Bayesian Inference of Deterministic Population Growth Models

Luiz Max F. de Carvalho*, Claudio J. Struchiner and Leonardo S. Bastos

Abstract Deterministic mathematical models play an important role in our understanding of population growth dynamics. In particular, the effect of temperature on the growth of disease-carrying insects. In this paper we propose a modified Verhulst – logistic growth – equation with temperature-dependent parameters. Namely, the growth rate r and the carrying capacity K are given by thermodynamic functions of temperature T, r(T) and K(T). Our main concern is with the problem of learning about unknown parameters of these deterministic functions from observations of population time series P(t,T). We propose a strategy to estimate parameters of r(T) and K(T) by treating the model output P(t,T) as a realization of a Gaussian process (GP) with fixed variance and mean function given by the analytic solution to the modified Verhulst equation. We then use Gibbs sampling, implemented using the recently developed rstan package of the R statistical computing environment, to approximate the posterior distribution of the parameters of interest. In order to evaluate the performance of our algorithm, we run a Monte Carlo study on a simulated example, where bias and coverage were calculated. We then proceed to apply the approach described herein to laboratory data of the Chagas disease vector, Rhodnius prolixus. Analysis of this data shows that the growth rate for the insect population under study achieves its maximum around 26 °C and the carrying capacity is maximum around 25 °C, suggesting that *R. prolixus* populations may thrive even in temperate climates.

Luiz Max F. de Carvalho
Program for Scientific Computing (PROCC) – Oswaldo Cruz Foundation, Rio de Janeiro – RJ – Brazil, e-mail: lmax.procc@gmail.com – *Expected speaker

Claudio J. Struchiner
Program for Scientific Computing (PROCC) – Oswaldo Cruz Foundation, Rio de Janeiro – RJ – Brazil, e-mail: stru@fiocruz.br

Leonardo S. Bastos
Program for Scientific Computing (PROCC) – Oswaldo Cruz Foundation, Rio de Janeiro – RJ – Brazil, e-mail: lsbastos@fiocruz.br

1 Introduction

Deterministic models of population growth have played a major role in our understanding of biological populations [1, 2, 3, 4].

These mechanistic models allow us to draw a pittoresque picture of the real world and capture the main features of the system(s) under study.

In many applications it is of interest to estimate parameters of these models using observed data. This document is about a Bayesian approach to this problem in the context of models for insect population growth. In what follows, we introduce the necessary notation and concepts to be used hereafter.

Consider a deterministic model $M(\cdot)$.

Let $y \in \mathscr{Y} \subset \mathbb{R}^n$ be the set of model inputs and $x \in \mathscr{X} \subset \mathbb{R}^p$ be the model outputs. The deterministic model $M(x;\theta) = y$, where $\theta \in \Sigma \subset \mathbb{R}^q$ is a q-dimensional parameter vector, completely specifies the relationship between x and y.

Here we are concerned with the problem of, having observed x and y, draw inference about θ. From a Bayesian perspective, we are concerned with obtaining the posterior distribution [3]:

$$p(\theta|x,y) \propto p(y,x|\theta)\pi(\theta) \tag{1}$$

$$\propto p(y|x,\theta)\pi(x|\theta)\pi(\theta) \tag{2}$$

$$\propto p(y|x,\theta)\pi(x)\pi(\theta) \tag{3}$$

where (3) follows from the assumption of a priori independence of the inputs and parameters.

We discuss several aspects of the uncertainty in such models and illustrate with a temperature-dependent Verhulst model, proposed in Zimmerman et al (2014) [5].

This paper is organized as follows. In Section 1 we outline the necessary theory and notation. Section 2 details the model, likelihood, priors and posteriors. A simulation study is presented in Section 2.4 and a discussion with our closing remarks is given in Section 3.

2 Logistic Growth with temperature-dependent parameters

Global temperature change may be an important factor on infectious diseases emergence [6]. Arthropod-borne diseases are particularly influenced by climate change, their vectors are very sensitive to temperature variation. In this example we explore the population growth of the Chagas disease vector *Rhodnius prolixus*.

