

Process noise distinguishes between indistinguishable population dynamics

Matthew J. Simpson^{*1}, Jacob M. Ryan¹, James M. McGree^{1,2}, and Ruth E. Baker³

¹School of Mathematical Sciences, Queensland University of Technology (QUT), Brisbane, Queensland 4001, Australia

²ARC Centre of Excellence for Mathematical and Statistical Frontiers, QUT, Australia.

³Mathematical Institute, University of Oxford, Oxford, OX2 6GG, United Kingdom

January 29, 2019

Abstract

Model selection is becoming increasingly important in mathematical biology. Model selection often involves comparing a set of observations with predictions from a suite of continuum mathematical models and selecting the model that provides the best explanation of the data. In this work we consider the more challenging problem of model selection in a stochastic setting. We consider five different stochastic models describing population growth. Through simulation we show that all five stochastic models gives rise to classical logistic growth in the limit where we consider a large number of identically prepared realisations. Therefore, comparing

^{*}To whom correspondence should be addressed. E-mail: matthew.simpson@qut.edu.au

mean data from each of the models gives indistinguishable predictions and model selection based on population-level information is impossible. To overcome this challenge we extract *process noise* from individual realisations of each model and identify properties in the process noise that differ between the various stochastic models. Using a Bayesian framework, we show how process noise can be used successfully to make a probabilistic distinction between the various stochastic models. The relative success of this approach depends upon the identification of appropriate summary statistics and we illustrate how increasingly sophisticated summary statistics can lead to improved model selection, but this improvement comes at the cost of requiring more detailed summary statistics.

Keywords: logistic growth; model selection; population dynamics

1 Introduction

The logistic growth equation,

$$\frac{dC(t)}{dt} = \lambda C(t) \left[1 - \frac{C(t)}{K} \right], \quad (1)$$

with solution,

$$C(t) = \frac{KC(0)}{C(0) + e^{-\lambda t}(K - C(0))}, \quad (2)$$

is probably the most widely used model of population dynamics in the fields of mathematical biology and mathematical ecology (Edelstein-Keshet 2005; Murray 2002). In this continuum framework, $C(t) \geq 0$ is the number, or number density, of individuals, $\lambda > 0$ is the low-density growth rate and $K > 0$ is the carrying capacity of the environment. The key feature of the logistic growth model is that when the initial population density is small compared to the carrying capacity, $C(0)/K \ll 1$, the solution takes the form of a sigmoid curve that monotonically increases from $C(0)$ and asymptotes to $\lim_{t \rightarrow \infty} C(t) = K^-$.

For small initial densities, $C(0)/K \ll 1$, the logistic growth model can be approximated by a simpler exponential growth model since $C(t) \sim C(0)e^{\lambda t}$ for sufficiently small t (Warne et al. 2017). In the exponential growth model, the per capita growth rate is a positive constant, $(dC(t)/dt)(1/C(t)) = \lambda$, that is independent of the population density. This feature leads to unbounded growth as t increases. In contrast, the per capita growth rate for the logistic model, $(dC(t)/dt)(1/C(t)) = \lambda[1 - C(t)/K]$, is a linearly decreasing function of $C(t)$, such that the per capita growth rate is zero at $C(t) = K$, which leads to bounded populations at large time. This kind of transition between low-density exponential growth and high-density saturation are the two key elements of the logistic growth model that make it so appealing since these features are widely observed in many applications in biology and ecology (Edelstein-Keshet 2005; Murray 2002).

In a biological context, images in Figure 1(a)-(b) show the growth of a population of prostate cancer cells in two-dimensional cell culture at $t = 0$ and $t = 48$ h, respectively (Jin et al. 2016). This kind of growth process becomes limited by available space in the monolayer as the density increases, as illustrated in Figure 1(c). This reduced growth rate at larger densities is often modelled using Equations (1)-(2) (Browning et al. 2018; Cai

et al. 2007; Sengers et al. 2007; Tremel et al. 2009). In an ecological context, the image in Figure 1(d) shows a population of polyps growing on the surface of an oyster shell. Similar to the population of cells in Figure 1(a)-(b), the net growth rate of the polyps reduces as the density increases, as illustrated in Figure 1(e). This kind of saturation can be accurately predicted using the logistic growth model (Melica et al. 2014) as illustrated in Figure 1(e). Many other kinds of ecological growth processes, such as populations of plants, are also modelled using Equations (1)-(2) (Bolker and Pacala 1997; Law et al. 2003). Far more complicated organism-level growth phenomena also exhibit saturation behaviour as illustrated by images in Figure 1(f)-(g) showing the growth of a Shiba Inu puppy at age 12 and 30 weeks, respectively. The time series showing the weight of the puppy, given in Figure 1(h), takes the form of a sigmoid curve that can be modelled using Equations (1)-(2). We note, however, that while whole organism-level growth can be approximately modelled using the classical logistic equation, as illustrated in Figure 1(h), other generalised logistic growth laws are often used to study organism-level growth (Frascoli et al. 2014; Goriely 2017; Tsoularis and Wallace 2002).

Instead of working with a continuum-level description of population dynamics, such as Equations (1)-(2), it is often the case in mathematical biology and mathematical ecology that stochastic discrete models are used (Bolker and Pacala, 1997; Codling et al. 2008; Ermentrout and Edelstein-Keshet 1993; Lambert et al. 2018; Law et al. 2003; Maclaren et al. 2015). Individual-based stochastic models are sometimes preferred over continuum models since continuum models can not provide information about individual-level behaviour and continuum models do not capture or predict stochastically (Frascoli et al. 2013; Treloar et al. 2014; O’Dea and King 2012).

In this section we describe five different discrete models of population dynamics and explain their relationship to Equations (1)-(2). In our description of these models we pay careful attention to illustrate the different mechanisms inherent in each model, and we describe how each model can be simulated to produce stochastic realisations. Using results of these stochastic realisations, we demonstrate that each of the different discrete models gives rise to indistinguishable logistic growth in the limit of a very large number of identically prepared realisations. Therefore, when we are dealing with noisy individual realisations of the discrete models we are considering random variables. In contrast, when

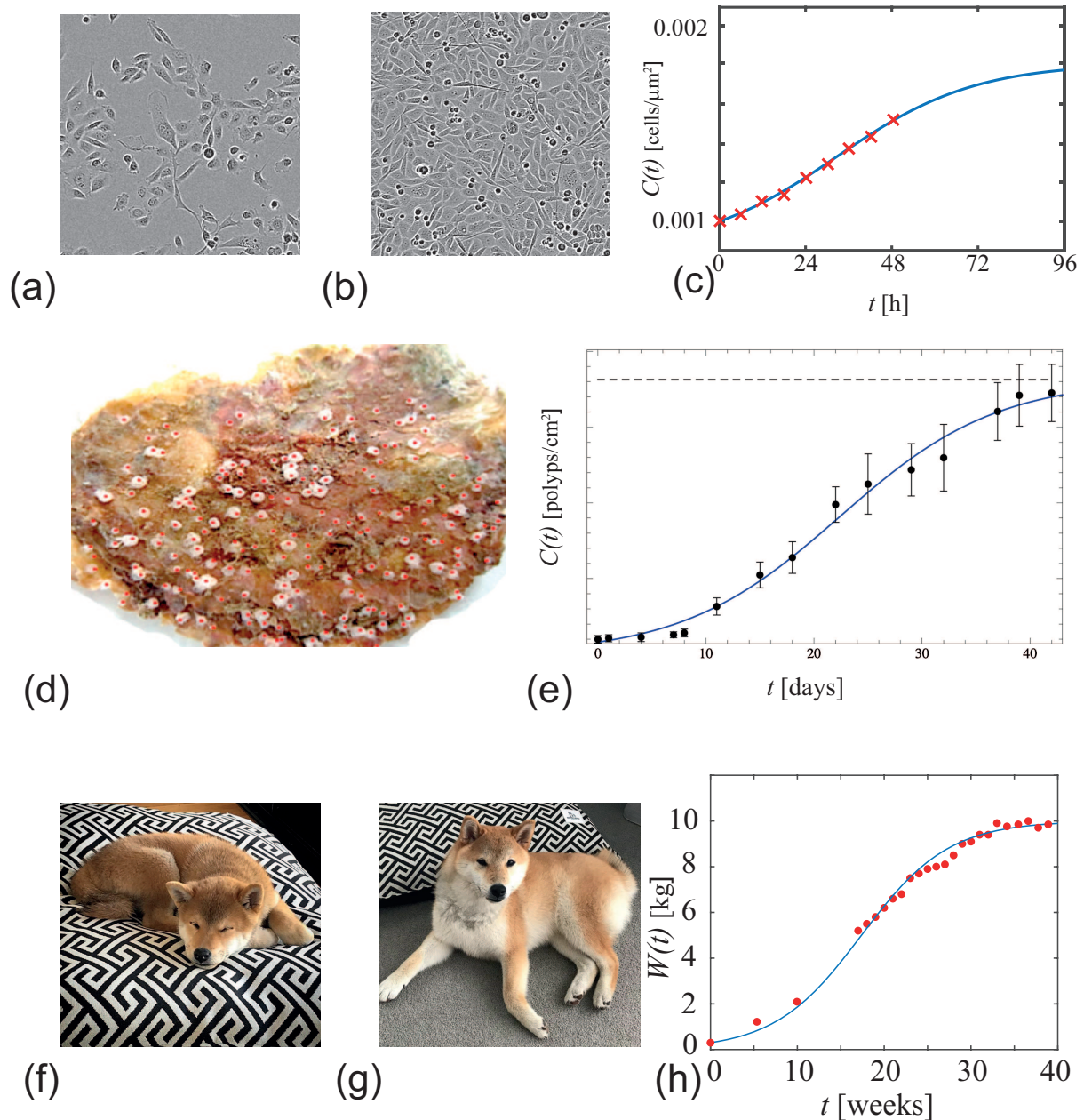


Figure 1: **Examples of logistic growth-type phenomena.** (a)-(b) shows images of a cell proliferation assay performed with prostate cancer cells [23]. The images in (a) and (b) correspond to $t = 0$ and $t = 48$ h, respectively, and each image shows a region that is approximately $450 \mu\text{m}^2$ in area. The plot in (c) shows experimental estimates of cell density, $C(t)$, (red crosses) superimposed on a plot of Equation (2) with $C(0) = 0.001$ cells/ μm^2 , $K = 0.002$ cells/ μm^2 and $\lambda = 0.04$ /h. (d) shows an oyster shell, upon which a population of polyps (red) are growing. The plot in (e) shows experimental estimates of the polyp density, $C(t)$, superimposed on a plot of Equation (2) with $C(0) = 0.081$ polyps/ cm^2 , $K = 1.81$ polyps/ cm^2 and $\lambda = 0.006$ /h. Images in (f) and (g) show photographs of a Shiba Inu puppy *Frankie*, aged 12 and 35 weeks, respectively. Data in (h) show time series measurements of *Frankie*'s weight, $w(t)$ (red dots) superimposed on a plot of Equation (2) with $C(0) = 0.35$ kg, $K = 10.7$ kg and $\lambda = 0.00113$ /h. Images and data in (a)-(c) are reproduced with permission from Jin et al. (2016). Images and data in (d)-(e) are reproduced with permission from Melica et al. (2014). Images in (f) and (g), and data in (h) were obtained by the first author.

we consider the averaged outcome of a large number of identically prepared realisations of the discrete models we are dealing with deterministic variables. To make this distinction clear we denote the number of individuals in any realisation from any one of the five discrete models as $\mathcal{C}(t)$, and the average number of individuals constructed using a large number of identically prepared realisations as

$$\langle \mathcal{C}(t) \rangle = \frac{1}{N} \sum_{n=1}^N \mathcal{C}_n(t), \quad (3)$$

where $\mathcal{C}_n(t)$ is the number individuals in the n^{th} identically prepared realisation of a particular stochastic model at time t , and N is the total number of realisations used to construct the averaged density.

