Journal of Theoretical Biology 535 (2022) 110998

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

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

ARTICLE INFO

Article history:
Received 21 October 2021
Revised 22 December 2021
Accepted 24 December 2021
Available online 29 December 2021

Keywords:
Gompertz growth
Identifiability analysis
Logistic growth
Parameter estimation
Richards growth

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 in biological modelling without necessarily considering whether parameters are identifiable. Standard practices that do not consider parameter identifiability can lead to unreliable or imprecise parameter estimates and potentially misleading mechanistic interpretations. For example, using the Gompertz model, the estimate of the time scale of coral re-growth is 625 days when we estimate the initial density from the data, whereas it is 1429 days using a more standard approach where variability in the initial density is ignored. While tools developed here focus on three standard sigmoid growth models only, our theoretical developments are applicable to any sigmoid growth model and any appropriate data set. MATLAB implementations of all software are available on GitHub.

1. Introduction

Classical sigmoid growth models play a critical role in many areas of ecology and population biology (Murray, 2002). 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 (Murray, 2002).

There are many mathematical models of biological growth with a range of resulting sigmoid growth curves (Tsoularis and Wallace, 2002; Browning et al., 2017). Within a continuum modelling framework, this general class of mathematical models can be written as

$$\frac{dC}{dt} = rCf(C), \tag{1}$$

where $C(t) \geqslant 0$ is the population density, $t \geqslant 0$ 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 (Jin et al., 2016b). There are many potential choices of crowding function, with typical choices of $f(C)$ being decreasing functions, $df/dC < 0$, with $f(K) = 0$, where $K > 0$ is the 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

* Corresponding author.
E-mail address: matthew.simpson@qut.edu.au (M.J. Simpson).

https://doi.org/10.1016/j.jtbi.2021.110998
0022-5193/

constant $\beta > 0$, which gives rise to the Richards' growth model. Key features of the logistic, Gompertz and Richards' growth models are summarised in Figs. 1 and 2.

Note that estimates of r, and the interpretation of these estimates, depend upon the choice of $f(C)$, and we will provide a concrete example to illustrate this point in Section 2.2. This is an important observation given that some previous approaches to parameter estimation for sigmoid growth models have used an *ad hoc* procedure by treating early time data separately from late time data (Browning et al., 2017; Warne et al., 2017). In this simplistic approach, early time data is used to estimate r from low-density, near-exponential growth at early times for data where $C(0)/K \ll 1$. Then, given the estimate of r, late time data, where crowding effects are pronounced, is used to estimate parameters in the crowding function, $f(C)$. However, as we show, considering the entire time series and estimating r and parameters in $f(C)$ simultaneously confirms that the estimate of r depends on the choice of $f(C)$, and so any interpretation of the estimate of r also depends upon the choice of $f(C)$. We will explore this point more in Section 2.2.

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 (Maini et al., 2004; Sherratt and Murray, 1990; Jin et al., 2016a; Warne et al., 2017; Browning et al., 2021), as well as various ecological population dynamics across a massive range of scales ranging from jellyfish polyps on oyster shells (Melica et al., 2014) to continental-scale dynamics of human populations (Steele et al., 1998). The Gompertz growth model is routinely used to study *in vivo* tumour dynamics (Benzekry et al., 2014; Laird, 1964; Sarapata and de Pillis, 2014) as well as various ecological applications such as reef shark population dynamics (Hisano et al., 2011) and the dynamics of coral populations (Thompson et al., 2020). While the logistic and Gompertz models are probably the most commonly-used sigmoid growth models, various extensions have been proposed (Tsoularis and Wallace, 2002; West et al., 2001; Tjørve and Tjørve, 2010; Gerlee, 2013; Jin et al., 2018).

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

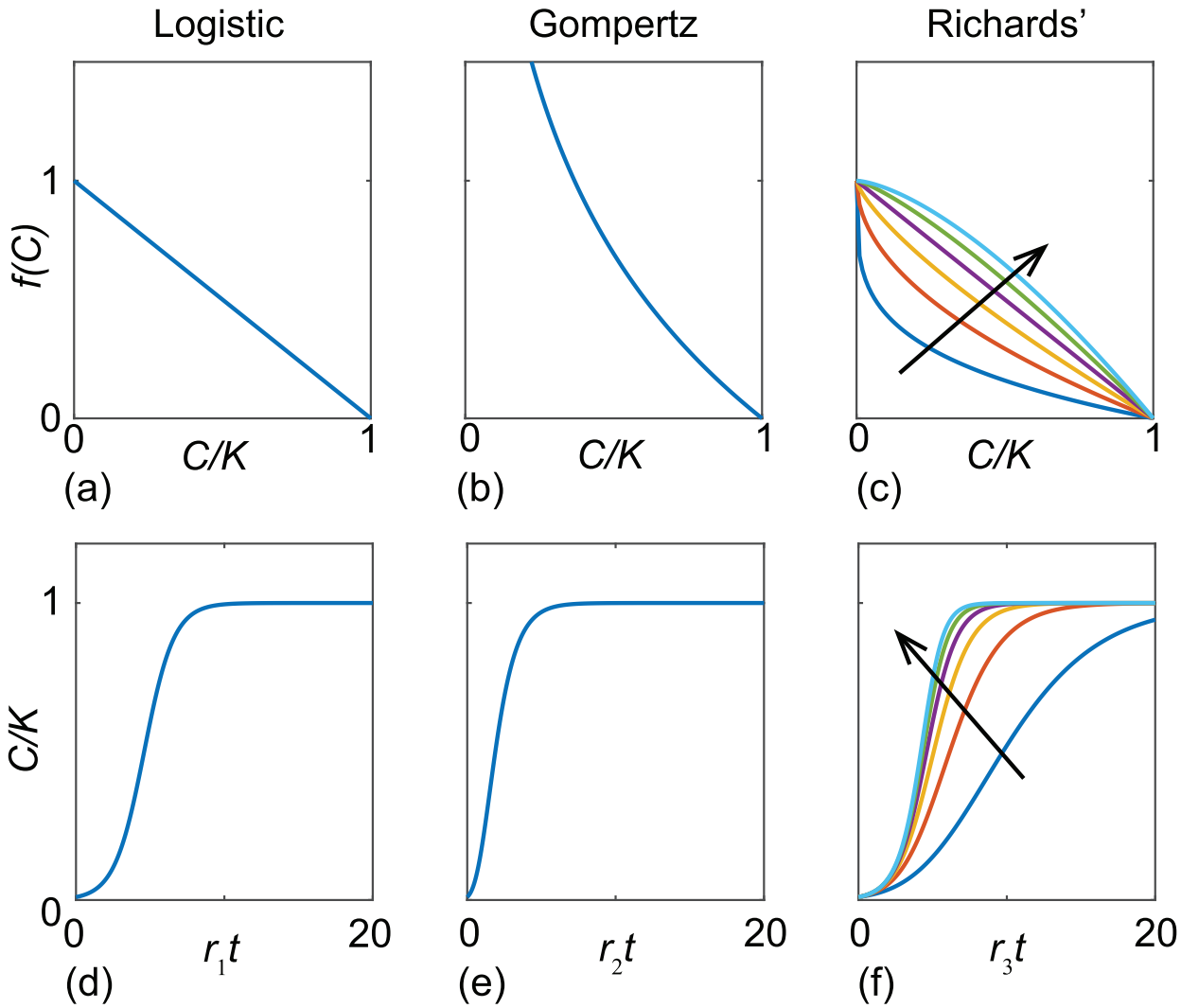

Fig. 1. (a)–(c) $f(C)$ as a function of C/K for the logistic, Gompertz and Richards' models, as indicated. (d)-(f) Solutions of the models in (a)–(c), respectively. Curves in (c) and (f) show various profiles for $\beta = 1/4, 1/2, 3/4, 1, 5/4$ and $3/2$, with the direction of increasing β indicated.