We introduce a modified logistic growth equation, also known as the Verhulst equation [5, 7]. The ordinary non-linear differential equation

$$\frac{dP}{dt} = r\left(1 - \frac{P}{K}\right)P \tag{4}$$

takes two parameters, the growth rate r and the carrying capacity K. For a given initial population condition N_0, an analytic solution is available for the number $P(t)$ of individuals at time t:

$$P(t) = \frac{K}{1 + \left(\frac{K - N_0}{N_0}\right)e^{-rt}} \tag{5}$$

In order to incorporate temperature-dependent behavior, we introduce temperature-dependent parameters, $r(T)$ and $K(T)$. The analytic solution in (5) is slightly modified to yield $P(t,T)$ for time t and temperature T:

$$P(t,T) = \frac{K(T)}{1 + \left(\frac{K(T) - N_0}{N_0}\right)e^{-r(T)t}} \tag{6}$$

To complete the model we must specify $K(T)$ and $r(T)$ as smooth functions of T. We model $K(T)$ and $r(T)$ as Gaussian kernels over T:

$$K(T) = b_K exp\left(-\frac{(T - a_K)^2}{b_K}\right) \tag{7}$$

$$r(T) = b_r exp\left(-\frac{(T - a_r)^2}{b_r}\right) \tag{8}$$

2.1 Likelihood

Assume $P(t,T)$ to be a Gaussian process with fixed variance τ^2. Recalling the notation in section 1, we have $x = \{T, t, N_0\}$, $y = \{P(t,T)\}$ and $\theta = \{a_K, b_K, c_K, a_r, b_r, c_r, \tau^2\}$. Additionally, note that the parameter vector θ can be split into the disjoint sets $\theta_K = \{a_K, b_K, c_K\}$ and $\theta_r = \{a_r, b_r, c_r\}$, which parametrize $K(T)$ and $r(T)$, respectively.

Let $\mathbf{y} = \{y_1, y_2, \ldots, y_N\}$ be an output vector with N measurements, which we observe directly. Moreover, let $\mathbf{t} = \{t_1, t_2, ..., t_N\}$ be the vector which contains the observed times of the observations and $\mathbf{T}$ the analogous vector for temperatures. We then specify

$$y_i | t_i, T_i, N_0, \theta \sim \mathcal{N}(\mu(t_i, T_i, N_0; \theta), \tau^2) \tag{9}$$

$$\mu(t_i, T_i, \theta) = \frac{K(T_i; \theta_K)}{1 + \left(\frac{K(T_i; \theta_K) - N_0}{N_0}\right)e^{-r(T_i; \theta_r)t_i}}, \quad \forall i = 1, 2, \ldots, N \tag{10}$$

which is equivalent to writing $y_i = M(t_i, T_i, N_0; \theta) + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \tau^2)$.

2.2 Priors

For $i = 1,2,..,K$, let θ_i denote each of the $K = 7$ model parameters. Assuming a priori independence, we have

$$\pi(\theta) = \prod_{i=1}^{K} \pi(\theta_i) \tag{11}$$

We then specify prior distributions for the θ as follows[1]:

$$\begin{aligned} a_K, a_r &\sim Normal(20,10) && (12)\\ b_K, b_r &\sim Gamma(4,5) && (13)\\ b_K &\sim Gamma(1,1000) && (14)\\ b_r &\sim Normal(1/2,2) && (15)\\ \tau^2 &\sim Gamma(1/10,10) && (16)\\ & && (17) \end{aligned}$$

These priors were formulated to reflect both model restriction (e.g., positive definitess) and biological knowledge about model parameters.

The a_K and a_r parameters mimic the mean parameter in a Gaussian distribution, and controls where the functions will achieve their maximum. We assume a normal distribution with moderate variance to provide a relatively uninformative accounting of placement in the positive side of the real line.

For b_K and b_r it is also necessary to ensure positive-definitess, which we achieve using Gamma priors. Our prior for b allows very "streched" forms of $K(T)$ and $r(T)$ (see Figure 2.5), while ensuring concavity. Finally, c_K and c_r control curve height for $K(T)$ and $r(T)$ respectively.