2 Discrete models

We will now describe and implement the five completely distinct stochastic models that we consider in this work. The first three models are *spatially implicit* models meaning that we consider the population of individuals as being spatially well-mixed but we do not explicitly consider the spatial location of any particular individual (Geritz and Kisdi, 2012). In contrast, the last two models are *spatially explicit* and we explicitly track the position of each individual within the population during the simulations (Baker and Simpson, 2010; Simpson et al. 2010). In the spatially explicit models we initially enforce the individuals in the population to be well mixed. At the beginning of each simulation, individuals are placed uniformly, at random, taking care to ensure that no two individuals occupy the same location. Each of the five discrete models will be simulated using the Gillespie algorithm (Gillespie, 1977), and we will now describe these models and the key variables associated with each model.

2.1 Model 1: Spatially implicit birth-only

The first model considers a population of individuals, $\mathcal{C}(t)$, that undergoes a birth process where the net birth rate is a linearly-decreasing function of the total density. This gives

rise to a per capita growth rate of

$$\frac{1}{\mathcal{C}(t)} \frac{d\mathcal{C}(t)}{dt} = b_1 \left(1 - \frac{\mathcal{C}(t)}{\kappa_1} \right), \quad (4)$$

where $b_1 > 0$ is the low density birth rate and $\kappa_1 > 0$ is the carrying capacity of the environment (Geritz and Kisdi, 2012).

2.2 Model 2: Spatially implicit birth-death

The second model considers a population of individuals, $\mathcal{C}(t)$, that undergoes a combined birth-death process. Here, the net birth rate is a linearly-decreasing function of the total density and the death rate is constant. This gives rise to a per capita growth rate of

$$\frac{1}{\mathcal{C}(t)} \frac{d\mathcal{C}(t)}{dt} = b_2 \left(1 - \frac{\mathcal{C}(t)}{\kappa_2} \right) - d_2, \quad (5)$$

where $b_2 > 0$ is the low density birth rate, $d_2 > 0$ is the death rate, and $\kappa_2 > 0$ is the carrying capacity of the environment (Geritz and Kisdi, 2012). Since we are interested in populations that grow, as opposed to populations that become extinct, we will focus on parameter choices where $d_2/b_2 \ll 1$.

2.3 Model 3: Spatially implicit birth-death-annihilation

The third model considers a population of individuals, $\mathcal{C}(t)$, that undergoes a combination of birth, death and annihilation processes. Here, the birth and death rates are constants, and the annihilation mechanism can be thought of representing the case where two individuals compete with each other for some kind of resource, leading to the destruction of one of the individuals. Together, these three mechanisms give rise to a per capita growth rate of

$$\frac{1}{\mathcal{C}(t)} \frac{d\mathcal{C}(t)}{dt} = b_3 - d_3 - a_3\mathcal{C}(t), \quad (6)$$

where $b_3 > 0$ is birth rate, $d_3 > 0$ is the death rate and $a_3 > 0$ is the rate of loss due to annihilation (Geritz and Kisdi, 2012). Again, since we are interested in the case of growing populations we will focus on parameter choices where $d_3/b_3 \ll 1$ and $a_3/b_3 \ll 1$.

2.4 Model 4: Spatially explicit birth-only

Unlike the first three models where we do not consider the spatial location of individual agents in the population, we now consider a spatially explicit model where individuals within the population reside on a three-dimensional square lattice, with lattice spacing Δ , and the domain is a cube with side length L so that the total number of lattice sites is $(L/\Delta)^3$. All simulations involve periodic boundary conditions and the model is an *exclusion process* meaning that each lattice site can be occupied by, at most, a single individual (Liggett, 1999; Simpson et al. 2010). All simulations are initialised by choosing a particular number of individuals, $\mathcal{C}(0)$, and distributing them uniformly, at random, across the lattice. Each time we initialise the model we take care to ensure that each site is occupied by no more than a single agent. During the simulations agents undergo one of two types of events. First, agents undergo movement events, at rate $m_4 > 0$ (Baker and Simpson, 2010). During a potential movement event, the agent will attempt to move to one of the six nearest neighbour lattice sites with equal probability. If the chosen target site is occupied then that potential event will be aborted. Second, agents undergo birth events, at rate $b_4 > 0$. During a potential birth event the mother agent will attempt to place a daughter agent on one of the six nearest neighbour lattice sites with equal probability. If those chosen target site is occupied then that potential proliferation event is aborted (Baker and Simpson, 2010).

Previously, Baker and Simpson (2010) explain how to construct the continuum limit of this model and show that, for this well-mixed (translationally invariant) initial condition the continuum limit, written in terms of the per capita growth rate, can be written as

$$\frac{1}{\mathcal{C}(t)} \frac{d\mathcal{C}(t)}{dt} = b_4 \left(1 - \frac{\mathcal{C}(t)}{\kappa_4} \right), \quad (7)$$

where $b_4 > 0$ is birth rate and $\kappa_4 = (L/\Delta)^3$ is the total number of lattice sites. We note that while the movement rate, m_4 , does not appear explicitly in the continuum limit description, we have the additional requirement that $b_4/m_4 \ll 1$ for the continuum limit description to hold, and more details about this condition are given by Baker and Simpson (2010).

2.5 Model 5: Spatially explicit birth-death

The fifth and final model we consider is an extension of the spatially explicit birth only model, except that now we consider both birth and death events. Again, we consider a spatially explicit model where individuals within the population reside on a three-dimensional square lattice, with lattice spacing Δ , and the domain is a cube with side length L so that the total number of lattice sites is $(L/\Delta)^3$. All simulations involve periodic boundary conditions and the model is an exclusion process meaning that each lattice site can be occupied by, at most, a single individual. All simulations are initialised by choosing a particular number of individuals, $\mathcal{C}(0)$, and randomly distributing them across the lattice, taking care to ensure that each site is occupied by no more than a single agent. During the simulations agents undergo three types of events. Firstly, agents undergo movement events, at rate $m_5 > 0$, in exactly the same way as for model 4. Second, agents undergo birth events, at rate $b_5 > 0$, in exactly the same way as for model 4. Finally, agents undergo death events, at rate $d_5 > 0$, where agents are simply removed from the lattice.

Again, Baker and Simpson (2010) explain how to construct the continuum limit of this model and show that for this translationally invariant initial condition the continuum limit, written in terms of the per capita growth rate, is given by

$$\frac{1}{\mathcal{C}(t)} \frac{d\mathcal{C}(t)}{dt} = b_5 \left(1 - \frac{\mathcal{C}(t)}{\kappa_5} \right) - d_5, \quad (8)$$

where $b_5 > 0$ is birth rate, $d_5 > 0$ is the death rate, and $\kappa_5 = (L/\Delta)^3$ is the total number of lattice sites. Again, while the movement rate, m_5 , does not appear explicitly in the continuum limit description, we have the additional requirement that $b_5/m_5 \ll 1$ and $d_5/m_5 \ll 1$ for the continuum limit description to hold (Baker and Simpson, 2010). Furthermore, since we are interested in population growth, we focus on parameter choices with $d_5/b_5 \ll 1$.

3 Results and Discussion

3.1 Re-scaling and population-level indistinguishability

Each of the models 1-5 are now simulated using the Gillespie algorithm, and in each case we always observe that the population evolves from $\mathcal{C}(0)$ to approach some long-time positive steady state population density $\lim_{t \rightarrow \infty} \mathcal{C}(t) > 0$. Some re-scaling of the parameters in Equations (4)-(8), according to the relationships summarised in Table 1, suggests that despite the major individual-level differences in each of the five discrete models, the average behaviour we expect to see from all five stochastic models is neatly described by Equations (1)-(2).

Table 1: Re-scaling of the parameters in each discrete model.

Discrete Model	λ	K
1	b_1	κ_1
2	$b_2 - d_2$	$\frac{\kappa_2(b_2 - d_2)}{b_2}$
3	$b_3 - d_3$	$\frac{b_3 - d_3}{a_3}$
4	b_4	$\left(\frac{L}{\Delta}\right)^3$
5	$b_5 - d_5$	$\left(\frac{L}{\Delta}\right)^3 \frac{b_5 - d_5}{d_5}$

With this framework we are now in a position to select biologically-relevant parameter values in Equations (1)-(2) and perform a suite of simulations from the five discrete models, using the equivalent parameters in Table 1, and make a comparison of the predictions of the stochastic and continuum models. To select the parameter values in Equations (1)-(2) we focus on population growth in the context of biological cells (Browning et al. 2018; Cai et al. 2007; Sengers et al. 2007). We note that a typical *in vivo* cell doubling time is approximately 20-30 h, giving an estimate of $\lambda = 0.02 - 0.03$ /h (Jin et al. 2016). In contrast, two-dimensional *in vitro* experiments, with plentiful nutrients and oxygen, are associated with a much faster doubling time of approximately 14 h, giving an estimate

of $\lambda = 0.05$ /h (Jin et al. 2016). Therefore, in this work we will consider biologically relevant upper and lower bounds on the value of λ . In the main document we consider results for $\lambda = 0.01$ /h and we present a second set of results for $\lambda = 0.05$ /h in the Supplementary Material document. In all cases we set $C(0) = 100$ and $K = 1000$ so that we consider net population growth of almost an order of magnitude. Furthermore, when we present results graphically we always plot $\mathcal{C}(t)/K$ and $C(t)/K$ as a function of time so that the results are presented in terms of population densities relative to the carrying capacity.

Results in Figure 2 show single realisations of each stochastic model and we compare the evolution of $\mathcal{C}(t)/K$ from each realisation with $C(t)/K$ from Equations (1)-(2). In each case we see that stochastic models are noisy but the agreement with the continuum solution is clear. This match between the simulation data and the solution of the continuum model suggests that if we are provided with data showing $\mathcal{C}(t)$ or $\mathcal{C}(t)/K$, as in Figure 1(c), (e) and (h), it would be very difficult to distinguish between which of the five discrete models provides the best explanation of the experimental data.

