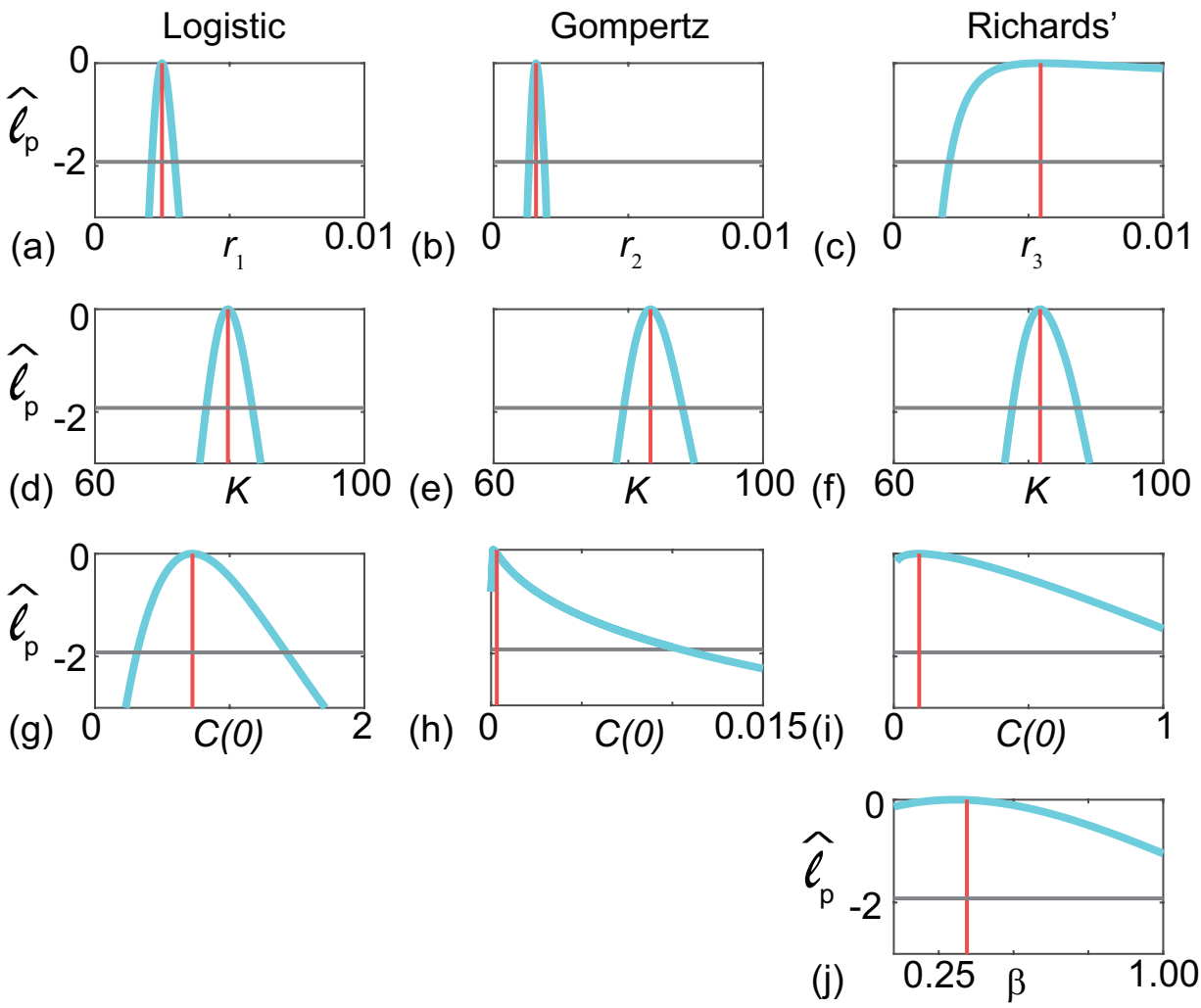

Figure 5: Univariate profile likelihoods for parameters in the logistic, Gompertz and Richards' model, as indicated. In each case the profile is shown (solid blue) together with the MLE (vertical red) and a horizontal line at -1.92.