The choice for c_K is essentially arbitrary, since little is known about natural populations of *Rhodnius prolixus* in terms of carrying capacity. We thus parameterize $\pi(b_K)$ to reflect rough projections taking into account the number of laid eggs. Since r in equation 4 can theoretically assume negative values, we used a prior for c_r that allows negative values with very low probability.

Now, let θ^r and θ^K be the partitions of the parameter vector, which in combination with the deterministic forms given in (7) and (8), these priors induce prior distributions on r and K for each fixed temperature T, $\pi^*(r) = r(T;\pi(\theta^r))$ and $\pi^*(K) = K(T;\pi(\theta^K))$ respectively. These prior distributions can be easily approximated using Monte Carlo sampling and are presented in Figure 2.5 (dashed lines).

[1] Please note that throughout this text we assume Gamma distributions are parameterised in terms of shape and scale – as opposed to rate – and Normal distributions are parameterised in terms of mean and standard deviation.

2.3 Posterior

Hence, from (3) and assuming a priori independence of **t** and **T**

$$p(\theta|y,t,T) \propto p(y|\theta,t,T)\pi(t)\pi(T)\pi(\theta) \tag{18}$$

is the desired posterior.

We then used the **stan** [8] package of the R Statistical Computing Environment [9] to approximate (18) through Gibbs sampling. Code implementing this posterior approximation is available from the authors upon request.

The MCMC was run for $50,000$ iterations until convergence which was assessed by inspecting the trace- and autocorrelation plots and potential scale reduction factor. After discarding the burn-in we aproximate the posterior distribution of $P(t,T)$ using Monte Carlo sampling as follows. Let Q be the number of samples.

1. Construct a grid of values for t and T;
2. For each $q = 1,2,...,Q$ draw a vector $\theta^{(q)} = \{a_K^{(q)}, b_K^{(q)}, c_K^{(q)}, a_r^{(q)}, b_r^{(q)}, c_r^{(q)}\}$ from the posterior distribution of the parameters;
3. Evaluate $M(t,T,N_0;\theta^{(q)})$ to get $P(t,T)^{(q)}$

From these Q samples, we can compute several quantities of interest, for instance the posterior mean of the population for a particular temperature at a given time. Let t_j and T_j represent a particular pair of temperature and time. Then the posterior mean of $P(t_j,T_j)$ is

$$\mathbb{E}\left[P(t_j,T_j)\right] = \frac{1}{Q}\sum_{q=1}^{Q} P(t_j,T_j)^{(q)} = \frac{1}{Q}\sum_{q=1}^{Q} M(t_j,T_j,N_0;\theta^{(q)}) \tag{19}$$

The posterior median and quantiles can be obtained in a similar fashion.

2.4 Simulation Study

We conducted a Monte Carlo simulation study in order to evaluate the approach proposed here. Simulation was carried out as follows: for each m in a total of M simulation steps,

1. Fix θ, τ and N_0;
2. Construct a grid of values for t and T;
3. For each point in the grid, sample from a normal distribution with mean $M(T,t,N_0;\theta)$ and variance τ, generating a data set $P^{(m)}$;
4. Using $50,000$ iterations of Gibbs sampling MCMC, obtain an estimate $\hat{\theta}^{(m)}$ of model parameters.

In this paper, we use both the *a posteriori* mean and median as point estimates for θ. With these results at hand, we are able to compute the normalized squared bias and mean squared error (MSE) for each parameter, as well as nominal coverage for the 95% credibility intervals. The normalized squared bias for each parameter is defined as

$$B(\theta_i) = \theta_i^{-1}\mathbb{E}\left[\hat{\theta}_i - \theta_i\right]^2 \tag{20}$$

Let $Z_i^{(m)}$ be the indicator variable that assumes 1 if the mth 95% credibility interval contains the true value of θ_i and zero otherwise. Coverage is defined as

$$C(\theta_i) = \mathbb{E}\left[Z_i^{(m)}\right] = \frac{1}{M}\sum_{j=1}^{M} Z_{ij} \tag{21}$$

Table 1 shows the bias, MSE and coverage for each parameter using the posterior mean as a point estimate. Parameter values were chosen so as to reflect a biologically sound behavior for $P(t,T)$.