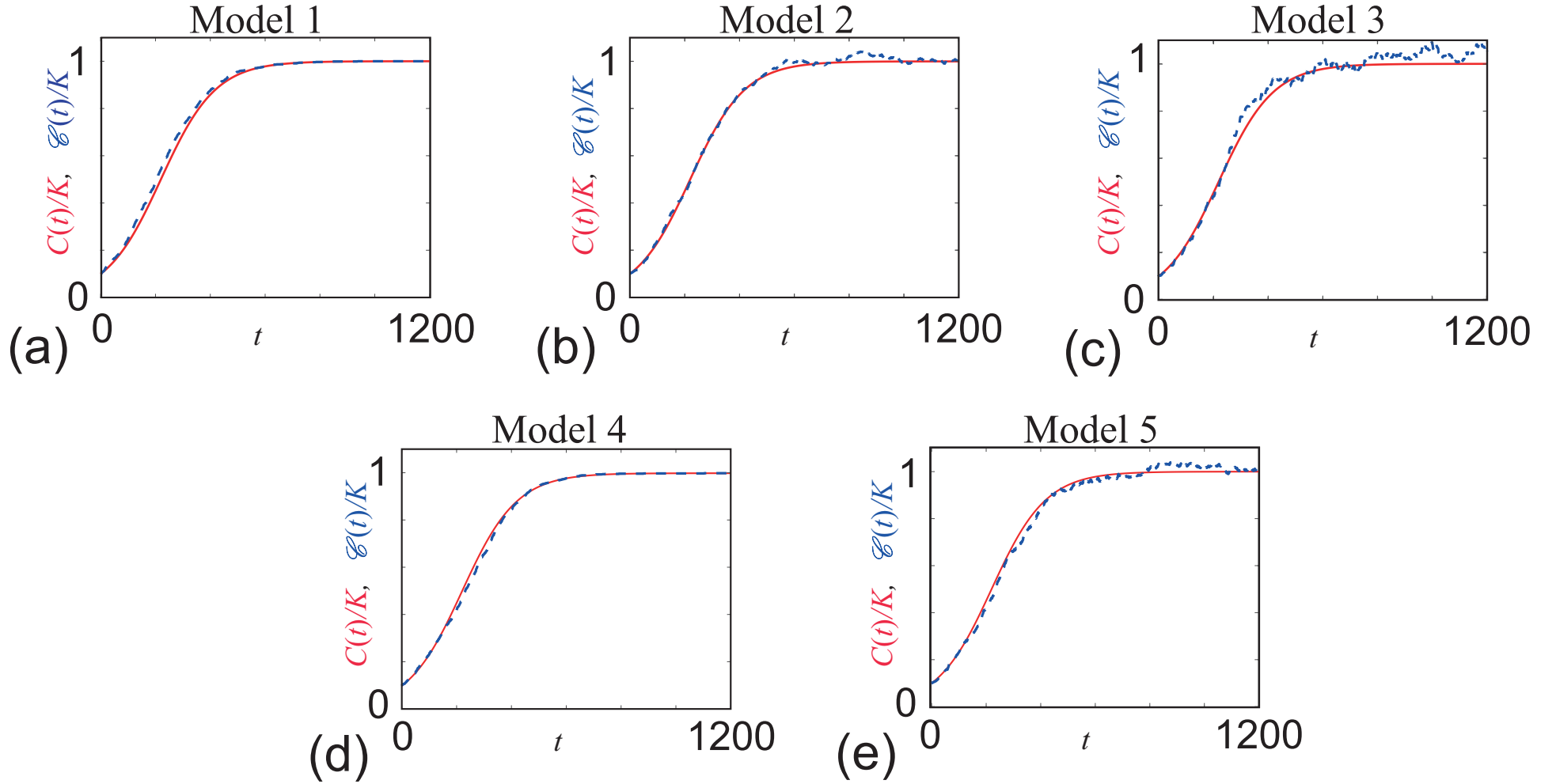


Figure 2: Comparison of the solution of the classical logistic equation, $C(t)/K$, with single realisations of the five different stochastic models, $C(t)/K$. The continuum-discrete match for models 1-5 are shown in (a)-(e), respectively, with $C(0) = 100$, $K = 1000$ and $\lambda = 0.01$ /h. Parameters in the five discrete models are: $b_1 = 0.01$ /h, $\kappa_1 = 1000$; $b_2 = 0.0125$ /h, $d_2 = 0.0025$ /h, $\kappa_2 = 1250$; $b_3 = 0.0125$ /h, $d_3 = 0.0025$ /h, $a_3 = 0.00001$ agents/h; $L = 10$, $\Delta = 1$, $b_4 = 0.01$ /h, $m_4 = 1$ /h; $L = 10$, $\Delta = 1$, $b_5 = 0.0125$ /h, $d_5 = 0.0025$ /h, $m_5 = 1$ /h.

The challenge of model selection becomes far more acute if we take the standard approach to interpret noisy data from a stochastic model and consider averaging the results from the discrete model over many identically prepared realisations. In any single realisation of any of the discrete models, events occur at random times and the time between events is exponentially distributed (Gillespie, 1977). To construct averaged density profiles we first interpolate the discrete time population data to give a continuous description of $\mathcal{C}_n(t)$. To interpolate the discrete time data we use MATLABs previous interpolation scheme (Mathworks 2019a) to give continuous representations of $\mathcal{C}_n(t)$ for $0 \leq t \leq 1200$. We then construct averaged density profiles by averaging the values of $\mathcal{C}_n(t)$ at 1201 equally spaced times, $t = 0, 1, 2, \dots, 1200$, according to Equation (3). Results in Figure 3 compare the solution of Equations (1)-(2) with averaged data from each discrete model, $\langle \mathcal{C}(t) \rangle$, with $N = 1000$. In this case we see that $C(t)$ and $\langle \mathcal{C}(t) \rangle$ are visually indistinguishable at this scale.

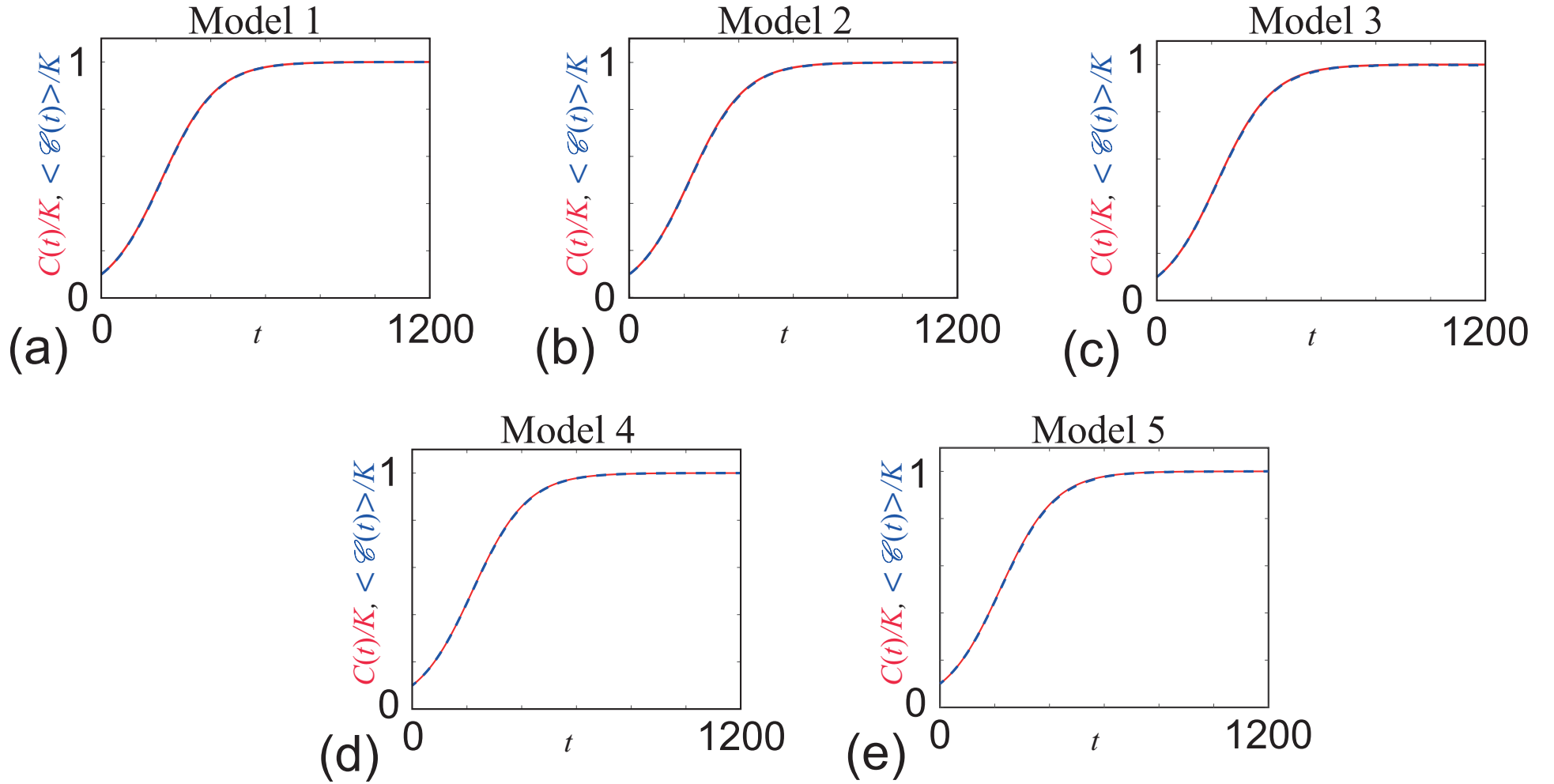


Figure 3: Comparison of the solution of the classical logistic equation, $C(t)/K$, with averaged data from 1000 identically prepared realisations of the five different stochastic models, $\langle C(t) \rangle / K$. The continuum-discrete match for models 1-5 are shown in (a)-(e), respectively, with $C(0) = 100$, $K = 1000$ and $\lambda = 0.01$ /h. Parameters in the five discrete models are: $b_1 = 0.01$ /h, $\kappa_1 = 1000$; $b_2 = 0.0125$ /h, $d_2 = 0.0025$ /h, $\kappa_2 = 1250$; $b_3 = 0.0125$ /h, $d_3 = 0.0025$ /h, $a_3 = 0.00001$ agents/h; $L = 10$, $\Delta = 1$, $b_4 = 0.01$ /h, $m_4 = 1$ /h; $L = 10$, $\Delta = 1$, $b_5 = 0.0125$ /h, $d_5 = 0.0025$ /h, $m_5 = 1$ /h.

Before we explore the question of whether it is possible to reliably distinguish between the five different stochastic models, it is worthwhile to point out that all plots in Figures 2 and 3 show the independent variable for $0 \leq t \leq 1200$ h. While in principle it is possible to perform simulations for a longer period, the solution of Equation (1) gives $C(1200)/K \approx 0.9999$ for $\lambda = 0.01$ /h with $C(0)/K = 0.1$. Therefore, this truncated time interval captures virtually the entire dynamics of the population and we do not consider any longer periods of time for this choice of λ and $C(0)$.

Overall, the results in Figure 3 have two important consequences:

1. for each discrete model, we see that $\langle \mathcal{C}(t) \rangle \rightarrow C(t)$ as the number of identically prepared realisations of the discrete model becomes sufficiently large, as expected (Baker and Simpson 2010). Additional results (Supplementary Material) confirms that this is also true for larger λ . Additional results (not shown) confirm that this also holds when we vary $C(0)$;
2. if we consider generating $\langle \mathcal{C}(t) \rangle$ using a large number of identically prepared realisations, it is not possible to use this averaged data to distinguish which of the five very different stochastic models gave rise to that averaged population data.

3.2 Individual-level distinguishability

Given these observations, the task in this study is to attempt to distinguish between the five different stochastic models. The question of *model selection* is becoming increasingly important in the field of mathematical biology (Jin et al. 2017) and mathematical ecology (Johnson and Omland, 2004). For example, Sarapata and de Pillis (2014) compare a range of continuum sigmoid growth laws with data describing the temporal growth of various types of tumours and illustrate that some of the experimental datasets can be modelled using more than one choice of continuum model. Gerlee (2013) presents a thorough discussion of several different continuum models that are used to predict and explain tumour growth. Similarly, West et al. (2001) attempts to define a universal growth law to describe temporal tumour growth in various animal models. While these three previous studies focus on ordinary differential equations and temporal behaviour,

similar and more complicated questions of model selection are relevant when considering combined spatial and temporal processes, such as collective cell spreading (Treloar et al. 2014). These questions have been partially explored in the context of modelling collective cell spreading processes by attempting to select the best model among a suite of distinct continuum partial differential equation models (e.g. Maini et al. 2004a; Maini et al. 2004b; Sengers et al. 2007; Sherratt and Murray 1990; Warne et al. 2018).