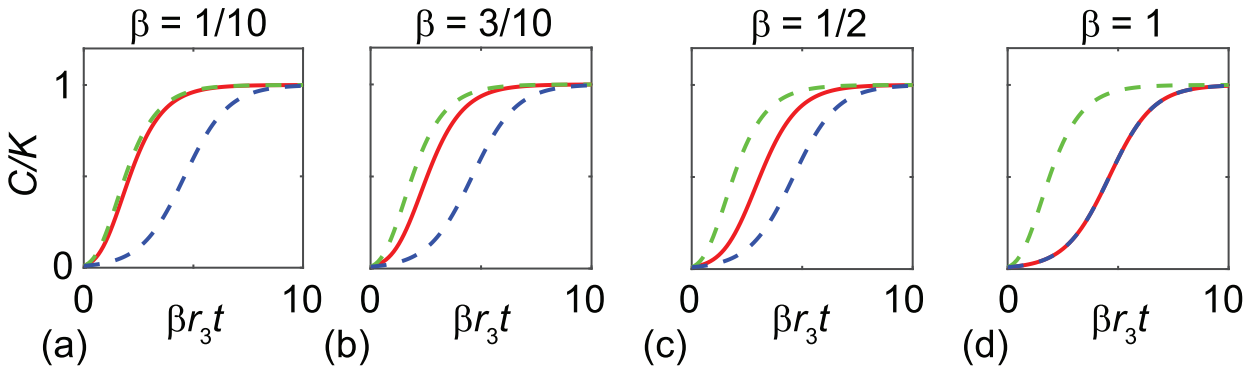

Fig. 2. Inter-model comparison. (a)–(d) various solutions of the logistic (blue dashed), Gompertz (green dashed) and Richards' (solid red) models with $\beta = 1/10, 3/10, 1/2$ and 1, as indicated, with $C(0)/K = 1/100$. The Gompertz solution is shown with $r_2 = \beta r_3$ and the logistic solution is shown with $r_1 = r_3$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

addressed in the nonlinear regression literature for many years (Bates and Watts, 1988; Huet et al., 2004; Ritz and Streibig, 2008; Ross1990,; Seber and Wild, 2003), 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 (Thompson et al., 2020; Warne et al., 2021), while the logistic growth model is widespread within the cell biology modelling community without necessarily considering other options (Maini et al., 2004; Sherratt and Murray, 1990). While such discipline-specific preferences do not imply that mathematical models and their parameterisation within those disciplines are incorrect or invalid, working solely within the confines of discipline-specific traditions 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 identifiability* (Audoly et al., 2001; Chiş et al., 2011a; Maclaren and Nicholson, 2020; Raue et al., 2009; Raue et al., 2013). 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 (Maclaren and Nicholson, 2020; Raue et al., 2009). Practical identifiability involves the ability to estimate parameters to sufficient accuracy given finite, noisy data (Chiş et al., 2011a; Maclaren and Nicholson, 2020; Raue et al., 2009). Methods of identifiability analysis are often used in the systems biology literature where there are many competing models available to describe similar phenomena (Eisenberg and Hayashi, 2014; Meshkat et al., 2018), and these methods provide insight into the trade-off between model complexity and data availability (Browning et al., 2020b). We also consider the often-neglected question of whether basic statistical assumptions required for the validity of identifiability analysis hold. Such checking of statistical validity includes, for instance, testing the assumption of an independent, additive error model by checking for correlation in the residuals after model fitting (Seber and Wild, 2003). In particular, we follow the ideas of likelihood-based frequentist inference outlined in (Pawitan, 2001; Cox, 2006; Sprott, 2008) 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 the Akaike Information Criteria (AIC) (Akaike, 1974) are increasingly used as alternatives to null hypothesis testing to select between models or carry out model averaging in ecology (Burnham et al., 2011; Dormann et al., 2018; Johnson and Omland, 2004). However, the underlying justifications for this trend have been criticised (Guthery et al., 2005; Murtaugh, 2014; Stephens et al., 2005). 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 the 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 implement to interpret. Furthermore, while measures such as the AIC have a predictive focus and interpretation (Akaike, 1974; Dormann et al., 2018), 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' (Shmueli, 2010). We emphasise the differences between these two approaches, by example, in Section 2.3.

The data we consider describes re-growth of hard coral cover on the Great Barrier Reef, Australia, after some external disturbance, such as the impact of a cyclone (eAtlas, 2021). 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 (Hughes et al., 2019). 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 Fig. 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 experimental replicates (e.g. in *in vitro* cell biology (Jin et al., 2016a; Warne

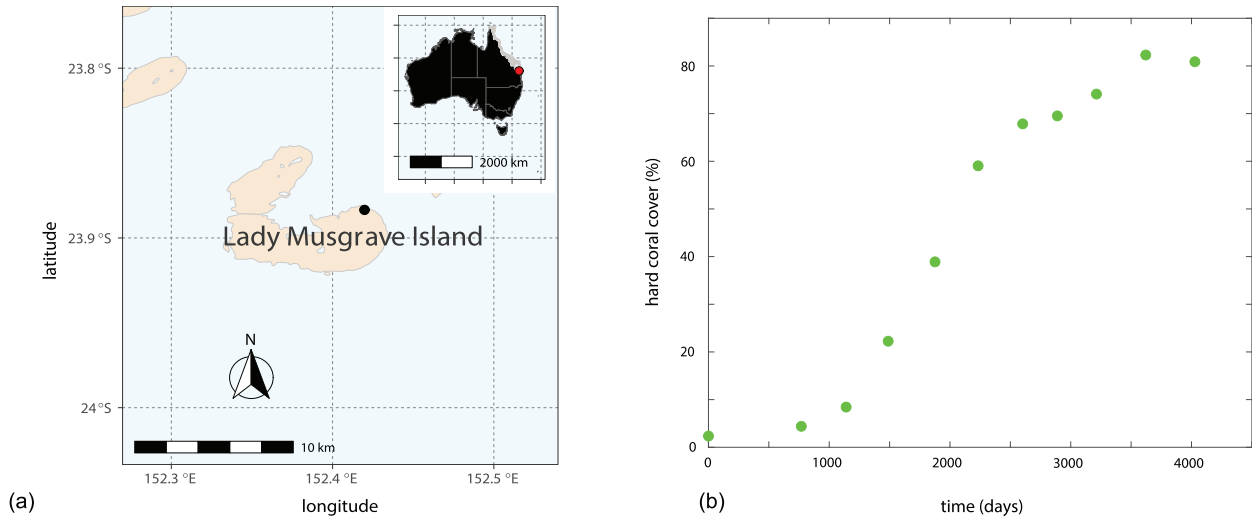

Fig. 3. (a) Location of Lady Musgrave Island (black disc) off the East coast of Australia (inset, red disc). (b) Field data showing the % area covered by hard corals (green discs) as a function of time after some external disturbance. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

et al., 2018)), it is not possible to construct and average this kind of reef-scale re-growth data across many identically prepared experiments simply because each reef will only experience one cyclonic event at a time, and each reef is composed of different coral components. Therefore, here we take the most fundamental approach and work with a single trace data set. This means that, unlike cell biology where identically-prepared experiments are routinely performed and averaged to provide a pre-estimate of the variability in the data (Simpson et al., 2020), it is not possible to pre-estimate the variability in the single trace data for coral re-growth. We overcome this difficulty by simultaneously estimating parameters in the various mathematical models as well as the parameter in a noise model which gives us a quantitative estimate of the variability in the data as part of the model calibration procedure. Within the coral reef modelling community, it is typical to model the kind of sigmoid curve illustrated in Fig. 3 with a Gompertz growth model (Thompson et al., 2020; Warne et al., 2021), 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 methodology and computational tools 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 re-growth 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 (Warne et al., 2017). We will provide further discussion on this point in Section 3.

2. Results and discussion

2.1. Mathematical models

We consider three mathematical models of population growth:

– *Logistic model*

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

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

– *Gompertz model*

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

The Gompertz model has three free parameters, $\theta = (r_2, K, C(0))$, and a logarithmic crowding function, $f(C) = \log(K/C)$, so that $f(K) = 0$ and $f(C) \to \infty$ as $C \to 0^+$.

– *Richards' model*

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

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 $\beta > 0$, so that $f(0) = f(K) = 0$.

For each model the rate parameters have dimensions of 1/day, whereas K and $C(0)$ are 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 Fig. 1 to make it clear that the question of model selection is subtle as all three models lead to qualitatively similar solutions. Indeed, the solutions of each model capture the same trends in the data, shown in Fig. 3, which means that these models are adequate to describe this data set. In addition to being qualitatively similar, the three models are related quantitatively, as shown in Fig. 2 where various solutions with different β are superimposed. Setting $\beta = 1$ in Fig. 2(d) confirms that the Richards' growth model simplifies to the logistic growth model, as expected. Further, comparing solutions in Fig. 2(a)–(c) confirms that the solution of the

Richards' model approaches the solution of the Gompertz 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 (Tsoularis and Wallace, 2002), 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 the Richards' growth model with the simpler Gompertz growth model.

