Running Head:

TAKING ERROR INTO ACCOUNT WITH ABC

Title:

Taking Error Into Account When Fitting Models Using Approximate Bayesian Computation

Authors:

Elske van der Vaart[a,b*], vdrvaart@uvt.nl, +31 13 466 31 60

Dennis Prangle[c], dennis.prangle@newcastle.ac.uk

Richard M. Sibly[a], r.m.sibly@reading.ac.uk

[a]School of Biological Sciences, University of Reading, Harborne Building, University of Reading, Whiteknights, Reading, Berkshire, RG6 6AS, United Kingdom

[b]Cognitive Science and Artificial Intelligence, Tilburg University, School of Humanities, PO Box 90153, 5000 LE Tilburg, the Netherlands

[c]School of Mathematics and Statistics, Newcastle University, Herschel Building, Newcastle University, Newcastle upon Tyne, NE1 7RY, United Kingdom

Abstract (maximum 200 words)

Stochastic computer simulations are often the only practical way of answering questions relating to ecological management. However, due to their complexity, such models are difficult to calibrate and evaluate. Approximate Bayesian Computation (ABC) offers an increasingly popular approach to this problem, widely applied across a variety of fields. However, ensuring

the accuracy of ABC's estimates has been difficult. Here, we obtain more accurate estimates by incorporating estimation of error into the ABC protocol. We show how this can be done where the data consist of repeated measures of the same quantity and errors may be assumed to be normally distributed and independent. We then derive the correct acceptance probabilities for a probabilistic ABC algorithm, and update the 'coverage test' with which accuracy is assessed. We apply this method – which we call 'error-calibrated ABC' – to a toy example and a realistic 14-parameter simulation model of earthworms that is used in environmental risk assessment. A comparison with exact methods and the diagnostic 'coverage test' show that our approach improves estimation of parameter values and their credible intervals for both models.

Keywords

ABC, IBM, approximate Bayesian computation, individual-based model, parameter estimation

Introduction

Stochastic computer simulations are increasingly used to make realistic predictions about real world ecological processes (Hartig et al. 2011); from the survival of shorebirds (West et al. 2002) to the effects of climate change (Zurell et al. 2012) and the invasiveness of plants (Nehrbass and Winkler 2007). Because such models attempt to simulate all relevant aspects of a real physical system, they often involve many parameters, some of which will be difficult to set correctly. Understanding the overall uncertainty introduced by these unknown parameter values is crucial, especially when the final objective of these models is to assess the possible consequences of management decisions, such as the translocation of vulnerable species (Lethbridge and Strauss 2015) or the placement of wind turbines (Nabe-Nielsen et al. 2014).

Approximate Bayesian Computation, or ABC, is a promising technique for estimating parameter values together with their credible intervals. Standard Bayesian methods explore properties of the multivariate posterior distribution over the parameters (Gelman et al. 2013), often by sampling parameter vectors from it. This posterior distribution specifies the degree of support for different parameter vectors given the model, data and prior knowledge about the values the parameters are likely to take. Sampling from the exact posterior is not always feasible, leading to the development of approximate Bayesian methods, such as ABC.

Originally developed within population genetics (Tavaré et al. 1997, Pritchard et al. 1999, Beaumont et al. 2002), ABC is now widely used, with recent applications to, for example, range expansions (Rasmussen and Hamilton 2012), infectious diseases (Kosmala et al. 2016), and forest dynamics (Lagarrigues et al. 2015). However, ensuring the accuracy of ABC's estimates remains difficult. Here, we improve the estimation process for cases where the data consists of repeated measures of the same quantity, such as a time series. We do this using Wilkinson (2013)'s insight that accurate estimates can be obtained if the form of the error – the distribution of the differences between model outputs and data – is incorporated into the ABC protocol.