These previous studies in model selection are very different to the problem that we consider in this work. In these previous studies the authors compare a suite of distinct continuum models and attempt to use experimental data to select the most appropriate model. In contrast, here we consider a suite of different stochastic models, each of which has the same continuum limit description, and we attempt to distinguish between the different stochastic models. Therefore, our task is to select between different stochastic models when the continuum limit is mathematically indistinguishable. This is a challenging task, especially when we consider that standard protocols in the mathematical biology literature for interpreting results from stochastic models is to construct averaged model results by averaging results from many identically prepared realisations of the stochastic model (e.g. Baker et al. 2010; Bruna and Chapman 2012; Bruna and Chapman 2014; Dyson et al. 2013; Flegg et al. 2013; Keeling and Eames 2005). Similarly, standard protocols for interpreting noisy experimental data is to deal with averaged quantities (Jin et al. 2016; Melica et al. 2014; Pozzobon and Perré 2018; Sengers et al. 2007; Vo et al. 2015). Clearly, averaging provides no advantage when the continuum limit description of various discrete models are identical. To deal with this complication we will take the opposite approach and instead of focusing on averaging noisy data, we will examine properties of the *process noise* and explore the extent to which the inherent process noise facilitates model selection.

Here, we focus on the process noise associated with each stochastic model. To do this we consider the quantity

$$\varepsilon(t) = \mathcal{C}(t) - \langle \mathcal{C}(t) \rangle, \quad (9)$$

which is a time-dependent measure of the process noise that can be calculated and visualised for any realisation of any one of the five stochastic models we consider. Typical plots of $\varepsilon(t)$ are given in Figure 4 for each of the five stochastic models, and in each case

we show four different realisations of $\varepsilon(t)$ for each model. Clearly each realisation of $\varepsilon(t)$ acts like a random variable, however some model-specific trends in $\varepsilon(t)$ are clear. One obvious trend is that the value of $\varepsilon(t)$ for Model 3 appears to deviate further from zero than all other models. Another trend in the $\varepsilon(t)$ data is that $\varepsilon(t)$ for Model 1 and Model 4 appear to decay to zero at late times for each realisation whereas $\varepsilon(t)$ for Models 2, 3 and 5 does not. Our aim now is to explore whether it is possible to use these individual-level differences to provide a probabilistic distinction between the five models. Furthermore, we will also explore which features of the $\varepsilon(t)$ signal to use to make this distinction as clear as possible. To this end we will take a Bayesian approach (Gelman et al. 2003; Warne et al. 2019b).

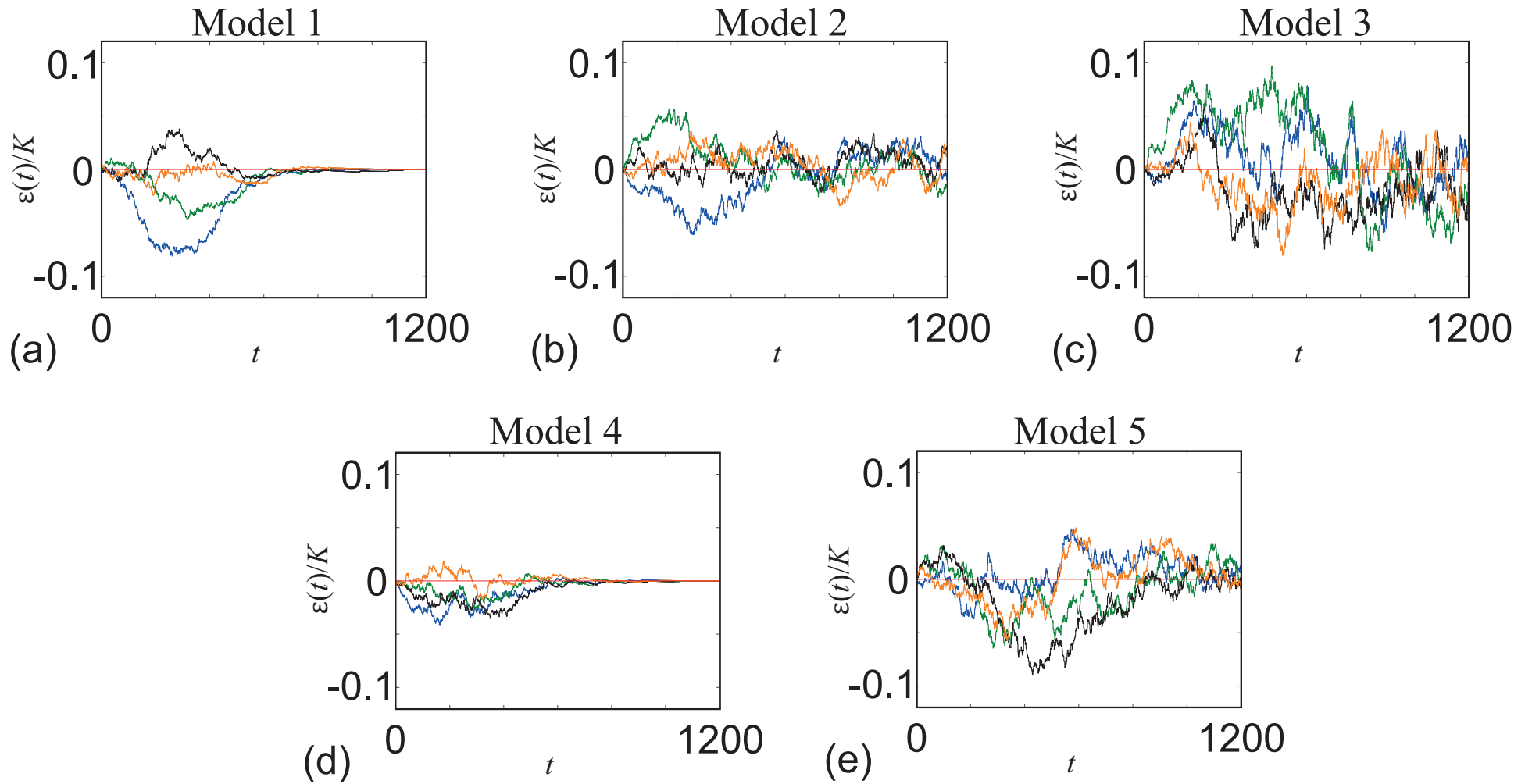


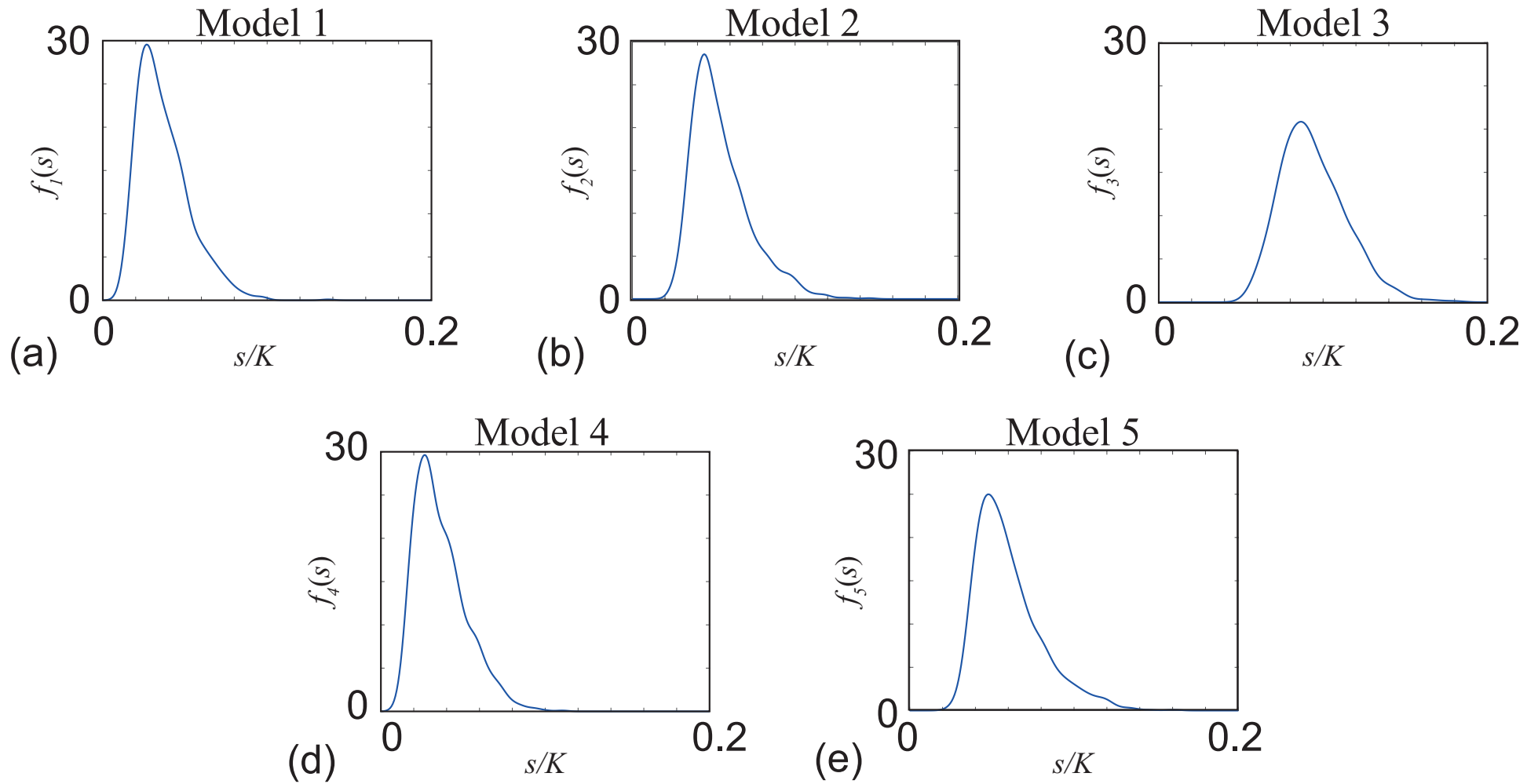
Figure 4: **Time series data showing $\varepsilon(t)/K$ for four realisations for each of the five models.** The time series data for models 1-5 are shown in (a)-(e), respectively, with $\mathcal{C}(0) = 100$, $K = 1000$ and $\lambda = 0.01$. For each stochastic model, four typical realisations of $\varepsilon(t)$ are shown in green, blue, black and orange.

3.3 Bayesian approach to distinguish models using process noise

As we will show, the key aspect of using process noise to distinguish between the five stochastic models is to use the process noise signal, $\varepsilon(t)$, to construct summary statistics to facilitate making a reliable distinction. For the purpose of clarity we will first illustrate some key features using a simple summary statistic, and then consider refining our choice of summary statistic to refine our model selection. Perhaps the simplest way to summarise the noise signature is use the maximum deviation from zero noise,

$$s = \max_{t \in [0, 1200]} |\varepsilon(t)|. \quad (10)$$

This choice of summary statistic allows us to replace each time series, $\varepsilon(t)$, with a single scalar value, s , and we will now explore the extent to which using this information enables us to reliably distinguish between the five stochastic models. To achieve this we take the 1000 identically prepared realisations of each stochastic model that we used to construct averaged density data in Figure 3 and we calculate s for each of the realisations. Using MATLABs `ksdensity` function (Mathworks, 2019b) we convert the discrete distribution of s for each model into a smooth, approximate density distribution, as shown in Figure 5.