Table 1 Bias assessment using the simulated data set – posterior mean. 1 – 'True' value used for simulation. 2 – Bias divided by the true parameter value. 3 – Coverage of the 95% credibility intervals.

Parameter	Value[1]	Posterior Mean	Bias[2]	MSE	Coverage[3]
a_K	30.00	29.44	0.01	7.99	0.93
a_r	23.00	22.71	0.00	3.31	0.86
b_K	10.00	13.08	0.95	450.26	0.94
b_r	15.00	16.82	0.22	16.77	0.88
c_K	700.00	692.17	0.09	7203.06	0.96
c_r	0.40	0.43	0.00	0.04	0.85
τ	3.16	4.89	0.94	67.02	0.88

The results for the posterior median as point estimator are presented at Table 2 and are largely in agreement with those for the posterior mean.

Table 2 Bias assessment using the simulated data set – posterior median. 1 – 'True' value used for simulation. 2 – Bias divided by the true parameter value. 3 – Coverage of the 95% credibility intervals.

Parameter	Value[1]	Posterior Median	Bias[2]	MSE	Coverage[3]
a_K	30.00	29.46	0.01	7.99	0.93
a_r	23.00	22.74	0.00	3.31	0.86
b_K	10.00	13.08	0.95	450.26	0.94
b_r	15.00	16.63	0.18	16.77	0.88
c_K	700.00	689.38	0.16	7203.06	0.96
c_r	0.40	0.42	0.00	0.04	0.85
τ	3.16	4.76	0.81	67.02	0.88

In Figure 2.4 we present prior and posterior distributions for model parameters used in this simulation study where we can notice the posteriors are substantially less diffuse than the priors.

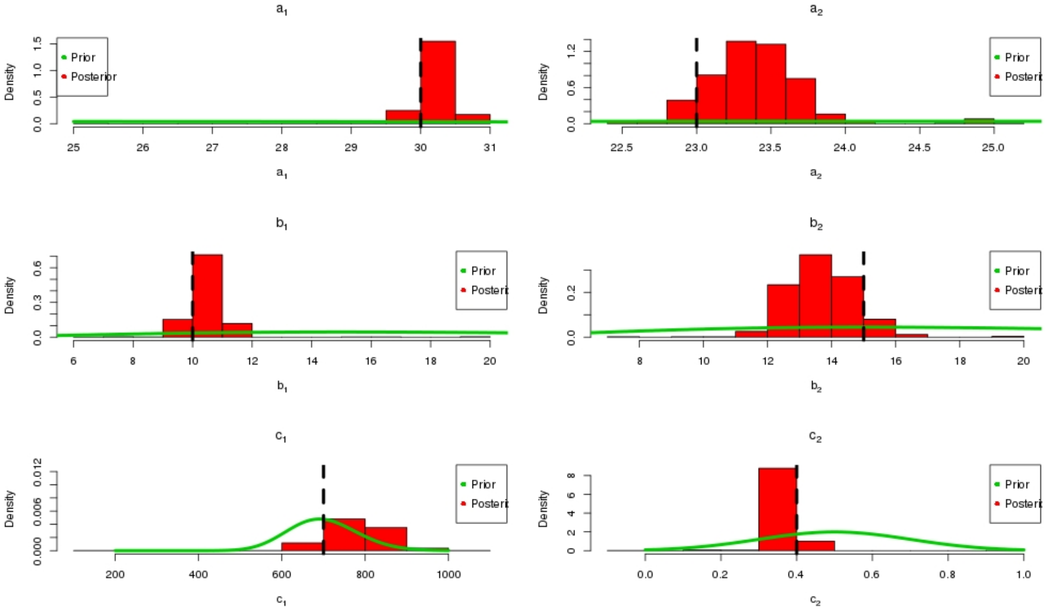


Fig. 1 Prior and posterior distributions of parameter for the simulated data. Dashed vertical lines mark true parameter values.