Bayesian inference generally requires an analytical likelihood, expressing how the likelihood of the data depends on the model parameters, but for mechanistic simulation models, this is often not possible. Instead, ABC is based on simulations using the model. By repeatedly sampling parameters from a model's prior, running the model, and then retaining the simulations closest the data according to some distance function, ABC can approximate a model's posterior with an

accuracy that depends on the distance allowed between model outputs and data. This version of ABC is referred to as 'rejection ABC'. However in many cases even the best-fitting model will not replicate the data exactly – even with the best parameters, there will always be some residual distance between the model and the data, due to either model misspecification, observational measurement error, or both. In these cases, taking error into account can greatly increase posterior accuracy. Accounting for different types of error is well established in deterministic modelling (e.g., Campbell 2006, Higdon et al. 2008, Goldstein and Rougier 2009), but Wilkinson (2013) was the first to consider it in the context of stochastic computer simulations and ABC.

Wilkinson's (2013) method assumes that the data measurements *D* can be considered as a realization of the model η run with its input parameters θ set at their best values, $\hat{\theta}$, plus an independent term ϵ representing error (Equation 1). If the distribution of ϵ is known, Equation 1 determines a probability distribution for *D* given the input value of $\hat{\theta}$. Therefore there is an associated likelihood function. However, for most simulators $\eta(\hat{\theta})$ is extremely complicated, so the likelihood function cannot be expressed as a simple mathematical formula. This means standard Bayesian or maximum likelihood methods cannot be used.

$$D = \eta(\hat{\theta}) + \epsilon \qquad \text{Equation 1}$$

The distribution of ϵ would ideally be based on a priori knowledge, with a principled decomposition into model and measurement error. However for many ecological applications, this is not practical. Any model concerned with the behavior of real organisms will have structural inadequacies that are difficult to formally characterise, and many models are validated against empirical data that was collected long ago, by other researchers, so that measurement

error is also unknown. In this paper, we present a simple approach to using Wilkinson's (2013) method in cases where the empirical data consist of many data points of the same type.

Using the difference between the observations and the model at its best-fitting parameter values, we parameterise a normally distributed estimate of the error, and then derive the corresponding optimal acceptance probabilities for a new 'error-calibrated ABC' algorithm. We illustrate the use of this new algorithm by analysing both a toy example and a complex computer simulation of earthworms (Johnston et al. 2014), which was developed for the purpose of pesticide risk assessment. This model was previously calibrated using 'rejection ABC' (van der Vaart et al. 2015), but a diagnostic 'coverage test' showed some inaccuracies in the posteriors. In this paper, we update this diagnostic so that it also takes error into account, and show that 'error-calibrated ABC' improves the estimation process for both the toy example and the earthworm simulation.

Methods

In previous work (van der Vaart et al. 2015) we implemented the most basic form of ABC, 'rejection ABC', using Algorithm 1. 'Rejection ABC' takes a sample of the parameter values needed to run the model from a prior distribution which expresses existing knowledge about what values each parameter is likely to take. The model is run with those parameter values, and then the process is repeated thousands of times with different sets of parameter values randomly drawn from the prior distribution. 'Rejection ABC' rejects all but the *m* best parameter values, i.e., the *m* values that produce model outputs closest to the data points. These are samples from an approximation to the Bayesian posterior distribution. The exact posterior distribution gives the degree of support for each parameter vector, combining prior information and model observations, and is used to produce univariate posterior distributions for each individual

parameter, as well as 95% credible intervals. The accuracy of the ABC approximation to the posterior can be assessed using 'coverage tests' (Prangle et al. 2013).

1. Repeat *n* times:
    a. Draw $\theta^i \sim \pi(\theta)$ (the prior distribution)
    b. Simulate $X^i \sim \eta(\theta^i)$ (the computer model)
2. Accept the *m* runs (θ^i, X^i) that minimise $\rho(X^i, D)$.

Algorithm 1. Original 'rejection ABC' algorithm used in van der Vaart et al. (2015).

The computer model is represented by $\eta(\theta)$, with output X and input parameters θ. This model is stochastic: repeated evaluations using the same input usually produce different outputs. Though our methods are also valid for deterministic models, better alternatives are available for those cases. X is a vector of model outputs which are to be compared with a data vector D. θ is a vector of model parameters, drawn from a prior distribution, $\pi(\theta)$. We often specify a prior distribution for each individual parameter and form the overall prior by an independence assumption. In total, n model runs are done, and ρ is the distance between the model output X and the data D. The m runs that minimise ρ are accepted and then the accepted (θ^i, X^i) pairs form a sample from an approximate posterior. Since both parameters and outputs are vectors, we use subscripts to denote particular components. For example, θ_j^i represents the jth parameter for model run i, while X_j^i represents the model output corresponding to the jth data point in model run i.