20

Figure 5: **Univariate density profiles for the five stochastic models with the simple univariate summary statistic, Equation (10).** (a)-(e) show density estimates, $f_i(s)$, as a function of the univariate summary statistic, s , given by Equation (10). Each density profile is constructed using 1000 identically prepared realisations as describing in Figure 3, and the smoothed density profile is obtained with MATLABs `ksdensity` function (Mathworks, 2019b). All results correspond to $\lambda = 0.01$ /h.

A qualitative comparison of the density profiles in Figure 5 confirms that these density profiles capture some of the putative differences that we described in Figure 4. For example, if we consider a maximum deviation of $s/K = 0.1$, the density associated with Model 1 is relatively small, $f_1(0.1) \approx 0$, whereas the density for Model 3 is relatively large, $f_3(0.1) \approx 20$. To make use of this difference in a Bayesian framework, suppose that we are given $f_i(s)$ for $i = 1, 2, \dots, 5$. If we then perform a single realisation of a randomly-selected model and calculate s from that realisation we apply Bayes theorem to give

$$\mathbb{P}(M_i|s) \propto \mathbb{P}(s|M_i)\mathbb{P}(M_i), \quad \forall i = 1, 2, \dots, 5, \quad (11)$$

where $\mathbb{P}(M_i|s)$ is the probability of model i given the summary statistic s , $\mathbb{P}(s|M_i)$ is the probability density of summary statistic s given model i , and $\mathbb{P}(M_i)$ is the prior model probability, which encodes our previous knowledge of which model is most appropriate for the data.

For all of our work we make the conservative assumption that the prior specifies that all models are equally likely, $\mathbb{P}(M_i) = 1/5$ for $i = 1, 2, \dots, 5$. We note that $\mathbb{P}(s|M_i)$ is the likelihood, and in this case we have $\mathbb{P}(s|M_i) = f_i(s)$ for $i = 1, 2, \dots, 5$. This means that for a single realisation of one of the five models chosen at random $\mathbb{P}(M_i|s)$ is simply proportional to $f_i(s)$, and we can calculate $\mathbb{P}(M_i|s)$ for each $i = 1, 2, \dots, 5$ by ensuring that $\sum_{i=1}^5 \mathbb{P}(M_i|s) = 1$. Again, for all our work, we construct the likelihood using a summary statistic, s , instead of the data, $\varepsilon(t)$. This choice is made on practical grounds as it is computationally efficient to work with lower dimensional summary statistics to summarise the data than it is to work with all data collected to construct $\varepsilon(t)$ (Maclaren et al. 2015; Maclaren et al. 2017; Lambert et al. 2018).

Instead of considering just one single realisation of a particular model chosen at random, a more realistic scenario is that we consider a small number of identically prepared realisations of a particular model. Since each identically prepared realisation of the unknown model, and the associated summary statistic, are independent, we obtain

$$\mathbb{P}(M_i|s) \propto \prod_{j=1}^J \mathbb{P}_j(s|M_i), \quad \forall i = 1, 2, \dots, 5, \quad (12)$$

under the conservative assumption that the prior specifies that each model is equally likely. Here, $\mathbb{P}_j(s|M_i)$ is the probability density that the summary statistic s , in the j^{th} identically prepared realisation is associated with model i . In this formulation we are considering J identically prepared realisations. Again, with this information we can calculate $\mathbb{P}(M_i|s)$ for each $i = 1, 2, \dots, 5$ by ensuring that $\sum_{i=1}^5 \mathbb{P}(M_i|s) = 1$.

To demonstrate the performance of this simple summary statistic we will now consider each of the five stochastic models in turn. For each model, $i = 1, 2, \dots, 5$, we take $J = 5$ identically prepared realisations of a randomly-chosen model to calculate $\mathbb{P}(M_i|s)$ according to Equation (12). We repeat this process 1000 times, giving us access to a distribution of estimates for $\mathbb{P}(M_i|s)$, from which we can calculate the sample mean and the 95% credible interval. Results in Figure 6(a) show $\mathbb{P}(M_i|s)$, for $i = 1, 2, \dots, 5$, for Model 1. Here, point estimates correspond to the sample mean of $\mathbb{P}(M_i|s)$ and the error bars denote the 95% credible interval. Figure 6(a) includes a horizontal line at $1/5$, indicating the prior distribution. In this case we see that $\mathbb{P}(M_2|s)$, $\mathbb{P}(M_3|s)$ and $\mathbb{P}(M_5|s)$ are all smaller than $1/5$, whereas $\mathbb{P}(M_1|s)$ and $\mathbb{P}(M_4|s)$ are greater than the prior values. This means that the simple summary statistic correctly indicates that Models 2, 3 and 5 are less likely to explain the summary statistic than Models 1 and 4. However, this process gives $\mathbb{P}(M_1|s) \approx \mathbb{P}(M_4|s)$, indicating that this summary statistic does not reliably distinguish between Models 1 and 4.

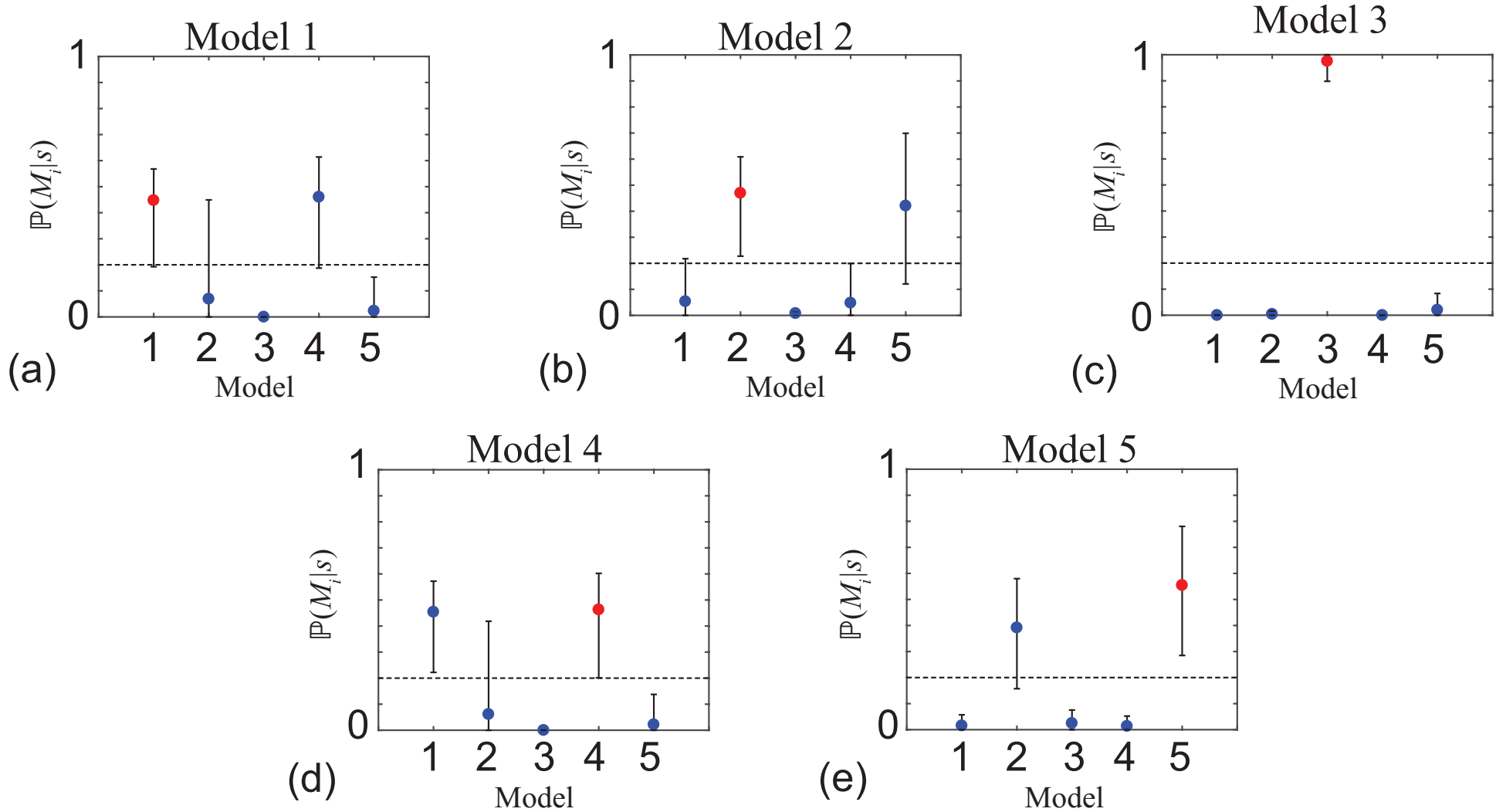


Figure 6: **Model identification with the simple univariate summary statistic, s , given by Equation (10).** Results in (a)-(e) show point estimates of $\mathbb{P}(M_i|s)$ for $i = 1, 2, \dots, 5$ and the uncertainty in the estimate is indicated by the error bars. The point estimates correspond to the sample mean and the error bar corresponds to the sample mean plus or minus one sample standard deviation, both calculated using 1000 identically prepared estimates of $\mathbb{P}(M_i|s)$ for $i = 1, 2, \dots, 5$ with $J = 5$. Each subfigure shows a horizontal line at $1/5$, indicating the prior distribution of $\mathbb{P}(M_i)$ for $i = 1, 2, \dots, 5$. In each subfigure, maximum value of $\max_{i=1,2,\dots,5} [\mathbb{P}(M_i|s)]$ is shown in red, and in each case this is the correct model choice. All results correspond to $\lambda = 0.01 /h$.

This process of calculating $\mathbb{P}(M_i|s)$, for $i = 1, 2, \dots, 5$ for each of the candidate models is repeated in Figure 6(b)-(e) for Models 2, 3, 4 and 5, respectively. Results in Figure 6(b) explore the ability of the simple summary statistic to distinguish Model 2 and we see that the process correctly identifies the Models 1, 3 and 5 are less likely to explain the data whereas Models 2 and 5 both give similar results. Results in Figure 5(c) explore the ability of the simple summary statistics to identify Model 3, and in this case we have a very promising result that $\mathbb{P}(M_3|s) \approx 1$, and all other candidate models have close to zero probability. Results in Figure 6(d) and (e) are similar to the results in Figure 6(a) and (b) since the simple summary statistic correctly assigns a lower probability to three of the candidate models, but is unable to reliably distinguish between two other candidate models.

In addition to simply calculating the sample mean and 95% credible intervals of our estimates of $\mathbb{P}_j(M_i|s)$, as reported in Figure 6, we can also visualise the entire distribution of 1000 estimates of $\mathbb{P}_j(M_i|s)$. Results in Figure 7 show histograms of $\mathbb{P}_j(M_i|s)$ for Models 1, 2, \dots , 5, respectively. In each case, we see that each histogram appears to be unimodal, which indicates that summarising these distributions in terms of the sample mean is useful and we see that some of the distributions, such as the histogram in Figure 7(c), are more peaked than the others, indicating a higher certainty in these results.