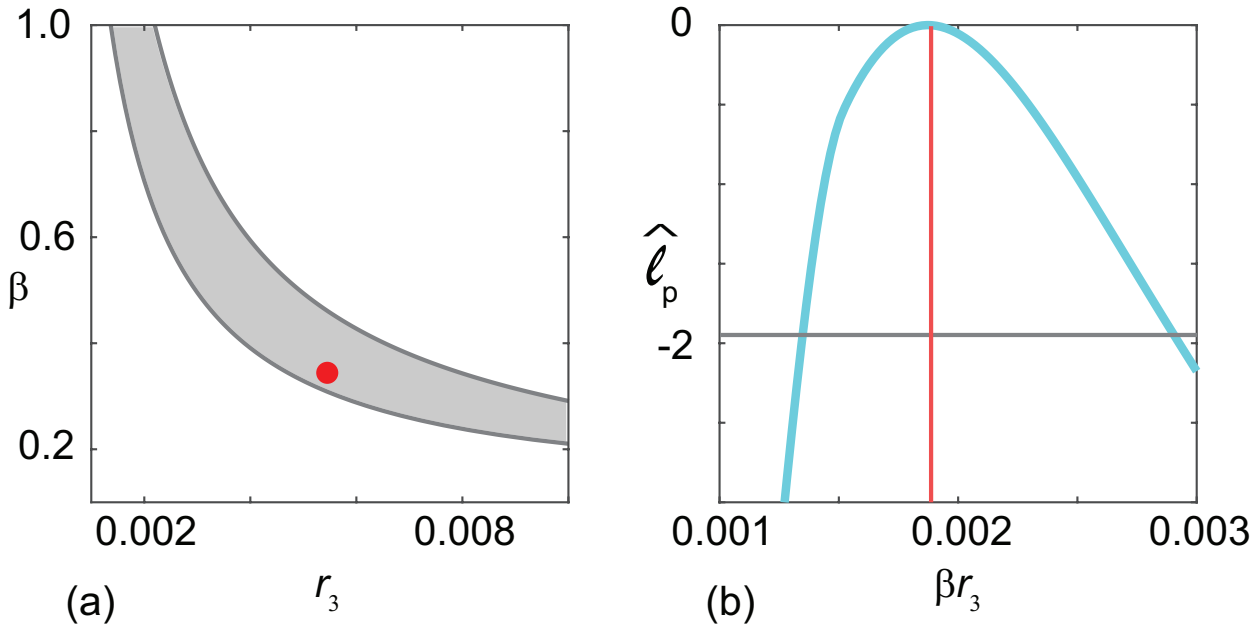

Figure 6: (a) Bivariate profiles for β and λ_3. (b) Univariate profile for the product βr_3. The MLE is $\widehat{\beta r_3} = 0.001874$ [0.0014, 0.0029].

4. Discussion

The crowding functions and solutions for each model depicted in Figure 1 make it clear that the question of model selection for sigmoid growth phenomena is paramount, since all three models lead to sigmoid solutions that capture the key features of the data in Figure 3(b). While the logistic, Gompertz and Richards' growth models are qualitatively similar, they are also quantitatively related as shown in Figure 2 where various solutions with different choices of 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' growth 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 [58], but it is not until we compare solutions directly that we gain a sense of how small β has to be before we establish the accuracy with which we can approximate the Richards' growth model with the simpler Gompertz growth model.

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 to conclude which model is preferable, but it is worthwhile noting some of the quantitative consequences of working with these three different