1.1. Coverage

Coverage tests were introduced by Prangle et al. (2013) to check the accuracy of estimated posterior distributions. The idea is to randomly draw a model output X^i from ABC's sample of accepted runs as the 'pseudo-data' X^0 for a new round of ABC. This does not require further

simulation runs, as the original runs can be re-used. The output of this new round of ABC is a set of accepted runs associated with X^0. Then, for each parameter j, we calculate the p_j^0, the proportion of accepted parameter values smaller than that which produced X^0. We then repeat the whole process many times, ending up with a sample of p_j^0 values for each parameter j. Intuitively these should be spread out between 0 and 1, and not 'bunched up' at either the middle or the extremes of the estimated posteriors. Ideally, the p_j^0 values have a Uniform(0,1) distribution (Prangle et al. 2013). Algorithm S1 in Appendix S1 gives the coverage algorithm that we first applied to our earthworm model (van der Vaart et al. 2015); unfortunately this produced non-uniform coverage for several parameters, motivating the work reported here.

1.2. Error-Calibrated ABC

In order to improve our estimation procedure, we used Wilkinson's (2013) version of ABC, which provides inference for the model given by Equation 1. How to choose π_ϵ, the probability density function for the error ϵ, is discussed in the next section. Now the acceptance step (2) of Algorithm 1 is replaced by a probabilistic version, where each (θ^i, X^i) pair is accepted with probability $\frac{\pi_\epsilon(D-X^i)}{c}$, where $\pi_\epsilon(D - X^i)$ is the probability density function of ϵ evaluated at $D - X^i$, and c is a constant chosen as the maximum of $\pi_\epsilon(D - X^i)$ (Wilkinson 2013).

1.3. Error Estimation

If π_ϵ were known it would be straightforward to implement Wilkinson's (2013) algorithm, though perhaps slow to produce adequate sample sizes, but in general π_ϵ is not known. However, if the data come from replicated experiments or time series it is possible to estimate π_ϵ from the differences between the data and the output of the best-fitting model.

To do this, we first find $\hat{X}$, the model output X^i which minimises $\rho(X^i, D)$. When all data are of the same type, $\rho(X^i, D)$ is the sum of all Euclidean distances between X^i and D. When the data are of *k* different types, all Euclidean distances are centered and scaled before summing. For example, in our earthworm model, where some data points concern growth and others concern reproduction, all Euclidean distances are centered and scaled by the mean and standard deviation of all Euclidean distances of that type. This ensures that the overall distance calculation is not dominated by scale differences between the data types.

We then assume that, for each data type, the errors on data points are independent of each other and drawn from a normal distribution with mean *0*, as in classical statistics; this is an assumption that we discuss in our conclusion. To estimate the standard deviation λ of this normal distribution, we take the standard deviation $\hat{\lambda}$ of all the $\hat{X}_j - D_j$ values that are of the same type. So, for example, for the earthworm model, $\hat{\lambda}_{growth}$ is equal to the standard deviation of all differences between the best-fitting model output $\hat{X}$ and the data *D* for all data points concerning growth, and $\hat{\lambda}_{reproduction}$ is equal to the standard deviation of all differences between the best-fitting model output $\hat{X}$ and the data *D* for all data points concerning reproduction.

Then, under our assumption of independent, normally distributed errors, the probability density function $\pi_\epsilon(D - X^i) \propto \prod_{j=1}^{l} \pi_{N(0,1)}\left(\frac{X_j^i - D_j}{\hat{\lambda}_{\tau(j)}}\right)$, where *l* is the number of data points, $\tau(j)$ is the type of the *j*th data point, and $\hat{\lambda}_\tau$ is the standard deviation of data points of type τ. In other words, the overall acceptance probability of a specific model run *i* can be calculated by multiplying the probability densities of each of the simulated data points being produced from

the empirical data, given the assumed error distribution. This density is quicker to compute via a transformation, giving $\pi_\epsilon(D - X^i) \propto \pi_{\chi_l^2}(s)\, s^{1-\frac{l}{2}}$, where $s = \sum_{j=1}^{l}\left(\frac{X_j^i - D_j}{\hat{\lambda}_{\tau(j)}}\right)^2$; i.e., the density of a chi-square distribution with l degrees of freedom evaluated at s, the summed squares of all normalised errors multiplied by a Jacobian term, $s^{1-\frac{l}{2}}$. Algorithm 2 shows the overall procedure, which we call 'error-calibrated ABC'.