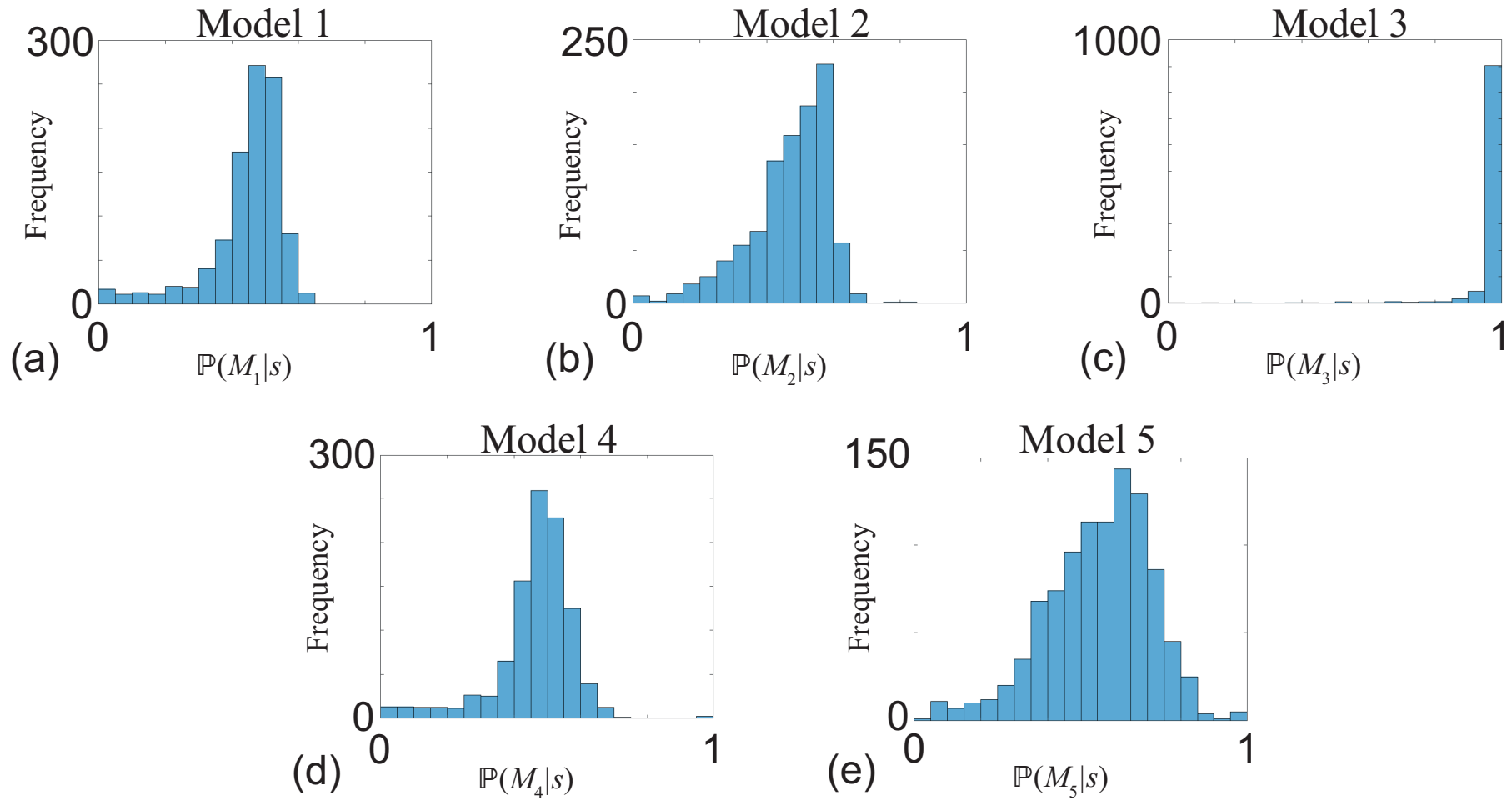


Figure 7: **Distribution of $\mathbb{P}(M_i|s)$ using the simple univariate summary statistic, s , given by Equation (10).** Results in (a)-(e) show distributions of estimates of $\mathbb{P}(M_i|s)$ calculated using 1000 identically prepared estimates of $\mathbb{P}(M_i|s)$ for $i = 1, 2, \dots, 5$ with $J = 5$. All results correspond to $\lambda = 0.01 / h$.

Results in Figures 6-7 are promising. The simplest possible summary statistic, Equation (10), can partially distinguish between the five stochastic models that are completely indistinguishable when we consider averaged data in Figures 2-3. We now explore our ability to improve these preliminary results by refining the summary statistic. A key feature of $\varepsilon(t)$, clearly evident in Figure 4, is that the temporal noise signature appears to depend on time. This suggests that improved results might be obtained by constructing a more detailed summary statistic that characterises both *early* and *late* features of $\varepsilon(t)$. To this end we partition the time interval considered, $0 < t < 1200$, into quartiles, and collect data describing the noise signature in the first quartile, $0 < t < 300$, and the fourth quartile, $900 < t < 1200$. Again, we summarise these two process noise signatures by the maximum deviation from zero, giving $s_1 = \max_{t \in [0, 300]} |\varepsilon(t)|$ and $s_4 = \max_{t \in [900, 1200]} |\varepsilon(t)|$. This means that instead of summarising the noise signature using a single number, s , we now summarise each stochastic simulation using vector with two components, (s_1, s_4) . Using the 1000 identically prepared realisations of each model to construct averaged density data in Figure 3 we calculate (s_1, s_4) for each and use the `ksdensity` function in MATLAB (Mathworks, 2019b) to form smooth, approximate bivariate density distributions associated with each stochastic model, as shown in Figure 8. Results in Figure 8 confirm that we have visually obvious differences in the bivariate distribution of (s_1, s_4) . Most notably these distributions clearly show the larger fluctuations inherent in Model 3, and we see that the late process noise in Models 1 and 4 are very different to the late process noise in Models 2 and 5.

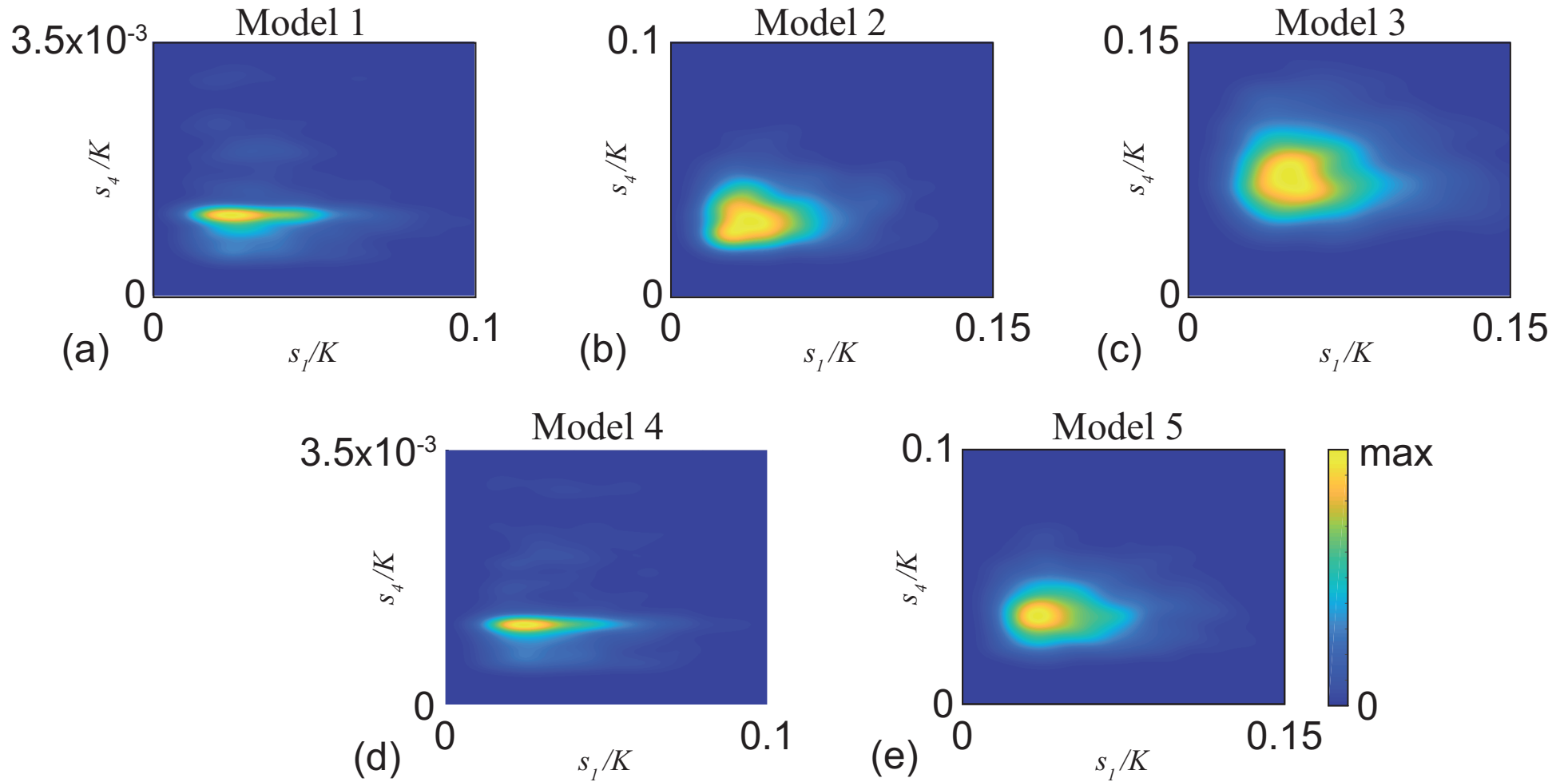


Figure 8: **Bivariate density profiles for the five stochastic models with $s_1 = \max_{t \in [0,300]} |\varepsilon(t)|$ and $s_4 = \max_{t \in [900,1200]} |\varepsilon(t)|$.** (a)-(e) show density estimates, $f_i(s_1, s_4)$, shown as a function of s_1/K and s_4/K , for $i = 1, 2, \dots, 5$, respectively. Each density profile is constructed using 1000 identically prepared realisations as describing in Figure 3, and the smoothed bivariate density profile is obtained with MATLABs `ksdensity` function (Mathworks, 2019b). All results correspond to $\lambda = 0.01$ /h.

Using the refined summary statistic densities in Figure 8, we estimate $\mathbb{P}(M_i|s)$ for each model by taking $J = 5$ identically prepared realisations of a randomly-chosen model according to Equation (12). Again, we repeat this process 1000 times, giving us a distribution of estimates for $\mathbb{P}(M_i|s)$, from which we can calculate the sample mean and the 95% credible interval. Results in Figure 9(a) show $\mathbb{P}(M_i|s)$, for $i = 1, 2, \dots, 5$, for Model 1. As for the results based on the simple summary statistic in Figure 6(a), the refined summary statistic gives $\mathbb{P}(M_2|s)$, $\mathbb{P}(M_3|s)$ and $\mathbb{P}(M_5|s)$ are all smaller than $1/5$, whereas $\mathbb{P}(M_1|s)$ and $\mathbb{P}(M_4|s)$ are greater than the prior values. This result correctly implies that the bivariate summary statistic assigns a low probability to Models 2, 3 and 5, and a higher probability to Models 1 and 4.