As we make clear in Figs. 1 and 2 the logistic, Gompertz and Richards' growth models are qualitatively similar, but quantitatively distinct. In this work we take the view that the models are phenomenological and the differences between the models are not necessarily indicative of differences in underlying ecological mechanisms. Different opinions about this have been expressed in the literature, with some studies taking the same point of view that these models are simply phenomenological, whereas other studies attempt to relate the differences between the models to mechanistic differences. For example, Sarapata and de Pillis (2014) show that the logistic growth model outperforms a range of other sigmoid growth models when describing the growth of a particular set of breast cancer data, whereas they find that the Gompertz model outperforms the logistic growth model when describing bladder cancer data. They do not, however, claim that these difference are related to mechanistic biological differences between the progression of breast cancer and bladder cancer. In contrast, the seminal work of Laird (1964) provided a plausible biological explanation to explain why *in vivo* tumour growth appear to be best described by the Gompertz model. Laird (1964) argues that the exponential decrease in the net rate of growth in Eq. (3) is related to the fact that tumour cells grow as a result of individual cells passing through a cell cycle with ever diminishing cell cycle speed. Gerlee (2013) provides a very thorough discussion of the relative merits of various sigmoid growth laws in tumour biology, concluding that the Gompertz growth model has dominated historical developments. Interestingly, Gerlee disagrees with Laird's explicit biological justification of the Gompertz model, suggesting that one reason the Gompertz model has been so widely used in the tumour biology literature is related to the ease with which the parameters can be estimated using a relatively straightforward graphical procedure. In this work we do not claim that one model is better than another model at describing the underlying ecological mechanisms leading to coral reef re-growth. Rather, our focus is on quantifying the trade-off between data availability and model complexity as a novel means of model selection.

2.2. Parameter estimation and model checks

Data in Fig. 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, $y_i^o \,|\, \theta \sim \mathcal{N}(y_i, \sigma^2)$, 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 time series of observations, and the above noise assumptions, the log-likelihood function is

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

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

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

subject to the bound constraints: $r_m > 0$ for $m = 1, 2, 3, K > 0, C(0) > 0, \beta > 0$ and $\sigma > 0$. The procedure for estimating $\widehat{\theta}$ involves maximising a multidimensional function, and there are many ways to achieve this numerically. For this work we find that MATLAB's fmincon (fmincon, 2021) is convenient since we can specify bound constraints, and undertake both a local search, based on a suitable estimate of $\widehat{\theta}$, and then double-check the result using a global search. The global search option is more computationally expensive, but helps to avoid potential limitations of relying solely upon a local search, such as finding a local maxima of $\ell(\theta \,|\, y_i^o)$.

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

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

Fig. 4(a)–(c) compares the data with each model evaluated at the MLE. Scaled residuals are given in Fig. 4(e)–(f), and Table 1 summarises $\widehat{\theta}$ for each model. Likelihood-based approaches to parameter estimation implicitly assume that the statistical model is well-specified and that the associated sufficient statistics capture all the information in the data (Cox, 2006). Hence here, for example, this implies that the deterministic model captures the mean trend of the data, while the error term is uncorrelated and approximately normally distributed. While nonlinear regression does not strictly require the normality assumption, correlated residuals can be a valuable indicator of model inadequacy in growth models and other time-series data (Seber and Wild, 2003). Hence, it is desirable to perform a post-estimation check on this assumption – first by simply visualising the residuals and then using a more formal test of residual correlation. In this work, we use the Durbin-Watson test, which tests the null hypothesis of no correlation in the errors against an alternative hypothesis of non-zero coefficient in a first-order autoregressive, AR(1), correlated errors model (Durbin and Watson, 1950; Seber and Wild, 2003). This test shows no indication of residual correlation across all cases in Fig. 4, though we note the sample size is relatively small.

The quality of the fit to the data is excellent for each model in Fig. 4(a)–(c), and by visualising the scaled residuals in Fig. 4(d)–(f) we see that the models are consistent with the standard assumption that the errors, and hence scaled residuals, are uncorrelated. Furthermore, the apparent uncorrelated structure of these residuals is consistent with the Durbin-Watson test (Durbin and Watson, 1950). However, this insightful step of explicitly checking that the statistical assumptions inherent in the estimation of $\widehat{\theta}$ are reasonable is often overlooked. We will return to this point later, where a different example leads to far more correlated scaled residuals.

At this point, confronted with the data-model comparisons in Fig. 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 approximate estimate of the time scale of re-growth to be 400, 630, and 180 days for the logistic, Gompertz and Richards' models, respectively. While these estimates are all of the same order of magnitude, the range of 180–630 days is relatively broad

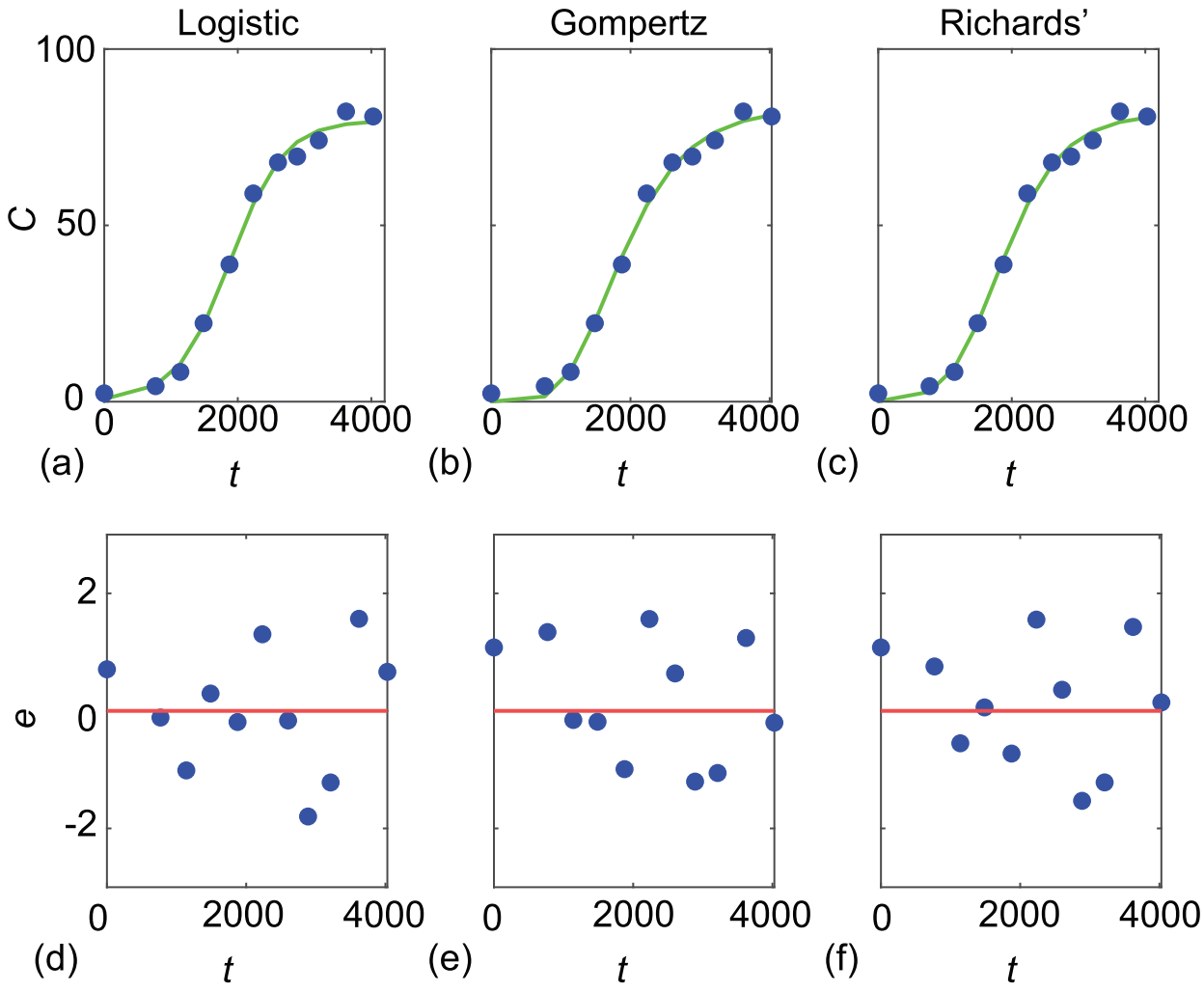

Fig. 4. (a)–(c) Observed data (blue discs) superimposed with MLE solutions of the logistic, Gompertz and Richards' models. (d)–(f) scaled residuals for each model. The Durbin-Watson test yields $DW = 1.8$ ($p = 0.34$), $DW = 2.0$ ($p = 0.49$), and $DW = 2.0$ ($p = 0.51$), for the logistic, Gompertz and Richards' models, respectively.

if we wanted to estimate 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. This example also confirms our earlier assertion that the choice of $f(C)$ impacts our estimate and interpretation of r_m, and this sensitivity can never be explored if only one particular model is used to interrogate the data.