1. Repeat *n* times:

- ***a.*** Draw $\theta^i \sim \pi(\theta)$
- ***b.*** Simulate $X^i \sim \eta(\theta^i)$

2. Find $\hat{X}$, the simulated value that minimises $\rho(X^i, D)$.

3. For each data type *k*, calculate $\hat{\lambda}_k$, the standard deviation of all corresponding $\hat{X}_j - D_j$.

4. Accept (θ^i, X^i) with probability $\frac{\pi_{\chi_l^2}(s)\, s^{1-\frac{l}{2}}}{c}$, where $s = \sum_{j=1}^{l}\left(\frac{X_j^i - D_j}{\hat{\lambda}_{\tau(j)}}\right)^2$ and c is equal to the maximum acceptance probability across all runs.

Algorithm 2. New 'error-calibrated ABC' algorithm.

1.4. Error-Calibrated Coverage

Finally, to assess the accuracy of this new algorithm, we update our coverage test, as shown in Algorithm 3, where d = 200, following Prangle et al. (2013). The main change is that the 'pseudo-data' is no longer directly equal to a best-fitting model runs but to a model run plus estimated noise π_ϵ making it more like the empirical data.

1. Add noise to all simulation results X^i, creating new pairs (θ^i, W^i). In

particular, add $\boldsymbol{N}(\mathbf{0}, \hat{\boldsymbol{\lambda}}^{\mathbf{2}}_{\boldsymbol{\tau(j)}})$ noise to X_j^i to get W_j^i, where $\hat{\lambda}_{\tau(j)}$ is the standard deviation of data points of type $\tau(j)$, i.e. of the same type as the jth data point.

2. For each of the d noisy (θ^i, W^i) that minimise $\rho(W^i, D)$:

a. Label as (θ^0, W^0) and do 'error-calibrated ABC' with $D = W^0$, using all remaining non-noisy model runs as the simulations:

i. Accept each (θ^i, X^i) according to its acceptance probability, using the $\hat{\lambda}$ values calculated in the original analysis.

b. For each parameter j:

i. Calculate p_j^0 , the sum of all acceptance probabilities with $\theta_j^i \leq \theta_j^0$ divided by the sum of all acceptance probabilities.

3. Plot the distribution of all p_j^0 values, and check for uniformity.

Algorithm 3. New coverage algorithm for 'error-calibrated ABC'.

1.5. Applications

To test this new 'error-calibrated ABC', we applied it first to a quadratic model where it is possible to calculate exact posteriors, and second to our earthworm simulation. In each case, we compared its results to those of 'rejection ABC', where we deterministically accepted the m runs with the highest acceptance probability according to Algorithm 1. For the quadratic model, the data consist of observations $D = \theta_1 + \theta_2 x + \theta_3 x^2$ plus noise $\epsilon = \boldsymbol{N}(\mathbf{0}, 100)$, evaluated for x values 1, 2, …, 10 with the true $\theta_1 = -2$, $\theta_2 = 1$ and $\theta_3 = 2$. The simulator η has the same form without the error; to estimate the values of the θ_1, θ_2, and θ_3 parameters we took 10^5 samples of $[\theta_1, \theta_2, \theta_3]$ where each of θ_1, θ_2, and θ_3 were drawn from independent $\boldsymbol{N}(\mathbf{0}, 9)$ priors. This means

that, for this simple example, the simulator is deterministic rather than stochastic. Exact posteriors were calculated using Bayesian regression; see Textbox S1.