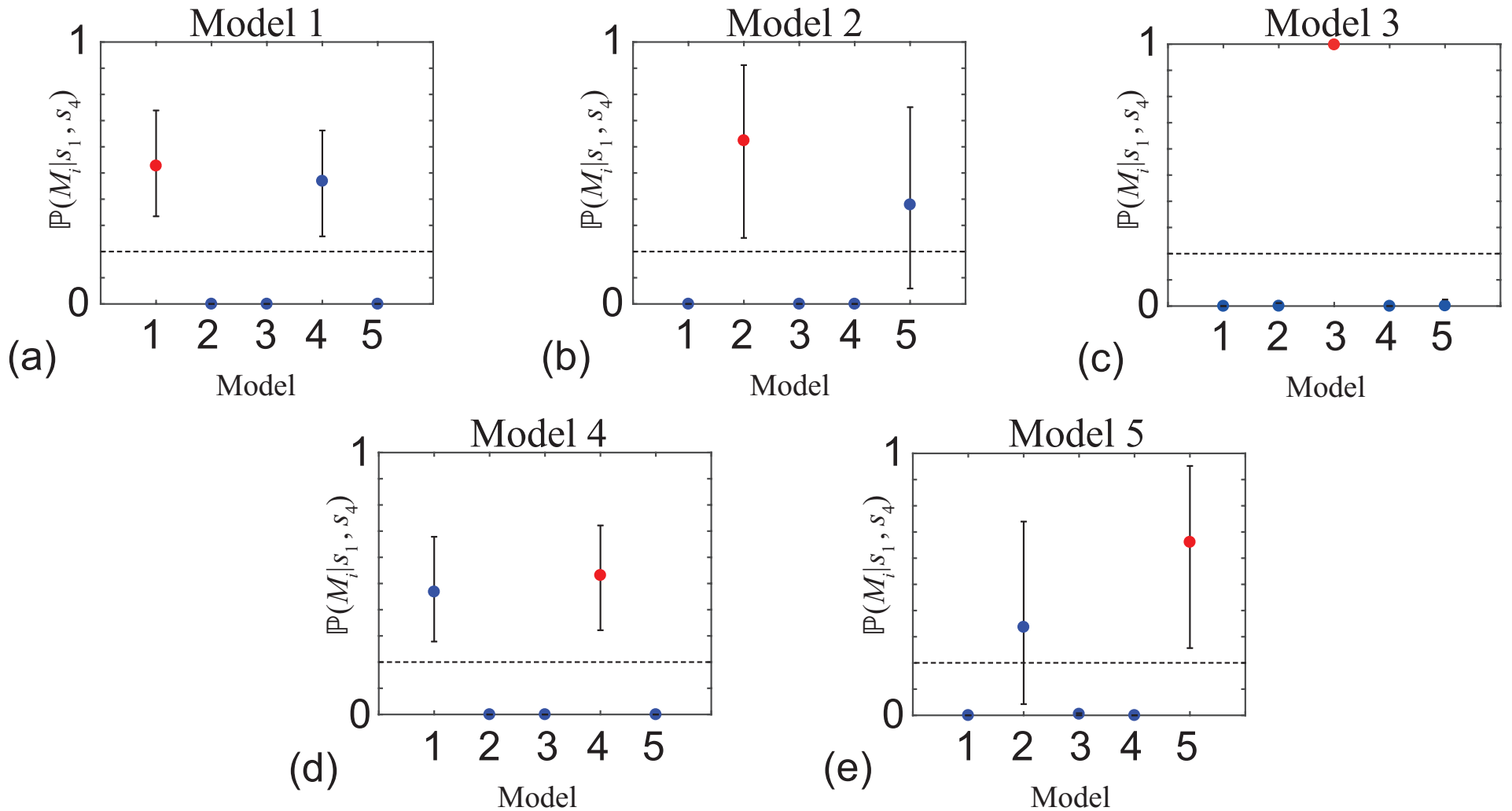


Figure 9: **Model identification with the bivariate summary statistic, (s_1, s_2) .** Results in (a)-(e) show point estimates of $\mathbb{P}(M_i | s)$ for $i = 1, 2, \dots, 5$ and the uncertainty in the estimate is indicated by the error bars. The point estimates correspond to the sample mean and the error bar corresponds to the sample mean plus or minus one sample standard deviation, both calculated using 1000 identically prepared estimates of $\mathbb{P}(M_i | s)$ for $i = 1, 2, \dots, 5$ with $J = 5$. Each subfigure shows a horizontal line at $1/5$, indicating the prior distribution of $\mathbb{P}(M_i)$ for $i = 1, 2, \dots, 5$. In each subfigure, maximum value of $\max_{i=1,2,\dots,5} [\mathbb{P}(M_i | s)]$ is shown in red, and in each case this is the correct model choice. All results correspond to $\lambda = 0.01$ /h.

Comparing results in Figure 6 and Figure 9 confirms that the additional information in the bivariate summary statistic leads to an improved ability to distinguish between some of the models. For example, if we compare results in Figure 6(b) with the results in Figure 9(b) we see that both summary statistics give point estimates of $\mathbb{P}(M_i|s) \approx 0$ and $\mathbb{P}(M_i|s_1, s_4) \approx 0$ for $i = 1, 3$ and 4 , correctly identifying that Models 1, 3 and 4 are unlikely to explain the data. Results in Figure 6(b) indicate that the simple univariate summary statistic leads to point estimates with $\mathbb{P}(M_2|s) \approx \mathbb{P}(M_5|s)$ confirming that the univariate summary statistic does not reliably distinguish between Model 2 and Model 5. In contrast, results in Figure 9(b) indicate that $\mathbb{P}(M_2|s_1, s_4) > \mathbb{P}(M_5|s_1, s_4)$ confirming that the bivariate summary statistic correctly indicates that Model 2 gives the best explanation of the summary statistic. Again, instead of relying simply on point estimates and quartile distributions in Figure 9 for the improved summary statistic, results in Figure 10 show the distributions of $\mathbb{P}(M_i|s_1, s_4)$ for each model.

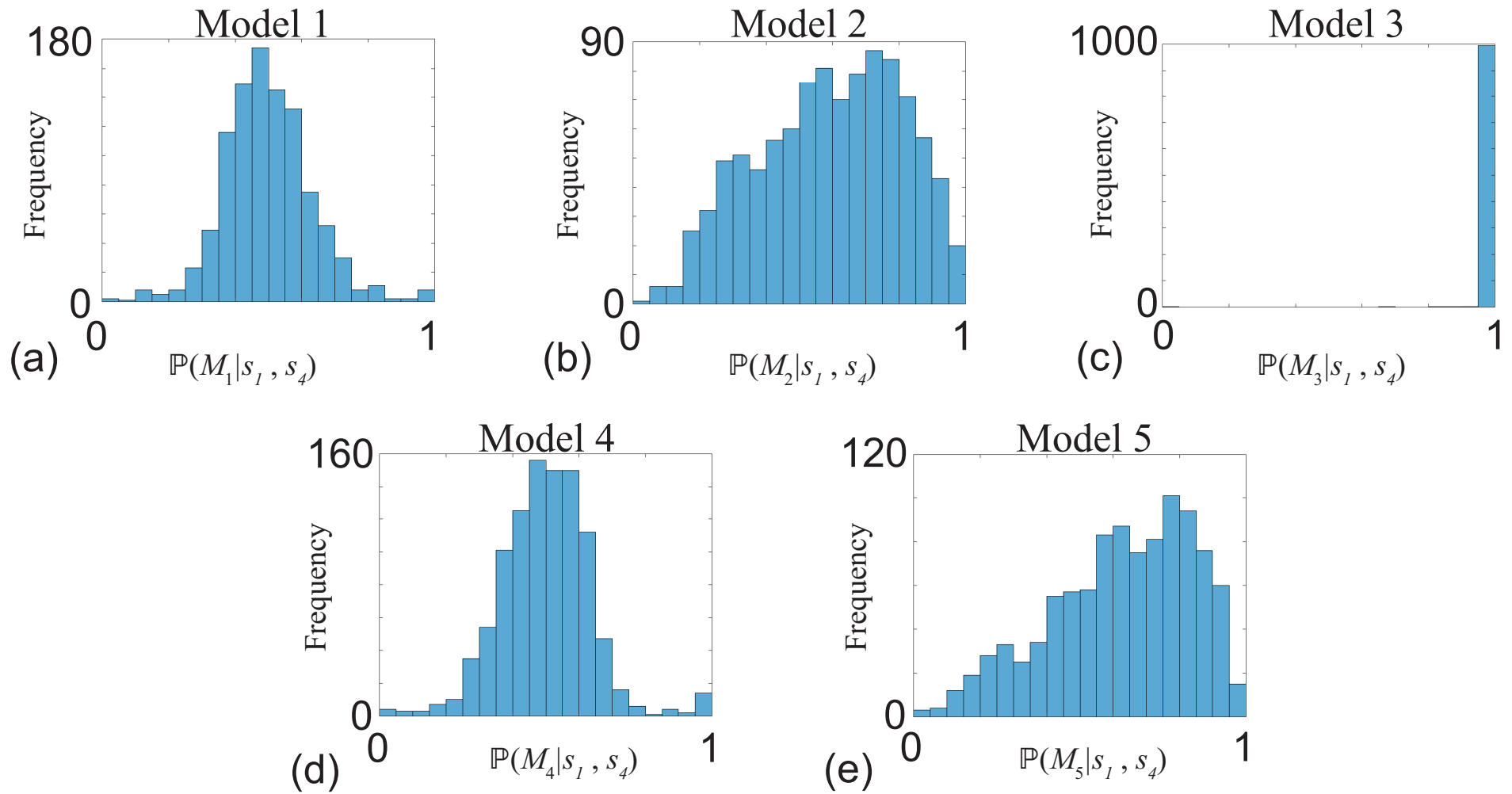


Figure 10: **Distribution of $\mathbb{P}(M_i | s)$ using the bivariate summary statistic, (s_1, s_4) .** Results in (a)-(e) show distributions of estimates of $\mathbb{P}(M_i | s)$ calculated using 1000 identically prepared estimates of $\mathbb{P}(M_i | s)$ for $i = 1, 2, \dots, 5$ with $J = 5$. All results correspond to $\lambda = 0.01$ /h.

4 Conclusions

The process of model selection is becoming increasingly of interest to the mathematical biology community. Unlike other areas of applied mathematics, such as fluid mechanics and solid mechanics, where there is consensus about what kinds of models are most appropriate, in mathematical biology it is often the case that several competing models are available to describe similar phenomena (Gerlee 2014; Sarapata and de Pillis 2014; Jin et al. 2016; Browning et al. 2017). The process of model selection often involves taking a set of observations and exploring the extent to which results from a suite of continuum models provide the best explanation of those data. Various approaches to model selection can be taken such as using maximum likelihood (Jin et al. 2014) or Bayesian approaches (Warne et al. 2018). In this work we explore a more challenging problem in model selection by working with a suite of stochastic models describing population growth processes. The main challenge is that we consider five distinct discrete models and show, through simulation, that each model is well described by the same continuum logistic growth model, given by Equations (1)-(2). If we take the usual approach and average data from many identically prepared realisations from each stochastic model that the averaged results are completely indistinguishable. This means that the usual tools used in model selection for continuum models cannot be used to distinguish between the five stochastic models that we consider.

To make progress we take a different approach and examine the *process noise* associated with each model and find that there are certain features of the process noise that appear to be different between the five different stochastic models. We demonstrate how to make a probabilistic distinction between the five models in a Bayesian framework by constructing appropriate summary statistics from the process noise. Our results show that even using the simplest possible summary statistic, the maximum deviation away from zero noise, allows us to reliably distinguish between some of the models. In particular, we find that Model 3 involving spatially implicit birth-death-annihilation is remarkably easy to distinguish from the other four models using the simplest possible summary statistic. Furthermore, we find that certain groups of models are also easy to distinguish using this simple approach. As expected, when we extend the summary statistic to include information about both the *early* and *late* portions of the process noise, we

find that our ability to distinguish between the five models is enhanced. Therefore, while all five models are completely indistinguishable using population-level information, we show that individual-level information can be used in a Bayesian framework to provide a probabilistic distinction between the various models.