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 an infinite amount of ideal, noise-free data (Eisenberg and Hayashi, 2014; Meshkat et al., 2018). GenSSI software (Chiş et al., 2011b; Ligon et al., 2018) confirms that all three sigmoid growth models are structurally identifiable. In brief, GenSSI uses Lie derivatives of the ODE model to generate a system of equations relating the parameters to the observed quantity, and the solvabil-

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]

ity properties of this system provides information about global and local structural identifiability as well as non-identifiability (Chiş et al., 2011a). The fact that GenSSI shows that the logistic, Gompetz and Richards' models are structurally identifiable is very insightful since it means that any identifiability issues we encounter relate to the question of *practical identifiability* which deals with the more usual scenario of working with finite, noisy data, as in Fig. 3(b) (Raue et al., 2009; Raue et al., 2013; Simpson et al., 2020).

We use a profile likelihood-based approach to explore practical identifiability (Simpson et al., 2021). 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

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

In Eq. (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

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

In all cases we implement this optimisation using the fmincon function in MATLAB with the same bound constraints used to calculate $\widehat{\theta}$ (fmincon, 2021). A typical implementation would involve defining an interval of the interest parameter, say $r_1^- \leqslant r_1 \leqslant r_1^+$, and then discretising this interval uniformly. For example, we could consider a uniform mesh of N points so that we have $r_1(j) = r_1^- + (j-1) \times (r_1^+ - r_1^-)/(N-1)$, for $j = 1, 2, 3, \ldots, N$. For each point on this grid, we again use numerical optimisation methods to estimate values of $K, C(0)$ and σ that maximise $\widehat{\ell}(r_1, K, C(0), \sigma\,|\,y_i)$ for the fixed value of $r_1(j)$. Since numerical optimisation is often based on a local search, it is computationally convenient to take the estimates of $K, C(0)$ and σ for the ith point on the mesh as the initial guess for the adjacent points, $r_1(j \pm 1)$. Throughout this work we use uniformly spaced grids of $N = 100$ points, defined on problem-specific intervals.

Assuming an adequate model family, the likelihood function represents the information about θ contained in y_i^0, and the relative likelihood for different values of θ indicates the relative evidence for these parameter values (Sprott, 2008). A flat profile indicates non-identifiability and the degree of curvature is related to the inferential precision (Pawitan, 2001; Sprott, 2008). 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 (Pawitan, 2001; Royston, 2007), and the points of intersection are determined using linear interpolation. Following the likelihood-based frequentist ideas of Cox (2006) and Sprott (2008), 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 (Durbin and Watson, 1950) to assess residual correlation, noting that we have a simple trend model and relatively small sample sizes.

Fig. 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 Fig. 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.

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 (fmincon, 2021) on a uniform mesh of pairs of the interest parameter, and optimise out the nuisance parameters at each mesh point. In this case we consider a two-dimensional uniform mesh with N mesh points in each direction, $r_3(j) = r_3^- + (j-1) \times (r_3^+ - r_3^-)/(N-1)$ for $j = 1, 2, 3, \ldots, N$, and $\beta(k) = \beta^- + (k-1) \times (\beta^+ - \beta^-)/(N-1)$ for $k = 1, 2, 3, \ldots, N$. This gives us a two-dimensional uniform mesh of points $(r_3(j), \beta(k))$. For each point on the mesh we use numerical optimisation to estimate values of $K, C(0)$ and σ that maximise $\hat{\ell}(r_3, K, \beta, C(0), \sigma\,|\,y_i)$ for the fixed values of $(r_3(j), \beta(k))$ at that mesh point. Again, since numerical optimisation is often based on a local search, it is computationally convenient to take the estimates of $K, C(0)$ and σ at point $(r_3(j), \beta(k))$ on the mesh as the initial guess for the maximisation at adjacent points, $(r_3(j \pm 1), \beta(k \pm 1))$. The bivariate profile calculations here use $N = 100$, so that we work with a 100×100 mesh. For the bivariate profile the relative log-likelihood threshold of -3.00 gives approximate 95% confidence regions (Pawitan, 2001; Royston, 2007).

Using these methods we generate the bivariate profile in Fig. 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 within this broad, hyperbolic shaped region. To investigate further we return to compute a univariate profile by re-scaling the Richards' model so that we treat the product βr_3 as an interest parameter. The univariate profiles for this product (Fig. 6(b)) suggests that our data is sufficient to obtain precise estimates of the product βr_3, but that it does not identify β and r_3 individually. The univariate profile in Fig. 6(b) indicates that choosing the product $\beta r_3 = 1.9 \times 10^{-3}$ gives the best match to the data, but estimating this product is not helpful if we wish to identify β and r_3 separately. To highlight potential issues involving such types of non-identifiability we consider $\beta = 0.1$, then setting $r_3 = 1.9 \times 10^{-2}$ /day gives an excellent match to the data, and in this case the re-growth time scale is $1/r_3 = 53$ days. In comparison, if $\beta = 1$, then setting $r_3 = 1.9 \times 10^{-3}$ /day also gives an excellent match to the data, and in this case the re-growth time scale is $1/r_3 = 530$ days. This exercise provides concrete evidence that our data is insufficient to provide precise parameter estimates for the Richards' growth model since there are many parameter com-