For the earthworms, the observed data D consist of two types: 122 average body masses and 38 cocoon productions of earthworms living on experimental laboratory diets. In each case, five to ten earthworms were placed in small containers filled with cattle manure for food (Reinecke and Viljoen 1990, Gunadi et al. 2002, Gunadi and Edwards 2003). The model η is an individual-based model (or IBM) that simulates the growth and reproduction of individual earthworms according to established physiological principles (Sibly et al. 2013). Earthworms wriggle around randomly as they forage, and allocate assimilated energy to maintenance, growth, reproduction, and reserves, in a fixed order of priority; see Johnston et al. (2014). In total the model has fourteen parameters θ, given in Table S1 in Appendix S1. The priors for all parameters were lognormal, with means equal to previously determined literature values (see Johnston et al. (2014)) and standard deviations equal to 0.3536. This produces samples where 95% of the values lie between half and twice the literature values on the unlogged scale. We used ARCHER, the UK's national supercomputing service, to do 10^6 runs; see van der Vaart et al. (2015) for details.

1.6. Implementation

All ABC code and the quadratic example were implemented in *R* (R Core Team 2015). The earthworm model was built in NetLogo (Wilensky 1999), and *RNetLogo* was used to run NetLogo from *R* (Thiele et al. 2012). All statistical tests were corrected for multiple testing using Holm's method, and all code and simulation results were deposited in a figshare repository.[1]

[1] Link to be added before publication.

Results

For the quadratic example, 'error-calibrated ABC' estimated the standard deviation of the error, λ, to be 7.91, producing 330 acceptances. Figure 1A shows the model's resulting fit (a 'posterior predictive check'). The posteriors of all three parameters were not significantly different from those obtained by exact Bayesian regression (Figure 1B – D), and coverage plots were uniform (Figure 2A - C), suggesting accurate posteriors. By contrast, for 'rejection ABC' with m = 330 acceptances, all three posteriors were significantly different from those obtained by exact Bayesian regression, and coverage plots were 'U-shaped', with an excess of p values at the extremes (Figure S2). After further varying m from 100 to 1000 to 10000, we found that 'rejection ABC' was only accurate for m = 1000 (Figure S1 & Figure S2); see Figure 2D – F.

For the earthworms, 'error-calibrated ABC' initially accepted only the best-fitting run, which is necessarily accepted (see Algorithm 2). Using this best-fitting run, we verified that the error distributions were normal for both masses and cocoons (Figure S3), and we estimated their standard deviations to be 0.08 and 10.4 respectively. To increase the number of acceptances, we fixed λ_{mass} and $\lambda_{cocoons}$ at their original values, but otherwise reduced the data set to every 6th point; see the Discussion for rationale. Now, 108 runs were accepted, giving the posterior predictive check of Figure 3. Relative to the priors, 4 out of 14 posteriors were significantly narrowed (Figure S4), and coverage was uniform for all 14 (Figure S5).

In comparison, 'rejection ABC' with m = 100 acceptances narrowed five posteriors (Figure S6). For h, the half saturation coefficient, IG_m, the maximum ingestion rate, and M_m, the maximum mass, these posteriors were significantly different from those of 'error-calibrated ABC' (

Figure 4). IG_m and M_m, along with three other parameters, also produced non-uniform coverage, Figure S7. After varying m from 100 to 10^3, 10^4 and 10^5, we found that 'rejection ABC' never produced uniform coverage for all parameters at once (Figure S8), with E, the activation energy, for example, varying from 'U-shaped' at $m = 100$ to 'mountain shaped' at $m = 10^5$.

Discussion

We have shown how incorporating estimation of error into the ABC protocol can improve estimates of parameter values and their credible intervals. To do this we specified ABC acceptance probabilities for the case that errors are normally distributed and independent. Our 'error-calibrated ABC' implements a general methodology introduced by Wilkinson (2013). To diagnose the accuracy of our method, we updated Prangle et al.'s (2013) coverage test by adding the estimated error to the simulation runs used as 'pseudo-data', improving their realism.