2.5 *Rhodnius prolixus data*

We now turn our attention to a real-world data set. Chagas disease is an important tropical disease, transmitted by a blood sucking bug, *Rhodnius prolixus*. Temperature is key factor for both insect development and vector competence. We are thus interested in drawing inference on model parameters to understand the role of temperature in the insect's population dynamics.

In a laboratory experiment, $N_0 = 30$ females were observed in several temperatures, and the number of eggs laid was recorded for 35 days. We take the cumulative number of ecloded eggs for each temperature condition as a good approximation of $P(t,T)$, since all conditions (light, water, food, etc.) were optimal. The data thus consisted of $N = 350$ observations, from 10 temperature conditions for 35 days each.

In Figure 2.5 we show prior and posterior distributions for each parameter using this data. Results are similar to that of the simulated data, with substantially concentrated posteriors in comparison to the priors. Interestingly, while the prior expectation for b_K was 25, we obtained a posterior mean around 106 (Table 3), indicating that the variability of *Rhodnius* to temperature is much greater than we expected. See more on this at Section 3.

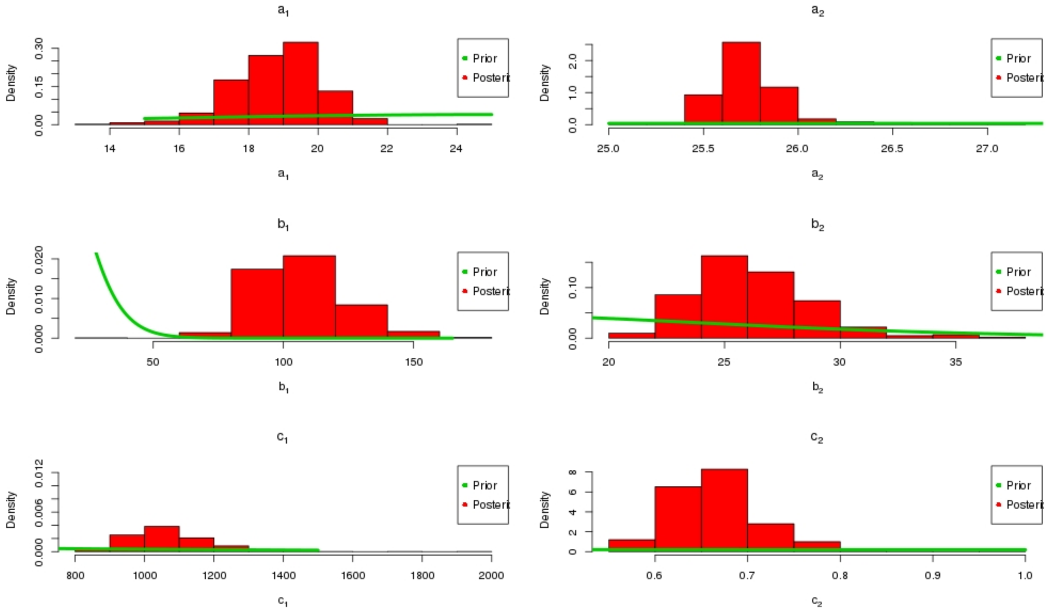
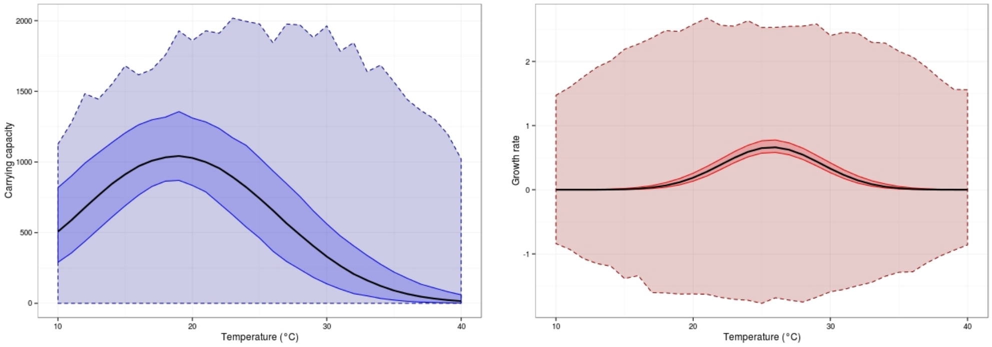


Fig. 2 Prior vs posterior distributions of parameter for the real data set.