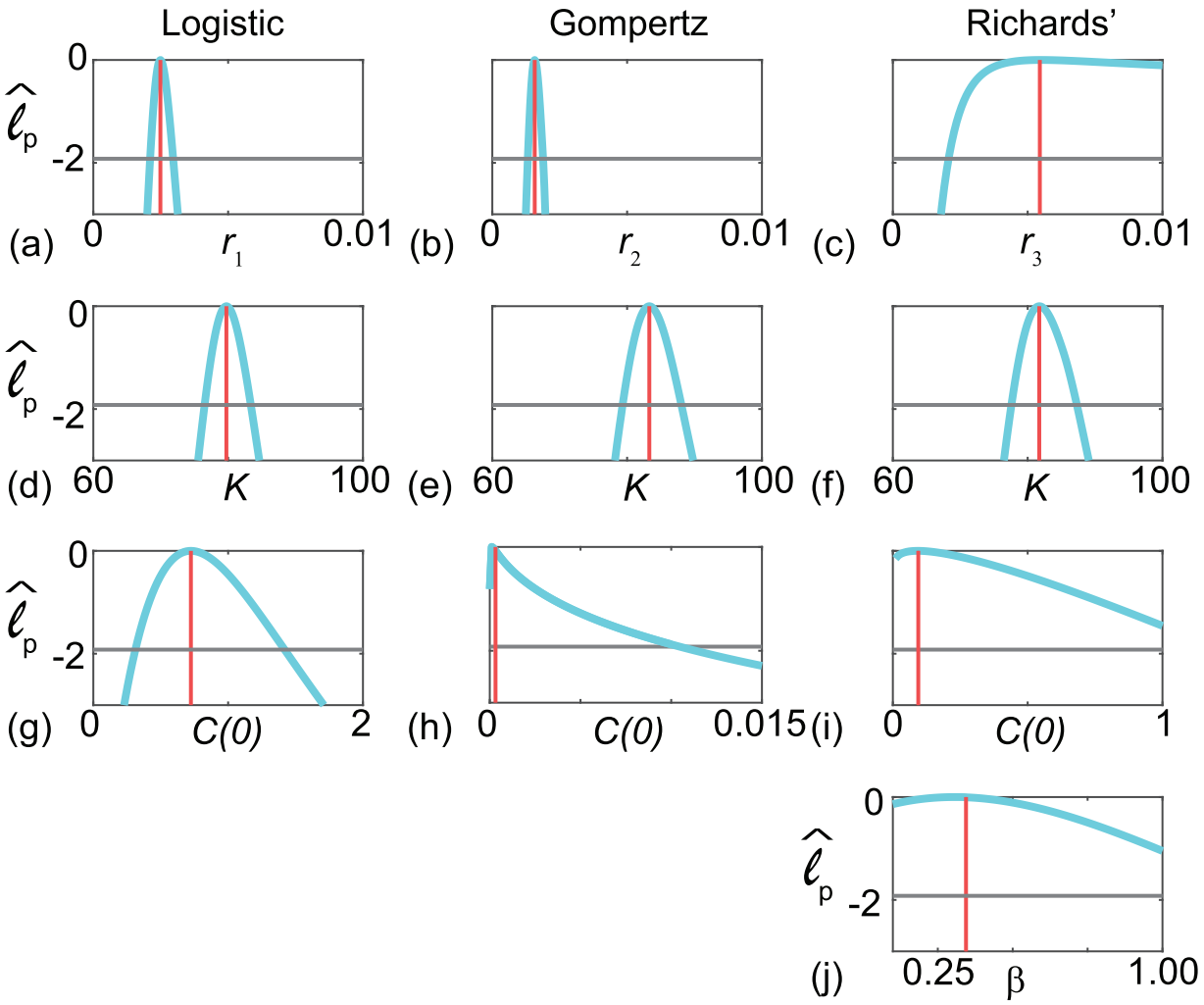

Fig. 5. Univariate profile likelihoods for parameters in the logistic, Gompertz and Richards' model, as indicated. In each case the profile is shown (solid blue) together with the MLE (vertical red) and a horizontal line at −1.92. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

binations that match the data and these parameter estimates should not be used to support decision making.

In summary, comparing the profiles between different models in Fig. 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 using parameter identifiability to guide model selection is not routinely considered, it provides a straightforward, computationally-efficient means of informing quantitative distinctions between

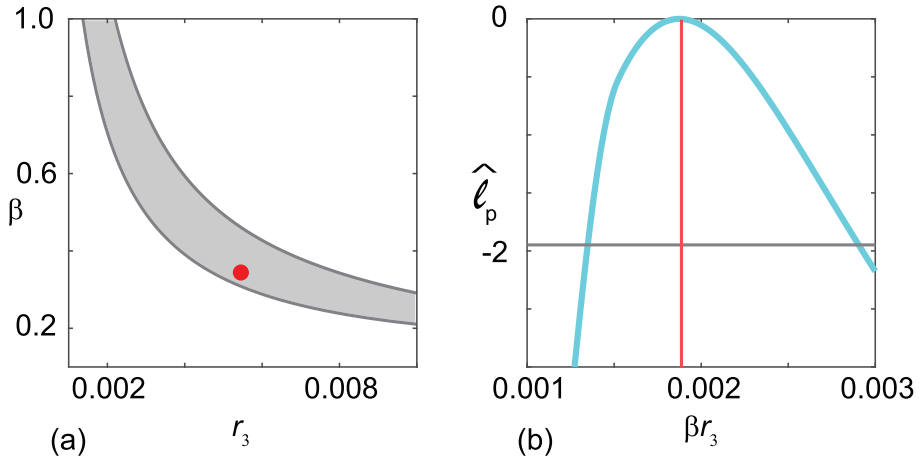

Fig. 6. (a) Bivariate profiles for β and r_3. (b) Univariate profile for the product βr_3. The MLE is $\widehat{\beta r_3} = 1.9 \times 10^{-3}$ $\left[1.4 \times 10^{-3}, 2.9 \times 10^{-3}\right]$.

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.

It is instructive to contrast our profile likelihood-based analysis of model selection with a more standard Information criteria-based analysis. A standard way to compare models is using the AIC (Akaike, 1974),

$$\text{AIC} = -2 \sup_{\theta} \ \ell(\theta|y_i) + 2k, \tag{12}$$

where k is the dimensionality of θ (Warne et al., 2018). Given a set of candidate models, the preferred model is the one with the minimum AIC. This means that the AIC rewards goodness of fit, measured by the likelihood function, but penalises model complexity through the $+2k$ term. This approach to model selection is very popular, and perhaps this popularity stems from the fact that this criterion is straightforward to implement and interpret. However, as we will now demonstrate, taking an AIC approach provides far less information than our profile likelihood-based analysis. The AIC approach ranks the models: (i) $\text{AIC} = -41.5$ for the logistic model; (ii) $\text{AIC} = -40.3$ for the Gompertz model; and (iii) $\text{AIC} = -37.4$ for the Richards' model. This ranking is consistent with our profile likelihood-based analysis, but it is obvious that the profile likelihood-based analysis provides a more nuanced and useful insight. For example, the profile likelihood-based analysis confirms that the logistic model is practically identifiable, whereas both the Gompertz and Richards' models encounter practical identifiability issues with this data. In particular, not only does the profile likelihood-based approach highlight issues of practical identifiability, it shows precisely which parameters are identifiable, which parameters are not-identifiable, and the nature of this non-identifiability. For instance, we find that $C(0)$ is one-sided identifiable for the Gompertz model and the product βr_3 is identifiable for the Richards' model whereas, individually, β and r_3 are not identifiable. In contrast, the AIC-based analysis provides no such detailed insight. Therefore, given the relative ease with which profile likelihoods can be calculated, we suggest that this approach ought to be more widely adopted because it provides far more useful insight into model selection.

All results in this section correspond to the data in Fig. 3(b) which describes the re-growth of coral at site 1 near Lady Musgrave Island. For completeness, we repeat this analysis for a second, independent data set describing the re-growth of coral at site 3 near Lady Musgrave Island, with full results reported in the Supplementary Material. In summary, we find the same outcome applies to this second data set, namely that the logistic growth model does not suffer from any practical identifiability issues, whereas both the Gompertz and Richards' growth model suffer from practical identifiability issues in this second set of data. Note that our MATLAB algorithms on GitHub contain both data sets and can be used to replicate all calculations in both the main and supplementary material documents.

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

Our observation that the initial density $C(0)$, for the Gompertz model (Fig. 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, $C(0) = y_1^o$ (Thompson et al., 2020; Warne et al., 2021). Fig. 7(a)–(c) compares the data with the model evaluated at the MLE with $C(0) = y_1^o$, and the scaled residuals are shown in Fig. 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.

The quality of the model-data match in Fig. 7(a)–(c) is clearly poorer than when we simultaneously estimate $C(0)$ as part of $\hat{\theta}$ (Fig. 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 Fig. 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 $\hat{r}_2$ 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 Fig. 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 Fig. 8(a)–(d) are generated by setting $r_3 = 5.5 \times 10^{-3}$, $K = 80, C(0) = 1$ 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 are plotted in Fig. 8 and Fig. 8(i)–(l), respectively, and 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 Fig. 8(m)–(p). These profiles again highlight certain relationships between the Richards', logistic and Gompertz growth models established in Fig. 4. In particular, for sufficiently small β in Fig. 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 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 Fig. 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 is relatively narrow, implying that it is possible to obtain meaningful and interpretable parameter estimates when working with a misspecified model. The remaining profiles in Fig. 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 uncorre-