There are many ways in which the work we present here can be extended. The most obvious line of extension would be to consider using different summary statistics in an attempt to further improve our model selection results. While all results in Figure 6 and Figure 9 are based on taking the maximum absolute value of ε , we also explored with other choices of summary statistic, such as working with the maximum value of ε , the minimum value of ε and measures of autocorrelation in ε . These additional results (not shown) did not improve our ability to distinguish between the give models over the results in Figure 6 and Figure 9 and so we choose to present results based on the simplest summary statistic only. Another way to potentially improve our ability to distinguish between the five models is to use higher dimensional summary statistics. Just as the results in Figure 9 for the bivariate summary statistic, (s_1, s_4) , improve our ability to distinguish between the models than the results in Figure 6 for the univariate summary statistic, s , we expect that introducing additional information into the summary statistic definition, and hence working with higher dimensional summary statistics, will improve our model selection results. Since we have already partially demonstrated that more detailed summary statistics can lead to improved model section in Figure 6 and Figure 9, we do not pursue further refinements along these lines here. Another point for consideration would be to relax our conservative assumption of working with uniform prior distributions and incorporating some uncertainty via the prior. A slightly different extension would be to make use of our observation that the process noise, $\varepsilon(t)$, is time dependent. This suggests that it would be possible to explore the question of the optimal time(s) to make observations (Dehideniya et al. 2018; Johnston et al. 2016; Overstall and McGree 2019; Warne et al. 2017). One final avenue for further exploration would be to examine how the concepts and techniques developed here for the classical logistic growth model would apply to other stochastic models in the mathematical biology and mathematical ecology literature, such as stochastic analogues of the well known SIR and SIS disease transmission models [1]. We leave such analysis for future consideration.

References

- [1] LJS Allen. A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. *Infectious Disease Modelling*. 2 (2017) 128-142.
- [2] RE Baker, CA Yates, R Erban. From microscopic to macroscopic descriptions of cell migration on growing domains. *Bulletin of Mathematical Biology*. 72 (2010) 719-762.
- [3] RE Baker, MJ Simpson. Correcting mean-field approximations for birth-death-movement processes. *Physical Review E*. 82 (2010) 041905.
- [4] B Bolker, SW Pacala. Using moment equations to understand stochastically driven spatial pattern formation in ecological systems. *Theoretical Population Biology*. 52 (1997) 179–197.
- [5] AP Browning, SW McCue, MJ Simpson. A Bayesian computational approach to explore the optimal duration of a cell proliferation assay. *Bulletin of Mathematical Biology*. 79 (2017) 1888–1906.
- [6] AP Browning, SW McCue, RN Binny, MJ Plank, ET Shah, MJ Simpson. Inferring parameters for a lattice-free model of cell migration and proliferation using experimental data. *Journal of Theoretical Biology*. 437 (2018) 251–260.
- [7] M Bruna, SJ Chapman. Diffusion of multiple species with excluded-volume effects. *The Journal of Chemical Physics*. 137 (2012) 204116.
- [8] M Bruna, SJ Chapman. Diffusion of finite-size particles in confined geometries. *Bulletin of Mathematical Biology*. 76 (2014) 947–982.
- [9] AQ Cai, KA Landman, BD Hughes. Multi-scale modeling of a wound-healing cell migration assay. *Journal of Theoretical Biology*. 245 (2007) 576–594.
- [10] E Codling, MJ Plank, S Benhamou. Random walk models in biology. *Journal of the Royal Society Interface*. 5 (2008) 813–834.
- [11] M Dehideniya, CC Drovandi, JM McGree. Optimal Bayesian design for discriminating between models with intractable likelihoods in epidemiology. *Computational Statistics & Data Analysis*. 124 (2018) 277–297.

- [12] L Dyson, PK Maini, RE Baker. Macroscopic limits of individual-based models for motile cell populations with volume exclusion. *Physical Review E*. 86 (2012) 031903.
- [13] L Edelstein-Keshet. *Mathematical Models in Biology*, 2nd Edition (2005) SIAM.
- [14] GB Ermentrout, L Edelstein-Keshet. Cellular automata approaches to biological modelling. *Journal of Theoretical Biology*. 160 (1993) 97-133.
- [15] MB Flegg, SJ Chapman, R Erban. The two-regime method for optimizing stochastic reaction–diffusion simulations. *Journal of the Royal Society Interface*. 9 (2012) 859–868.
- [16] F Frascoli, BD Hughes, MH Zaman, KA Landman. A computational model for collective cellular motion in three dimensions: general framework and case study for cell pair dynamics. *PLOS One*. 8 (2013) e59249.
- [17] F Frascoli, PS Kim, BD Hughes, KA Landman. A dynamical model for tumour immunotherapy. *Mathematical Biosciences*. 253 (2014) 50–62.
- [18] A Gelman, JB Carlin, HS Stern, DB Rubin, DB. *Bayesian Data Analysis*, 2nd ed (2003) CRC Press.
- [19] SAH Geritz, E Kisdi. Mathematical ecology: why mechanistic models? *Journal of Mathematical Biology*. 65 (2012) 1411–1415.
- [20] P Gerlee. The model muddle: in search of tumor growth laws. *Cancer Research*. 73 (2013) 2407–2411.
- [21] DT Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry* 81 (1977) 2340–2361.
- [22] A Goriely. *The Mathematics and Mechanics of Biological Growth*. Springer, New York.
- [23] W Jin, ET Shah, CJ Penington, SW McCue, LK Chopin, MJ Simpson. Reproducibility of scratch assays is affected by the initial degree of confluence: experiments, modelling and model selection. *Journal of Theoretical Biology*. 390 (2016) 136–145.
- [24] W Jin, ET Shah, CJ Penington, SW McCue, PK Maini, MJ Simpson. Logistic proliferation of cells in scratch assays is delayed. *Bulletin of Mathematical Biology*. 79 (2017) 1028–1050.

- [25] JB Johnson, KS Omland. Model selection in ecology and evolution. *Trends in Ecology & Evolution*. 19 (2004) 101–108.
- [26] ST Johnston, JV Ross, BJ Binder, DLS McElwain, P Haridas, MJ Simpson. Quantifying the effect of experimental design choices for *in vitro* scratch assays. *Journal of Theoretical Biology*. 40 (2016), 19–31.
- [27] MJ Keeling, KTD Eames. Networks and epidemic models. *Journal of the Royal Society Interface*. 2 (2005) 295–307.
- [28] R Law, DJ Murrell, U Dieckman. Population growth in space and time: spatial logistic equations. *Ecology* 84 (2003) 252–262.
- [29] B Lambert, AL MacLean, AG Fletcher, AN Combes, MH Little, HM Byrne. Bayesian inference of agent-based models: A tool for studying kidney branching morphogenesis. *Journal of Mathematical Biology*. 76 (2018) 1673–1697
- [30] TM Liggett. *Stochastic interacting systems: contact, voter and exclusion processes*. (1999) Springer, Berlin.
- [31] OJ Maclaren, HM Byrne, AG Fletcher, PK Maini. Models, measurement and inference in epithelial tissue dynamics. *Mathematical Biosciences & Engineering*. 12 (2015) 1321–1340.
- [32] OJ Maclaren, A Parker, C Pin, SR Carding, AJM Watson, AG Fletcher, HM Byrne, PK Maini. A hierarchical Bayesian model for understanding the spatiotemporal dynamics of the intestinal epithelium. *PLoS Computational Biology*. 13 (2017) e1005688.
- [33] Mathworks, 2019a. Gridded and scattered data interpolation, data gridding, piecewise polynomials. <https://au.mathworks.com/help/matlab/interpolation.html>. Accessed: January, 2019.
- [34] Mathworks, 2019b. Kernel smoothing function estimate for univariate and bivariate data. <http://www.mathworks.com/help/stats/ksdensity.html>. Accessed: January, 2019.

- [35] PK Maini, DLS McElwain, DI Leavesley. Traveling wave model to interpret a wound-healing cell migration assay for human peritoneal mesothelial cells. *Tissue Engineering*. 10 (2004) 475–482.
- [36] PK Maini, DLS McElwain, D Leavesley. Travelling waves in a wound healing assay. *Applied Mathematics Letters*. 17 (2004) 575–580.
- [37] V Melica, S Invernizzi, G Caristi. Logistic density-dependent growth of an *Aurelia aurita polyps* population. *Ecological Modelling*. 291 (2014) 1–5.
- [38] JD Murray. *Mathematical Biology. I: An Introduction*, 3rd Edition (2002) Springer, Heidelberg.
- [39] RD O’Dea, JR King. Continuum limits of pattern formation in hexagonal-cell monolayers. *Journal of Mathematical Biology*. 64 (2012) 579–610.
- [40] AM Overstall, JM McGree. Bayesian design of experiments for intractable likelihood models using coupled auxiliary models and multivariate emulation. *Bayesian Analysis* (2019) In press.
- [41] V Pozzobon, P Perré. Han’s model parameters for microalgae grown under intermittent illumination: Determined using particle swarm optimization. *Journal of Theoretical Biology*. 437 (2018) 29–35.
- [42] EA Sarapata, LG de Pillis. A comparison and catalog of intrinsic tumor growth models. *Bulletin of Mathematical Biology*. 76 (2014) 2010–2024.
- [43] BG Sengers, CP Please, RO Oreffo. Experimental characterization and computational modelling of two-dimensional cell spreading for skeletal regeneration. *Journal of the Royal Society Interface*. 4 (2007) 1107–1117.
- [44] JA Sherratt, JD Murray. Models of epidermal wound healing. *Proceedings of the Royal Society Series B*. 241 (1990) 29–36.
- [45] MJ Simpson, KA Landman, BD Hughes. Cell invasion with proliferation mechanisms motivated by time-lapse data. *Physica A: Statistical Mechanics and its Applications*. 389 (2010) 3779–3790.

- [46] KK Treloar, MJ Simpson, DLS McElwain, RE Baker. Are *in vitro* estimates of cell diffusivity and cell proliferation rate sensitive to assay geometry? *Journal of Theoretical Biology*. 356 (2014) 71–84.
- [47] A Tremel, A Cai, N Tirtaatmadja, BD Hughes, GW Stevens, KA Landman, AJ O'Connor. Cell migration and proliferation during monolayer formation and wound healing. *Chemical Engineering Science*. 64 (2009) 247–253.
- [48] A Tsoularis, J Wallace. Analysis of logistic growth models. *Mathematical Biosciences*. 179 (2002) 21–55.
- [49] BN Vo, CC Drovandi, AN Pettitt, MJ Simpson. Quantifying uncertainty in parameter estimates for stochastic models of collective cell spreading using approximate Bayesian computation. *Mathematical Biosciences*. 263 (2015) 133–142.
- [50] DJ Warne, RE Baker, MJ Simpson. Optimal quantification of contact inhibition in cell populations. *Biophysical Journal*. 113 (2017) 1920–1924.
- [51] DJ Warne, RE Baker, MJ Simpson. Using experimental data and information criteria to guide model selection for reaction-diffusion problems in mathematical biology. To appear, *Bulletin of Mathematical Biology* (2019a). Preprint at <https://doi.org/10.1101/444679>
- [52] DJ Warne, RE Baker, MJ Simpson. Simulation and inference algorithms for stochastic biochemical reaction networks: from basic concepts to state-of-the-art. To appear, *Journal of the Royal Society Interface* (2019b). Preprint at <https://arxiv.org/abs/1812.05759>
- [53] GB West, JH Brown, BJ Enquist. A general model for ontogenetic growth. *Nature*. 413 (2001) 628–631.