For our two example models, 'error-calibrated ABC' appears to have improved posterior accuracy: Coverage plots were uniform for all parameters, and for the quadratic case, results were indistinguishable from those of exact Bayesian regression. In both cases, 'rejection ABC' with an equivalent number of acceptances was demonstrably inaccurate. For the quadratic model, this could be corrected by accepting more runs, but for the earthworm IBM, 'rejection ABC' never produced uniform coverage for all parameters simultaneously. Thus, we conclude that 'error-calibrated ABC' offers a real improvement with respect to model calibration.

In essence, coverage checks for inaccuracies in ABC's posteriors by repeatedly applying the ABC protocol to 'pseudo data' for which the correct parameter values are known. Typically, a lack of uniformity can then be due to either error or inadequacy in the ABC protocol; most

notably, an incorrect acceptance rate. A standard coverage test assumes that the model is perfect, and that calibrated correctly, it can replicate the data exactly. However, our updated coverage test drops this assumption, by adding 'noise' drawn from the error model to all data points before using them as 'pseudo-data'. In a coverage test, surpluses in the tails of the coverage distribution, as in Figure 2D, imply that posteriors are too narrow, with too few runs accepted. At the other extreme, deficits in the tails of the coverage distribution, as in Figure 2F, imply that posteriors are too wide, with too many runs accepted. For this polynomial example, we know the error model is correct, so any lack of uniformity must be due to problems with the acceptance criteria.

For the earthworm model, the use of 'error-calibrated ABC' required two approximations: Firstly, it seems unlikely that its errors really are independent across observations and normally distributed. However, this assumption has often been made by ecologists deploying regression models, and would seem as justifiable here. For the future, it would be interesting to explore methods that incorporate correlations between successive errors, since these could reduce the degrees of freedom and so increase acceptance rates. Currently, as our second approximation, we had to remedy a lack of acceptances by reducing the data set to every 6^{th} data point. As the error distribution π_ϵ is multivariate normal with dimension equal to the number of data points, acceptance falls off exponentially as the number of data points increases. "Too much data" is a common problem in ABC, known as 'the curse of dimensionality'. It is generally addressed by summarizing data sets into as few as one or two 'summary statistics' (see, e.g., Blum et al. 2013). Addressing the issue by 'thinning out' a time series, as here, is not an established technique but has the same fundamental justification. While simple, it appears to work well in this case; visually the accepted runs still mimic the full data set (Figure 3), and for 'rejection

ABC', the posteriors estimated with the full and reduced data sets are similar (Figure S9).

Our overall approach is relatively simple, and does not make use of various sophistications already present in the literature. These include techniques for 'correcting' accepted parameter values on the basis of the resulting model fit, for example using regression (Beaumont et al. 2002), by estimating the error simultaneously with a model's parameters, as in ABCμ (Ratmann et al. 2009), by analysing time series data sequentially (Jasra 2015), or by sampling a model's parameters more efficiently, as in MCMC-ABC (Marjoram et al. 2003) or SMC-ABC (Sisson et al. 2007). We see the simplicity of 'error-calibrated ABC' as an attraction; more efficient sampling schemes are harder to implement and make it impossible to re-use runs for the purpose of calculating coverage. In these cases, 'error-calibrated ABC' offers an accessible approach to improving models' posteriors, with the additional benefit of explicitly accounting for error.

Acknowledgements

This work used the ARCHER UK National Supercomputing Service (http://www.archer.ac.uk) and was supported by NERC grant number NE/K006282/1. DP was supported by a Richard Rado postdoctoral fellowship from the University of Reading during much of this project. We thank A Meade for computational assistance, S Watson and the University of Reading's Bayesian reading group for discussion, and PJ van Leeuwen, R Everitt, M Beaumont, M Kosmala, J Barber and two anonymous referees for very helpful comments on the manuscript.

References

Beaumont, M. A., W. Zhang, and D. J. Balding. 2002. Approximate Bayesian computation in population genetics. Genetics **162**:2025 - 2035.