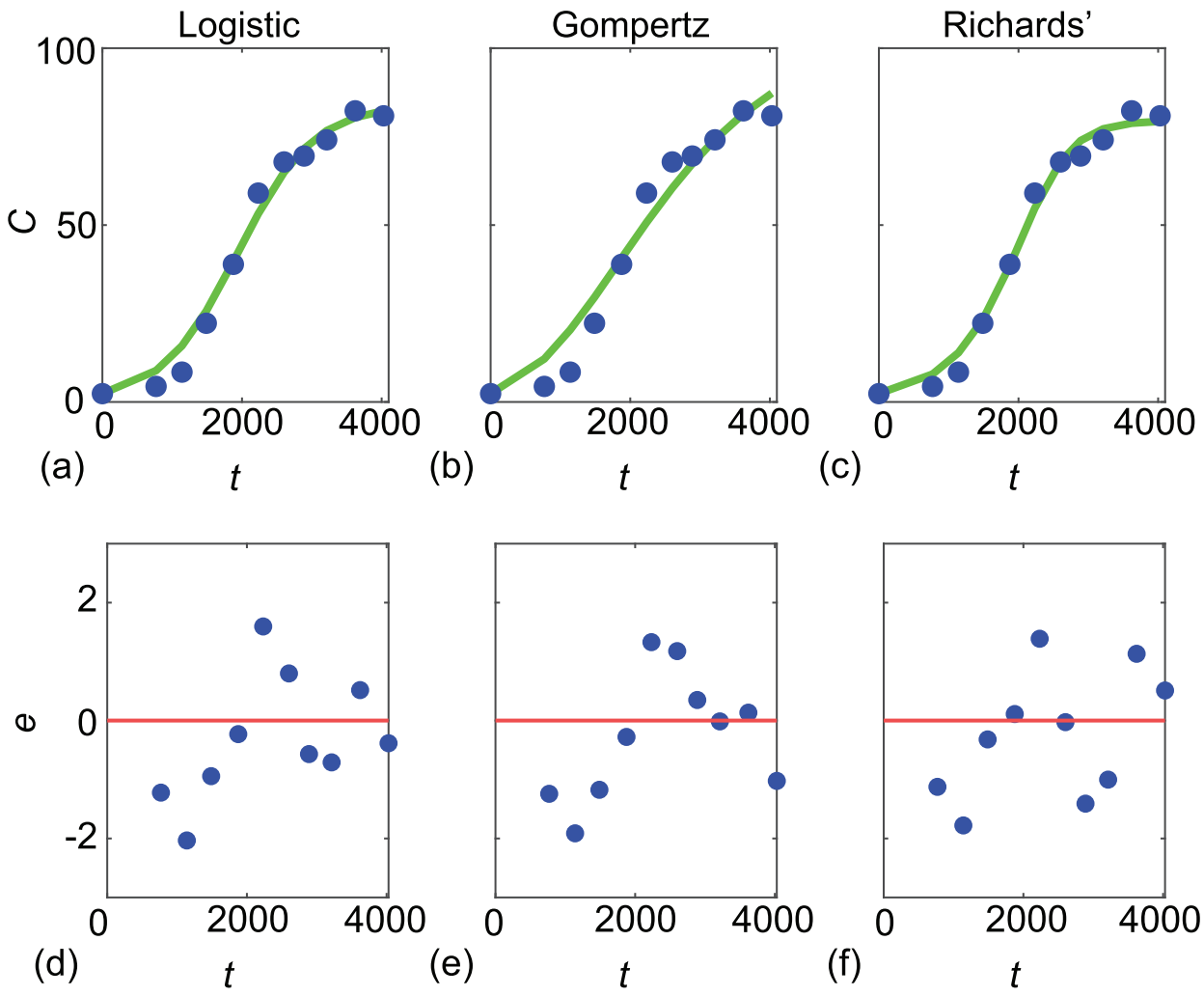

Fig. 7. (a)-(c) Observed data (blue discs) superimposed with MLE solutions of the logistic, Gompertz and Richards' models with $C(0) = y_1^o$. (d)-(f) scaled residuals for each model. The Durbin-Watson test yields $DW = 1.1$ ($p = 0.048$), $DW = 0.64$ ($p = 0.0045$), and $DW = 1.3$ ($p = 0.11$), for the logistic, Gompertz and Richards' models, respectively.

lated residuals and an excellent match to the data in Fig. 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 $(\beta\widehat{r_3} - \widehat{r_i})/r_3$ for $i = 1, 2$, that are plotted in Fig. 9(a) for the logistic and Gompertz models, respectively. As before in Fig. 8, the quantity $(\beta\widehat{r_3} - \widehat{r_2})/r_3$ approaches zero as β becomes sufficiently small, while the quantity $(\beta\widehat{r_3} - \widehat{r_1})/r_3$ is approximately zero when $\beta = 1$, as expected. Plotting these quantities with noise-free data in Fig. 9(a) illustrates how the MLEs for $\widehat{r_1}$ and $\widehat{r_2}$ vary with β without the influence of external noise. Analogous data in Fig. 9(b) shows the influence of corrupting the data with some noise, and we see that the underlying trends in these quantities are modestly impacted by incorporating noise into 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 Fig. 9(c) shows how the width of the profiles varies with β for

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]

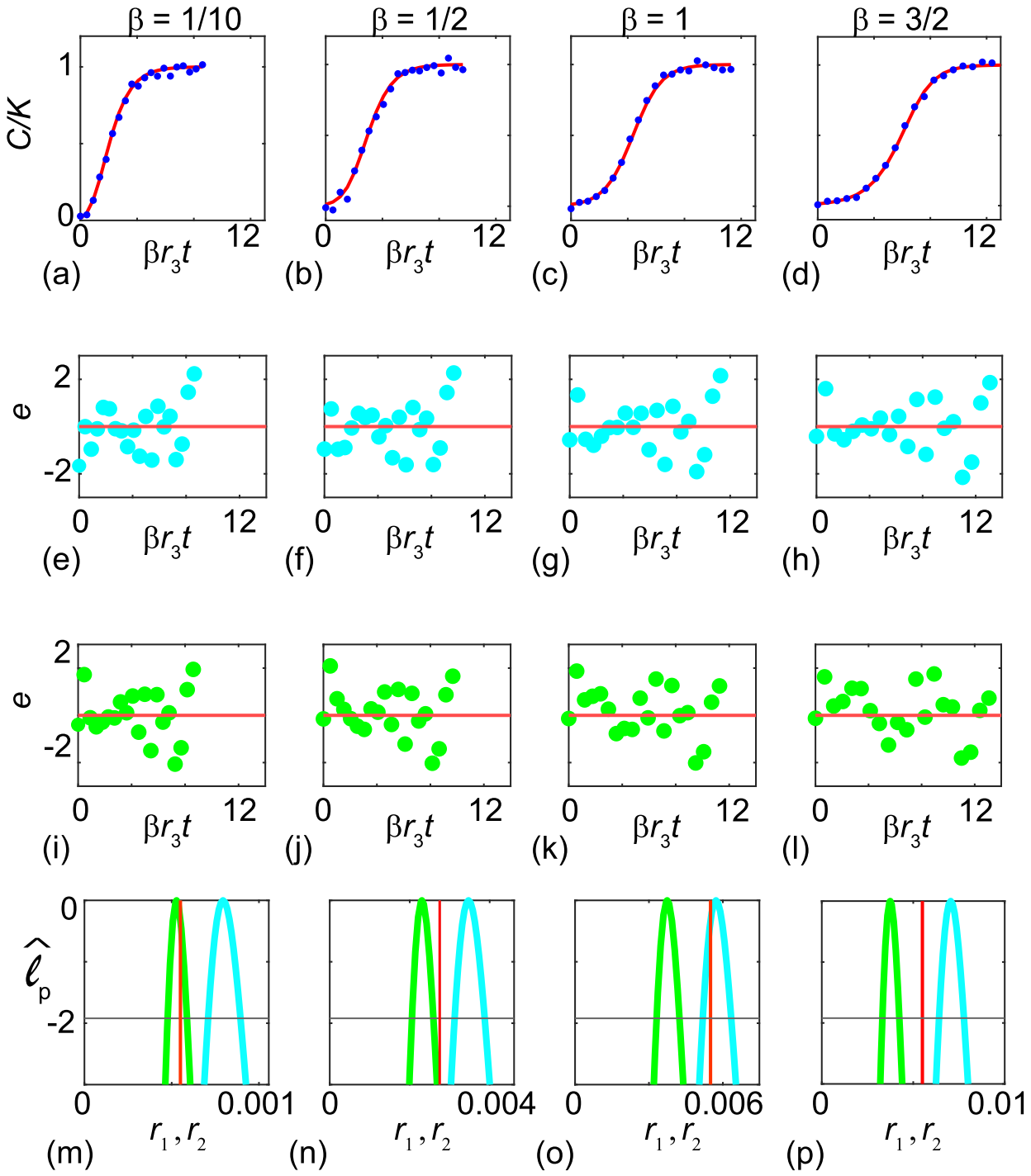

Fig. 8. (a)–(d) Solution of the Richards' growth curves (solid red) for the values of β indicated with $C(0) = 1, r_3 = 5.5 \times 10^{-3}$ and $K = 80$. Noisy data (blue dots) are obtained by sampling 20 equally-spaced observations in the interval $0 < t < T$ and adding normally distributed noise with zero mean and $\sigma = 2$. Here T is defined implicitly by $C(T) = 0.999K$. For each data we estimate $\widehat{\theta}$ using same methods outlined in the methods in Section 2.2. 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 $\beta r_3 = 5.5 \times 10^{-3}\beta$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

noise-free data, $\sigma = 0$. 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 Fig. 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 Fig. 9(c) are completely obscured by the incorporation of noise. Overall, one way to interpret the results in Figs. 8,9 is that the Richards' model for $\beta \leqslant 1$ can be reasonably well approximated by simpler models, whereas for cases where $\beta > 1$ the Richards' model may not be reliably interpreted using simpler models.