Figure 2.5 shows the prior and posterior distribution of the thermodynamic functions. These are easily obtained using Monte Carlo sampling from the prior and posterior distributions of θ, respectively.

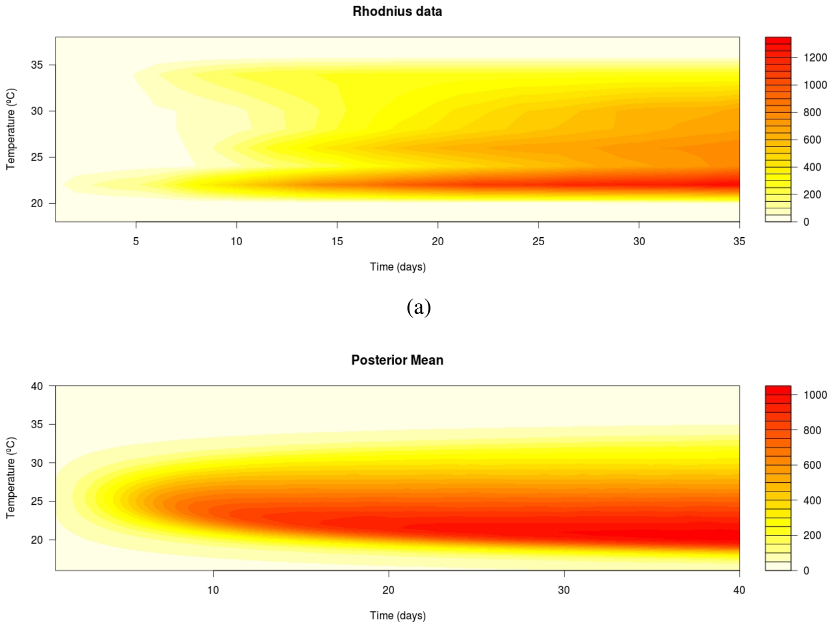


(a)

(b)

Fig. 3 Prior and posterior credibility intervals for the thermodynamic functions under for the real data set. We show $K(T)$ in (a) and $r(T)$ in (b). Dashed lines and lighter tones depict prior and solid lines and darker tones represent the posterior 95% credibility intervals. Thick solid lines mark the medians.

Finally, we present the posterior distribution for $P(t,T)$, obtained using the approach described in section 2.4. We also present the heat map of the original laboratory data described above. It can be noted that the posterior distribution allows a region of optimality for populational growth around $20-25\,^{\circ}$C. Each thermodynamic function, $r(T)$ and $K(T)$ can have its own point of maximum, however.



(a)

(b)

Fig. 4 Heat maps showing population through time and temperature. In (a) we show the laboratory data collected for *Rhodnius* and in (b) the posterior mean, obtained using Monte Carlo sampling from the posterior distribution of the parameters (Section 2.4).

In Table 3 we provide posterior estimates obtained for the real data along with prior expectations and 95% credibility intervals.

Table 3 Posterior inference results for the real set. We report posterior and prior expectations, along with the appropriate 95% credibility intervals. Five independent chains were run for $50,000$ iterations each with the first $25,000$ discarded as burn-in. Convergence was assessed using trace plots and the potential scale reduction factor.