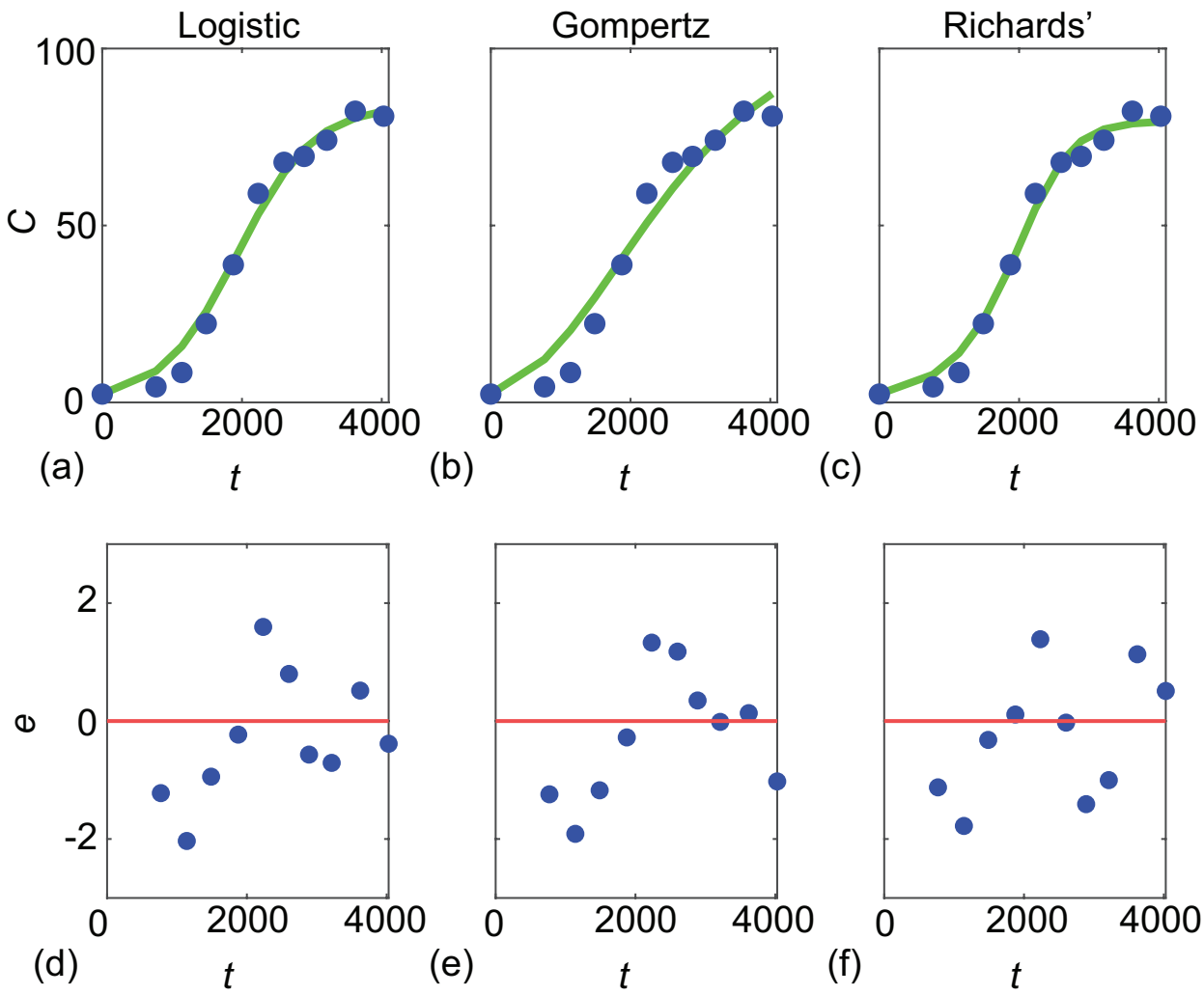

Figure 7: (a)-(c) Observed data (blue discs) superimposed with MLE solutions of the logistic, Gompertz and Richards' models with $C(0) = y_1^o$. (d)-(f) scaled residuals for each model. The Durbin-Watson test yields $DW = 1.0544$ ($p = 0.04841$), $DW = 0.6397$ ($p = 0.0045$), and $DW = 1.2919$ ($p =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

models. Our estimates of $\widehat{r_m}$ provide important ecological insight since these estimates of growth rates provide a direct measure of the time scale of re-growth, but here our estimates vary between the models. Simply taking 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 days for the logistic growth model, 625 days for the Gompertz growth model, and 183 days for the Richards' growth model. 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. This simple exercise demonstrates the importance of addressing model selection, which we shall now explore through the lens of parameter identifiability.

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, results for the Gompertz growth model show that r_2 and K are identifiable, but $C(0)$ is not. The upper bound of the 95% confidence interval for $C(0)$ is two orders of magnitude greater than the MLE, indicating that $C(0)$ is not well-identified by this data, and we will return to this point later. Profiles for the Richards' growth model indicate that while K is identifiable, the remaining parameters are not. In particular, we see that r_3 is one-sided identifiable, and profiles for $C(0)$ and β are relatively flat. Motivated by the presence of the product βr_3 in the exact solution, Equation (4), we show the bivariate profile in Figure 6(b) revealing that the 95% confidence region is hyperbolic. The univariate profile of the product βr_3 in Figure 6(b) suggests that our data is sufficient to obtain precise estimates of the product βr_3, but does not identify the two factors individually. Overall, 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 basic extensions of that model, namely the Gompertz and Richards' growth models, suffer from practical identifiability issues in this case. 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 tools related to 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.

Our observation that $C(0)$ for the Gompertz model (Figure 5(h)) is not well-identified by the data prompts us to briefly 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, $C(0) = y_1^o$ [59, 60]. The quality of the match with the data 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 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. Most strikingly 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. With these results, 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 in the case where we set $C(0) = y_1^o$ and neglect variability in $C(0)$. For completeness, univariate profiles for each model in the case where we set $C(0) = y_1^o$ are provided in the Supplementary Material (Section S1). These additional profiles, however, should not be interpreted too seriously since they are based on a flawed estimation procedure with correlated residuals, as shown in Figure 7. Further results in the Supplementary Material (Section S2) provide additional insight into the relationships between the three models and the potential issue of model family misspecification through additional analysis of synthetic data that is inspired by the findings in the main document.