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 prac-

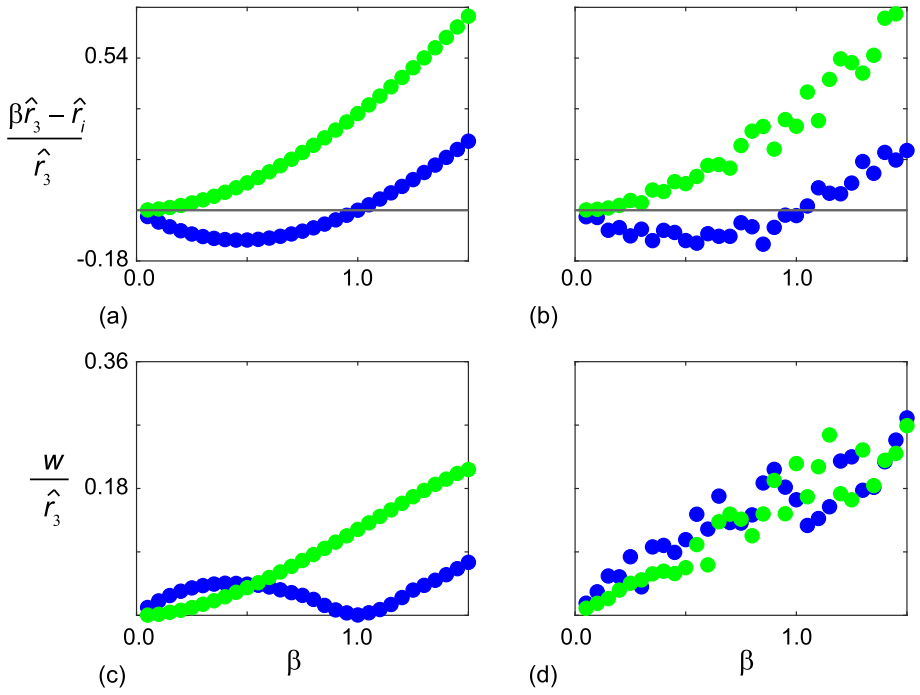

Fig. 9. (a)–(b) show quantities $(\beta\hat{r}_3 - \hat{r}_m)/\hat{r}_3$ for $m = 1, 2$ as functions of β for noise free data ($\sigma = 0$) and noisy data ($\sigma = 2$), respectively. (c)–(d) show the width of the profiles, w, plotted as functions of β for noise free data ($\sigma = 0$) and noisy data ($\sigma = 2$), 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. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

tical 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 (Gutenkunst et al., 2007). 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 (Akaike, 1974; Burnham et al., 2011; Bortz and Nelson, 2006; Johnson and Omland, 2004; Shmueli, 2010; Warne et al., 2018) and Bayesian approaches that use, for example, Bayes factors (Browning et al., 2020a; Kass and Raftery, 1995; Toni et al., 2009). Many of these established approaches, however, are difficult to interpret inferentially (Dormann et al., 2018) 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 the apparently simple sigmoid growth data can be accurately 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 more appealing than others. 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 (Thompson et al., 2020; Warne et al., 2021).

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 assumption 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 (Warne et al., 2017. Extending

our tools to deal with the question of experimental design is an open area that we expect will provide great insight into the question of optimal experimental design.

In conclusion, it is worth revisiting the point we made in the Introduction where we noted that the coral re-growth data we deal with is of high interest because the data describes the whole recovery trajectory, including late time data where the net growth rate is reduced as the density approaches the carrying capacity. In other applications, such as cell biology, usually only early time data is collected because maintaining cell culture for very long periods of time is difficult and expensive (Warne et al., 2017). In such cases, measurements only cover the early-time concave-up part of the sigmoid curve and do not include late-time, concave down data. This has important implications for our ability to estimate carrying capacity densities. Some of our previous work explored this challenge in detail, see Warne et al. (2017), for example (Warne et al., 2017). In such cases it is straightforward to use a profile likelihood-based method for model selection, but it would be more natural to focus on a different class of models that do not incorporate a carrying capacity density. One such mathematical model is the power law model, $\mathrm{d}C/\mathrm{d}t = rC^a$, where the parameters r and a are to be estimated from data (Sarapata and de Pillis, 2014). The practical identifiability of this model can be assessed using the same tools presented in this work for data that does not include long-time crowding effects.

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.

CRediT authorship contribution statement

Matthew J. Simpson: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Funding acquisition. **Alexander P. Browning:** Conceptualization, Methodology, Writing – review & editing. **David J. Warne:** Conceptualization, Methodology, Writing – review & editing. **Oliver J. Maclaren:** Conceptualization, Methodology, Writing – review & editing. **Ruth E. Baker:** Conceptualization, Methodology, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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. We thank two referees for their helpful suggestions.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19, 716–723. https://doi.org/10.1109/TAC.1974.1100705.

Audoly, S., Bellu, G., D'Angiò, L., Saccomani, M.P., Cobelli, C., 2001. Global identifiability of nonlinear models of biological systems. IEEE Trans. Biomed. Eng. 48, 55–65. https://doi.org/10.1109/10.900248.

Bates, D.M., Watts, D.G., 1988. Nonlinear regression analysis and its applications. John Wiley & Sons.

Benzekry, S., Lamont, C., Beheshti, A., Tracz, A., Ebos, J.M.L., Hlatky, L., Hahnfeldt, P., 2014. Classical mathematical models for description and prediction of experimental tumor growth. PLoS Comput. Biol. 10,. https://doi.org/10.1371/journal.pcbi.1003800 e1003800.

Bortz, D.M., Nelson, P.W., 2006. Model selection and mixed-effects modeling of HIV infection dynamics. Bull. Math. Biol. 68, 2005–2025. https://doi.org/10.1007/s11538-006-9084-x.

Browning, A.P., McCue, S.W., Simpson, M.J., 2017. A Bayesian computational approach to explore the optimal duration of a cell proliferation assay. Bull. Math. Biol. 79, 1888–1906. https://doi.org/10.1007/s11538-017-0311-4.

Browning, A.P., Jin, W., Plank, M.J., Simpson, M.J., 2020a. Identifying density-dependent interactions in cell populations. J. R. Soc. Interface 17, 20200143. https://doi.org/10.1098/rsif.2020.0143.

Browning, A.P., Warne, D.J., Burrage, K., Baker, R.E., Simpson, M.J., 2020b. Identifiability analysis for stochastic differential equations models in systems biology. J. R. Soc. Interface 17, 20200652. https://doi.org/10.1098/rsif.2020.0652.

Browning, A.P., Maclaren, O.J., Burnzli, P.R., Lanaro, M., Allenby, M.C., Woodruff, M. A., Simpson, M.J., 2021. Model-based data analysis of tissue growth in thin 3D printed scaffolds. J. Theor. Biol. 528,. https://doi.org/10.1016/j.jtbi.2021.110852 110852.

Burnham, K.P., Anderson, D.R., Huyvaert, K.P., 2011. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behav. Ecol. Sociobiol. 65 (1), 23–35. https://doi.org/10.1007/s00265-010-1029-6.

Chiş, O., Banga, J.R., Balsa-Canto, E., 2011a. Structural identifiability of systems biology models: a critical comparison of methods. PLoS One 6,. https://doi.org/10.1371/journal.pone.0027755 e27755.

Chiş, O., Banga, J.R., Balsa-Canto, E., 2011b. GenSSI: a software toolbox for structural identifiabilty analysis of biological models. Bioinformatics 18, 2610–2611. https://doi.org/10.1093/bioinformatics/btr431.

Cox, D.R., 2006. Principles of statistical inference. Cambridge University Press.

Dormann, C.F., Calabrese, J.M., Guillera-Arroita, G., Matechou, E., Bahn, V., Barto'n, K., Beale, C.M., 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. Ecol. Monogr. 88, 485–504. https://doi.org/10.1002/ecm.1309.

Durbin, J., Watson, G.S., 1950. Testing for serial correlation in least squares regression: I. Biometrika 37, 409–428.

eAtlas: Largest GBR coral reef survey data repository. Retrieved November 2021 eAtlas..

Eisenberg, M.C., Hayashi, M.A.L., 2014. Determining identifiable parameter combinations using subset profiling. Math. Biosci. 256, 115–126. https://doi.org/10.1016/j.mbs.2014.08.008.

Gerlee, P., 2013. The model muddle: In search of tumor growth laws. Cancer Res. 73, 2407–2411. https://doi.org/10.1158/0008-5472.CAN-12-4355.