310 Blum, M. G. B., M. A. Nunes, D. Prangle, and S. A. Sisson. 2013. A comparative review of
311 dimension reduction methods in approximate Bayesian computation. Statistical Science
312 **28**:189-208.
313 Campbell, K. 2006. Statistical calibration of computer simulations. Reliability Engineering &
314 System Safety **91**:1358-1363.
315 Gelman, A., C. J.B., H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2013. Bayesian
316 Data Analysis. 3rd edition. Chapman & Hall/CRC.
317 Goldstein, M., and J. Rougier. 2009. Reified Bayesian modelling and inference for physical
318 systems. Journal of Statistical Planning and Inference **139**:1221-1239.
319 Gunadi, B., C. Blount, and C. A. Edwards. 2002. The growth and fecundity of *Eisenia fetida*
320 (Savigny) in cattle solids pre-composted for different periods. Pedobiologia **46**:15-23.
321 Gunadi, B., and C. A. Edwards. 2003. The effects of multiple applications of different organic
322 wastes on the growth, fecundity and survival of *Eisenia fetida* (Savigny) (Lumbricidae).
323 Pedobiologia **47**:321-329.
324 Hartig, F., J. M. Calabrese, B. Reineking, T. Wiegand, and A. Huth. 2011. Statistical inference
325 for stochastic simulation models - theory and application. Ecology Letters **14**:816-827.
326 Higdon, D., J. Gattiker, B. Williams, and M. Rightley. 2008. Computer model calibration using
327 high-dimensional output. Journal of the American Statistical Association **103**:570-583.
328 Jasra, A. 2015. Approximate Bayesian Computation for a class of time series models.
329 International Statistical Review **83**:405-435.
330 Johnston, A. S. A., M. E. Hodson, P. Thorbek, T. Alvarez, and R. M. Sibly. 2014. An energy
331 budget agent-based model of earthworm populations and its application to study the
332 effects of pesticides. Ecological Modelling **280**:5-17.

Kosmala, M., P. Miller, S. Ferreira, P. Funston, D. Keet, and C. Packer. 2016. Estimating wildlife disease dynamics in complex systems using an Approximate Bayesian Computation framework. Ecological Applications **26**:295-308.

Lagarrigues, G., F. Jabot, V. Lafond, and B. Courbaud. 2015. Approximate Bayesian computation to recalibrate individual-based models with population data: Illustration with a forest simulation model. Ecological Modelling **306**:278-286.

Lethbridge, M. R., and J. C. Strauss. 2015. A novel dispersal algorithm in individual-based, spatially explicit Population Viability Analysis: A new role for genetic measures in model testing? Environmental Modelling & Software **68**:83-97.

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavaré. 2003. Markov chain Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences of the United States of America **100**:15324-15238.

Nabe-Nielsen, J., R. M. Sibly, J. Tougaard, J. Teilmann, and S. Sveegard. 2014. Effects of noise and by-catch on a Danish harbour porpoise population. Ecological Modelling **272**:242-251.

Nehrbass, N., and E. Winkler. 2007. Is the Giant Hogweed still a threat? An individual-based modelling approach for local invasion dynamics of *Heracleum mantegazzianum*. Ecological Modelling **201**:377-384.

Prangle, D., M. G. B. Blum, G. Popovic, and S. A. Sisson. 2013. Diagnostic tools for approximate Bayesian computation using the coverage property. Australian & New Zealand Journal of Statistics **56**:309-329.

354 Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. 1999. Population growth
355 of human Y chromosomes: A study of Y chromosome microsatellites. Molecular Biology
356 and Evolution **16**:1791-1798.
357 R Core Team. 2015. R: A Language and Environment for Statistical Computing. The R
358 Foundation for Statistical Computing, Vienna, Austria.
359 Rasmussen, R., and G. Hamilton. 2012. An approximate Bayesian computation approach for
360 estimating parameters of complex environmental processes in a cellular automata.
361 Environmental Modelling & Software **29**:1-10.
362 Ratmann, O., C. Andrieu, C. Wiuf, and D. M. Richardson. 2009. Model criticism based on
363 likelihood-free inference, with an application to protein network evolution. Proceedings
364 of the National Academy of Sciences **106**:10576-10581.
365 Reinecke, A. J., and S. A. Viljoen. 1990. The influence of feeding patterns on growth and
366 reproduction of the vermicomposting earthworm *Eisenia fetida* (Oligochaeta). Biology
367 and Fertility of Soils **10**:184-187.
368 Sibly, R. M., V. Grimm, B. T. Martin, A. S. A. Johnston, K. Kułakowska, C. J. Topping, P.
369 Calow, J. Nabe-Nielsen, P. Thorbek, and D. L. DeAngelis. 2013. Representing the
370 acquisition and use of energy by individuals in agent-based models of animal
371 populations. Methods in Ecology and Evolution **4**:151-161.
372 Sisson, S. A., Y. Fan, and M. A. Tanaka. 2007. Sequential Monte Carlo without likelihoods.
373 Proceedings of the National Academy of Sciences of the United States of America
374 **104**:1760-1765.
375 Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly. 1997. Inferring coalescence times
376 from DNA sequence data. Genetics **145**:505-518.