	Posterior Mean (95% C.I.)	Prior Mean (95% C.I.)
a_K	19.23 (17.56 – 21.09)	25.00 (5.40 – 44.60)
a_r	25.73 (25.44 – 26.10)	25.00 (5.40 – 44.60)
b_K	106.17 (75.25 – 137.31)	20.00 (5.44 – 43.84)
b_r	26.77 (22.59 – 32.19)	20.00 (5.44 – 43.84)
c_K	1023.32 (898.28 – 1165.40)	1000.00 (25.31 – 3688.87)
c_r	0.66 (0.58 – 0.76)	0.50 (-3.41 – 4.41)
τ	177.33 (166.10 – 191.78)	1.00 (0.00 – 9.78)

3 Discussion

Deterministic, differential-equation-based models are an important tool in Theoretical Biology, allowing us to study the behavior of biological systems using simple and easily interpretable equations. In this paper we adapt a classical population growth model, the Verhulst logistic equation, initially proposed in 1838, to accomodate temperature-dependent parameters. We then proceed to develop a Bayesian approach to learn about model parameters when population time series are available.

Our approach is based on the idea that both the growth rate r and the carrying capacity K are smooth functions of temperature, which we model as Gaussian kernels of the form presented in (7). This parameterisation is very flexible, allowing us to model a variety of temperature-response patterns. It is biologically motivated, since the response of the insect populations to temperature change should have a maximum point and substantially decrease at extremely high and extremely low temperatures [5].

The main idea is to model population through time as a Gaussian process with a deterministic mean function $M(\cdot)$ which is given by the solution to the differential equation (Eq. 6). The posteriors outlined in (3) and (18) allow for the incorporation of uncertainty regarding model inputs. Although this source of uncertainty is negligible in our setting due to carefully controlled experimental conditions, it could play an important role in studies dealing with field data.

It should also be noted that in this paper we assume independence between model inputs. This assumption may be irrealistic, since temperature is likely to depend on time in general. In our particular conditions, all experiments were performed at controlled environments, where temperature was kept constant throughout time. In the case of population time series obtained from field data, this assumption is likely not to hold.

From Tables 1 and 2 it can be noticed that our approach presents good frequentist properties, with high coverage probabilities of the credibility intervals for most parameters. The exception is the Gaussian process variance, τ^2 for which we could not recover the true value with good accuracy, albeit achieving good coverage. This result most likely stems from the restricting fixed variance assumption. Replacing the fixed variance by a smooth function of time $\tau(t)$ could greatly improve model fit and is an important future direction of research.

In conclusion, in this paper we provide insight on how to perform inference on the parameters of a complex, non-linear deterministic model of population growth when population time series are available. These parameters are directly interpretable and provide important information about the underlying biological dynamics. The framework proposed here can be adapted to a broad class of models, for example to learn about the parameters of epidemiological models using data on disease incidence. Moreover, the Bayesian approach allows for a complete treatment of uncertainty on model inputs and outputs.

References

1. W. G. Costello and H. M. Taylor, “Deterministic population growth models,” *The American Mathematical Monthly*, vol. 78, no. 8, pp. 841–855, 1971.
2. R. M. May *et al.*, “Simple mathematical models with very complicated dynamics,” *Nature*, vol. 261, no. 5560, pp. 459–467, 1976.
3. D. Poole and A. E. Raftery, “Inference for deterministic simulation models: the bayesian melding approach,” *Journal of the American Statistical Association*, vol. 95, no. 452, pp. 1244–1255, 2000.
4. C. S. Gillespie and A. Golightly, “Bayesian inference for generalized stochastic population growth models with application to aphids,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 59, no. 2, pp. 341–357, 2010.
5. L. T. Zimmermann, L. M. F. Carvalho, L. R. Vasconcellos, B. L. S., C. J. Struchiner, and A. H. Lopes, “Temperature-dependent oviposition and egg eclosion of chagas disease vector *Rhodnius prolixus*,” *Submitted*, 2014.
6. J. A. Patz, P. R. Epstein, T. A. Burke, and J. M. Balbus, “Global climate change and emerging infectious diseases,” *JAMA: the journal of the American Medical Association*, vol. 275, no. 3, pp. 217–223, 1996.
7. P.-F. Verhulst, “Notice sur la loi que la population suit dans son accroissement. correspondance mathématique et physique publiée par a,” *Quetelet*, vol. 10, pp. 113–121, 1838.
8. Stan Development Team, “Stan: A c++ library for probability and sampling, version 2.2,” 2014.
9. R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.