Gutenkunst, R., Waterfall, J., Casey, F., Brown, K., Myers, C., Sethna, J., 2007. Universally sloppy parameter sensitivities in systems biology models. PLoS Comput. Biol. 3,. https://doi.org/10.1371/journal.pcbi.0030189 e189.

Guthery, F.S., Brennan, L.A., Peterson, M.J., Lusk, J.J., 2005. Information theory in wildlife science: critique and viewpoint. J. Wildlife Manage. 69, 457–465.

Hisano, M., Connolly, S.R., Robbins, W.D., 2011. Population growth rates of reef sharks with and without fishing on the Great Barrier Reef: Robust estimation with multiple models. PLoS One 6,. https://doi.org/10.1371/journal.pone.0025028 e25028.

Hughes, T.P., Kerry, J.T., Baird, A.H., Connolly, S.R., Chase, T.J., Dietzel, A., Hill, T., Hoey, A.S., Hoogenboom, M.O., Jacobson, M., Kerswell, A., Madin, J.S., Mieog, A., Paley, A.S., Pratchett, M.S., Torda, G., Woods, R.M., 2019. Global warming impairs stock-recruitment dynamics of corals. Nature 568, 387–390. https://doi.org/10.1038/s41586-019-1081-y.

Huet, S., Bouvier, A., Poursat, M.-A., Jolivet, S., 2004. Statistical tools for nonlinear regression. Springer, New York.

Jin, W., Shah, E.T., Penington, C.J., McCue, S.W., Chopin, L.K., Simpson, M.J., 2016a. Reproducibility of scratch assays is affected by the initial degree of confluence: experiments, modelling and model selection. J. Theor. Biol. 390, 136–145. https://doi.org/10.1016/j.jtbi.2015.10.040.

Jin, W., Penington, C.J., McCue, S.W., Simpson, M.J., 2016b. Stochastic simulation tools and continuum models for describing two-dimensional collective cell spreading with universal growth functions. Phys. Biol. 13,. https://doi.org/10.1088/1478-3975/13/5/056003 056003.

Jin, W., McCue, S.W., Simpson, M.J., 2018. Extended logistic growth model for heterogeneous populations. J. Theor. Biol. 445, 51–61. https://doi.org/10.1016/j.jtbi.2018.02.027.

Johnson, J.B., Omland, K.S., 2004. Model selection in ecology and evolution. Trends Ecol. Evol. 19, 101–108. https://doi.org/10.1016/j.tree.2003.10.013.

Kass, R., Raftery, A., 1995. Bayes factors. J. Am. Stat. Assoc. 90, 773–795. https://doi.org/10.1080/01621459.1995.10476572.

Laird, A.K., 1964. Dynamics of tumour growth. Brit. J. Cancer 18, 490–502. https://doi.org/10.1038/bjc.1964.55.

Ligon, T.S., Frölich, F., Chiş, O., Banga, J.R., Balsa-Canto, E., Hasenauer, J., 2018. GenSSI 2.0: multi-experimental structural identifiability analysis of SBML models. Bioinformatics 34, 1421–1423. https://doi.org/10.1093/bioinformatics/btx735.

Maini, P.K., McElwain, D.L.S., Leavesley, D.I., 2004. Traveling wave model to interpret a wound-healing cell migration assay for human peritoneal mesothelial cells. Tissue Eng. 10, 475–482. https://doi.org/10.1089/107632704323061834.

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

fmincon: Find minimum of constrained nonlinear multivariable function. Retrieved November 2021 MATLAB: fmincon..

Melica, V., Invernizzi, S., Caristi, G., 2014. Logistic density-dependent growth of Aurelia aurita polyps population. Ecol. Model. 10, 1–5. https://doi.org/10.1016/j.ecolmodel.2014.07.009.

Meshkat, N., Sullivant, S., Eisenberg, M., 2018. Identifiability results for several classes of linear compartment models. Bull. Math. Biol. 8, 1620–1651. https://doi.org/10.1007/s11538-015-0098-0.

Murray, J.D., 2002. Mathematical biology I: An introduction. Heidelberg: Springer. (doi: 10.1007/978-3-662-08539-4)..

Murtaugh, P.A., 2014. In defense of P values. Ecology 95, 611–617.

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

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., Timmer, J., 2009. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics 25, 1923–1929. https://doi.org/10.1093/bioinformatics/btp358.

Raue, A., Kreutz, C., Theis, F.J., Timmer, J., 2013. Joining forces of Bayesian and frequentist methodology: a study for inference in the presence of non-identifiability. Philos. Trans. R. Soc. A 371, 20110544. https://doi.org/10.1098/rsta.2011.0544.

Ritz, C., Streibig, J.C., 2008. Nonlinear regression with R. Springer, New York.

Ross, G.J.S., 1990. Nonlinear estimation. Springer, New York.

Royston, P., 2007. Profile likelihood for estimation and confidence intervals. Stata J. 7, 376–387. https://doi.org/10.1177/1536867X0700700305.

Sarapata, E.A., de Pillis, L.G., 2014. A comparison and cataolg of intrinsic tumor growth models. Bull. Math. Biol. 76, 2010–2024. https://doi.org/10.1007/s11538-014-9986-y.

Seber, G.A.F., Wild, C.J., 2003. Nonlinear regression. Wiley-Interscience, New Jersey.

Sherratt, J.A., Murray, J.D., 1990. Models of epidermal wound healing. Proc. R. Soc. B 241, 29–36. https://doi.org/10.1098/rspb.1990.0061.

Shmueli, G., 2010. To explain or to predict? Stat. Sci. 25, 289–310. https://doi.org/10.1214/10-STS330.

Simpson, M.J., Baker, R.E., Vittadello, S.T., Maclaren, O.J., 2020. Parameter identifiability analysis for spatiotemporal models of cell invasion. J. R. Soc. Interface 17, 20200055. https://doi.org/10.1098/rsif.2020.0055.

Simpson, M.J., Browning, A.P., Drovandi, C., Carr, E.J., Maclaren, O.J., Baker, R.E., 2021. Profile likelihood analysis for a stochastic model of diffusion in heterogeneous media. Proc. R. Soc. A 477, 20210214. https://doi.org/10.1098/rspa.2021.0214.

Sprott, D.A., 2008. Statistical inference in science. Springer Science & Business Media.

Steele, J., Adams, J., Slukin, T., 1998. Modelling paleoindian dispersals. World Archaeol. 30, 286–305. https://doi.org/10.1080/00438243.1998.9980411.

Stephens, P.A., Buskirk, S.W., Hayward, G.D., Martinez Del Rio, C., 2005. Information theory and hypothesis testing: a call for pluralism. J. Appl. Ecol. 42, 4–12. https://doi.org/10.1111/j.1365-2664.2005.01002.x.

Tjørve, E., Tjørve, K.M.C., 2010. A unified approach to the Richards-model family for use in growth analyses: why we need only two model forms. J. Theor. Biol. 3, 417–425. https://doi.org/10.1016/j.jtbi.2010.09.008.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., Stumpf, M.P.H., 2009. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J. R. Soc. Interface 6. https://doi.org/10.1098/rsif.2008.0172. 187–120.

Tsoularis, A., Wallace, J., 2002. Analysis of logistic growth models. Math. Biosci. 179, 21–55. https://doi.org/10.1016/S0025-5564(02)00096-2.

Thompson, A., Martin, K., Logan, M., 2020. Development of the coral index, a summary of coral reef resilience as a guide for management. J. Environ. Manage. 271,. https://doi.org/10.1016/j.jenvman.2020.111038 111038.

Warne, D.J., Baker, R.E., Simpson, M.J., 2017. Optimal quantification of contact inhibition in cell populations. Biophys. J. 113, 1920–1924. https://doi.org/10.1016/j.bpj.2017.09.016.

Warne, D.J., Baker, R.E., Simpson, M.J., 2018. Using experimental data and information criteria to guide model selection for reaction-diffusion problems in mathematical biology. Bull. Math. Biol. 81, 1760–1804. https://doi.org/10.1007/s11538-019-00589-x.

Warne, D.J., Crossman, K.A., Jin, W., Mengersen, K., Osborne, K., Simpson, M.J., Thompson, A.A., Wu, P., Ortiz, J.-C., 2021. Identification of two-phase recovery for interpretation of coral reef monitoring data. J. Appl. Ecol. https://doi.org/10.1111/1365-2664.14039.

West, G.B., Brown, J.H., Enquist, B.J., 2001. A general model for ontogenetic growth. Nature 413, 628–631. https://doi.org/10.1038/35098076.