Thiele, J. C., W. Kurth, and V. Grimm. 2012. RNetLogo: An R package for running and exploring individual-based models implemented in NetLogo. Methods in Ecology and Evolution **3**:480-483.

van der Vaart, E., M. A. Beaumont, A. Johnston, and R. M. Sibly. 2015. Calibration and evaluation of individual-based models using Approximate Bayesian Computation. Ecological Modelling **312**:182-190.

West, A. D., J. Goss-Custard, R. A. Stillman, R. W. G. Caldow, S. E. A. L. D. Durell, and S. McGrorty. 2002. Predicting the impacts of disturbance on shorebird mortality using a behaviour-based model. Biological Conservation **106**:319-328.

Wilensky, U. 1999. NetLogo. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

Wilkinson, R. D. 2013. Approximate Bayesian Computation (ABC) gives exact results under the assumption of model error. Statistical Applications in Genetics and Molecular Biology **12**:129-141.

Zurell, D., V. Grimm, E. Rossmanith, N. Zbinden, N. E. Zimmerman, and B. Schröder. 2012. Uncertainty in predictions of range dynamics: Black grouse climbing the Swiss Alps. Ecography:590-603.

Figure Legends

Figure 1. Results for the quadratic example. A: Posterior check. Black points represent the data, the result of $\theta_1 + \theta_2 x + \theta_3 x^2$ plus $\boldsymbol{N}(\mathbf{0}, 100)$ noise, and the semi-transparent grey lines are the 'posterior predictive check', i.e., 100 random samples from runs accepted by 'error-calibrated

399 ABC'. B – D: Posterior distributions. Bars are 'error-calibrated ABC', lines are exact Bayesian
400 regression, all differences nonsignificant (Kolmogorov-Smirnov, $p > 0.01$). The true θ_1, θ_2 and θ_3
401 were -2, 1 and 2, respectively, marked on the x-axes by arrows; posteriors are centred differently
402 because of the added noise and the priors that were used. On the horizontal axes, ticks are placed
403 at the mean of the exact posterior density and three standard deviations above and below.
404
405 **Figure 2. Coverage for the quadratic example.** A - C: 'Error-calibrated ABC'. D – F:
406 'Rejection ABC' for parameter θ_3 at different acceptance rates *m*. Asterisks mark significant
407 departures from uniformity (Kolmogorov-Smirnov, $p < 0.01$).
408
409 **Figure 3. Body masses and cocoon productions in the earthworm experiments.** The black
410 lines show the empirical data (Reinecke and Viljoen 1990, Gunadi et al. 2002, Gunadi and
411 Edwards 2003), the thick grey line is the 'best-fitting run' and the semi-transparent grey lines are
412 the 'posterior predictive check', i.e., the output of 100 new simulations using random samples
413 from runs accepted by 'error-calibrated ABC'. Only every 6^{th} data point, marked by a circle, was
414 used in the analysis; those marked by a cross were removed to improve acceptance rates. Arrows
415 indicate when food was added (↑) or removed (↓). See van der Vaart et al. (2015) for details.
416
417 **Figure 4. Posterior distributions for the earthworm model.** Black lines show 'error-calibrated
418 ABC' accepting 108 runs; grey lines 'rejection ABC' accepting 100. Circles represent medians,
419 whiskers 95% credible intervals. Asterisks mark significant differences (Kolmogorov-Smirnov, p
420 < 0.01). All parameter values were scaled by dividing by the corresponding literature value.