In this work we present a framework for 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 model, and it is very often unclear which model is most appropriate to work with [22]. Various approaches are adopted to deal with model selection, from the very simplest and widespread approach of neglecting the question completely, to using more developed ideas such as information criteria [1, 5, 9, 28, 49, 63] and Bayesian approaches using Bayes factors [7, 29, 57]. Many of these established approaches, however, are difficult to interpret inferentially [15] 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 set, we show that a simple sigmoid-growth data set can be predicted by many different mathematical models without any obvious challenges. Only when we ask the question whether the parameters are identifiable does it becomes clear that some models are more attractive. In this case we show that a typical data set relating to sigmoid re-growth of hard corals can be modelled using logistic, Gompertz and Richards' growth models. However, both the Gompertz and Richards' growth models encounter identifiability issues, whereas there is no such issue with the logistic growth model. Attempts to overcome identifiability issues by the common but ad-hoc practice fixing of initial conditions leads to 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. While our work here focuses on three simple models and one sigmoid data set, our broad approach 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 [62]. 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] Akike 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] Austin SH, Robinson TR, Robinson WD, Ricklefs RE. 2011. Potential biases in estimating the rate parameter of sigmoid growth functions. Methods in Ecology and Evolution. 2. 43-51. (doi:10.1111/j.2041-210X.2010.00055.x)

[4] Aviv-Sharon E, Aharoni A. 2020. Generalised Logistic growth modeling of the COVID-19 pandemic in Asia. Infectious Disease Modelling. 5. 502-509. (doi: 10.1016/j.idm.2020.07.003).

[5] 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).

[6] 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).

[7] 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).

[8] 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).

[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] Certain G, Barraquand F, Gårdmark A. 2018. How do MAR(1) models cope with hidden nonlinearities in ecological dynamics? Methods in Ecology and Evolution. 9. 1975-1995. (doi: 10.1111/2041-210X.13021).

[11] 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).

[12] 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).

[13] Cortés E. 2016. Perspectives on the intrinsic rate of population growth. Methods in Ecology and Evolution. 7. 1136-1145. (doi: 10.1111/2041-210X.12592).

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

[15] 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).

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

[17] eAtlas: Largest GBR coral reef survey data repository. Retrieved June 2021 eAtlas.

[18] Edelstein-Keshet L. 2005. Mathematical models in biology. SIAM, Philadelphia. (doi: 10.1137/1.9780898719147).

[19] 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).

[20] 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).

[21] Grueber, CE, Nakagawa, S, Laws, RJ, Jamieson, IG. 2011. Multimodel inference in ecology and evolution: challenges and solutions. Journal of Evolutionary Biology. 24(4), 699–711. (doi: 10.1111/j.1420-9101.2010.02210.x).

[22] 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).

[23] 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:).

[24] 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).

[25] 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).

[26] 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).

[27] Jin W, McCue SW, Simpson MJ. 2018. Extended logistic growth model for heterogeneous populations. Journal of Theoretical Biology. 445, 51-61. (doi: 10.1016/j.jtbi.2018.02.027).

[28] 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).

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

[30] Kot M. 2003. Elements of mathematical ecology. Cambridge University Press, Cambridge. (doi: 10.1017/CBO9780511608520).

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

[32] 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).

[33] 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).

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

[35] fmincon: Find minimum of constrained nonlinear multivariable function. Retrieved June 2021 fmincon.

[36] 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).

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

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

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

[40] Ortiz J-C, Wolff NH, Anthony KRN, Devlin M, Lewis S, Mumby PJ. 2018. Impaired recovery of the Great Barrier Reef under cumulative stress. Science Advances. 4. eaar6127. (doi: 10.1126/sciadv.aar6127).

[41] Osborne K, Thompson AA, Cheal AJ, Emslie MJ, Johns KA, Jonker MJ, Logan M, Miller IR, Sweatman HPA. 2017. Delayed coral recovery in a warming ocean. Global Change Biology. 23. 3869-3881. (doi: 10.1111/gcb.13707).

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

[43] 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).

[44] 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).

[45] Richit LA, Bonatto C, da Silva RV, Grzybowski JMV. 2019. Prognostics of forest recovery with r.recovery GRASS-GIS module: an open-source forest growth simulation model based on the diffusive-logistic equation. Environmental Modelling and Software. 111. 108-120.

(doi:10.1016/j.envsoft.2018.10.002).

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

[47] 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).

[48] 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).

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

[50] 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).

[51] 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).

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

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

[54] 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)

[55] 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).

[56] Tjørve KMC, Tjørve E. 2017. The use of Gompertz models in growth analyses, and new Gompertz-model approach: An addition to the unified-Richards family. PLoS ONE. 12. e0178694. (doi: 10.1371/journal.pone.0178691).

[57] 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).

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

[59] Thompson AA, Dolman AM. 2010. Coral bleaching: one disturbance too many for near-shore reefs of the Great Barrier Reef. 2010. Coral Reefs. 29. 637-648. (doi: 10.1007/s00338-009-0562-0).

[60] 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).

[61] 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).

[62] 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).

[63] 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).

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

[65] 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).