Hindawi
Discrete Dynamics in Nature and Society
Volume 2022, Article ID 8969903, 14 pages
https://doi.org/10.1155/2022/8969903

Research Article

Model Selection and Parameter Estimation for an Improved Approximate Bayesian Computation Sequential Monte Carlo Algorithm

Yue Deng,[1] Yongzhen Pei,[2] Changguo Li,[3] and Bin Zhu[1]

[1]*School of Software, Tiangong University, Tianjin 300387, China*
[2]*School of Mathematical Science, Tiangong University, Tianjin 300387, China*
[3]*Department of Basic Science, Army Military Transportation University, Tianjin 300161, China*

Correspondence should be addressed to Yongzhen Pei; peiyzh_team@sina.com

Received 24 April 2022; Accepted 12 June 2022; Published 30 June 2022

Academic Editor: Abdellatif Ben Makhlouf

Model selection and parameter estimation are very important in many fields. However, the existing methods have many problems, such as low efficiency in model selection and inaccuracy in parameter estimation. In this study, we proposed a new algorithm named improved approximate Bayesian computation sequential Monte Carlo algorithm (IABC-SMC) based on approximate Bayesian computation sequential Monte Carlo algorithm (ABC-SMC). Using the IABC-SMC algorithm, given data and the set of two models including logistic and Gompertz models of infectious diseases, we obtained the best fitting model and the values of unknown parameters of the corresponding model. The simulation results showed that the IABC-SMC algorithm can quickly and accurately select a model that best matches the corresponding epidemic data among multiple candidate models and estimate the values of unknown parameters of model very accurately. We further compared the effects of IABC-SMC algorithm with that of ABC-SMC algorithm. Simulations showed that the IABC-SMC algorithm can improve the accuracy of estimated parameter values and the speed of model selection and also avoid the shortage of ABC-SMC algorithm. This study suggests that the IABC-SMC algorithm can be seen as a promising method for model selection and parameter estimation.

1. Introduction

In many fields of engineering and science, researchers or engineers are dealing with model selection and comparison problems. The selection of the most suitable model among several competitive models is the necessary basis to determine whether the data can accurately estimate and predict the characteristics of data. In reality, it may be a challenge to choose a model that best matches the real data among some similar models, because it requires a deep understanding of the nature of things, in addition, if the parameters in the similar models are also unknown and the reliability of model selection is questionable. To get a reliable model, it is necessary to estimate the values of unknown parameters in the model. Infectious diseases that have occurred in recent years have significantly affected public health and the economy. Therefore, it is very important to perform model selection and parameter estimation in similar models in the process of infectious disease analysis, prediction, and control.

More attention has been paid to model selection and parameter estimation in recent decades. Given several potential models and one or more sets of data, the model selection should be able to select the best fitting model and estimate the values of unknown parameters in the model, to better fit the data. Several approaches have been applied to model selection, among which the Bayesian method is the most popular. The Bayesian theory is a very comprehensive approach and has universal applicability to the method of inferring models from data. Many different examples illustrated the application of the Bayesian methods [1–6]. When the likelihood function is very complex or difficult to

calculate, the Markov chain Monte Carlo (MCMC) method that can obtain approximate posterior distributions of parameters through sampling is successfully applied in model selection [7]. A practical solution combining particle filter model identification algorithm with real-time measurement system was proposed [8]. Skilling [9, 10] has proposed nested sampling (NS) as an alternative way of handling model selection and parameter estimation. In [11], when the distance between observed data and simulated data is the smallest, the observed data in likelihood are replaced by simulated data in the ABC algorithm. The authors in [12, 13] introduced the application of approximate Bayesian computation based on the sequential Monte Carlo (ABC-SMC) algorithm in model selection and parameter estimation. The advantage of ABC-SMC algorithm is that the prior distributions of parameters are adaptive, so it can study the complex posterior distributions of parameters more effectively. Reference [14] offered a recalibration posterior processing method that satisfies the coverage attribute to improve the quality of posterior distributions of parameters of ABC algorithm. Several criteria have been proposed to deal with the goodness of fit between the candidate model and data when dealing with model selection, such as AIC [15, 16], weighted Bayesian information criterion (BIC) [17], and Bayes factors, but all criteria are an approximation of Bayes factors [18]. These criteria are related to the marginal likelihood approximation and are also commonly used in the Bayesian inference [19, 20]. AIC is still the most widely used information criterion for ranking models among IT methods.

The ABC-SMC algorithm is a classic algorithm that provides the possibility to select the most suitable model among multiple competing models and estimate the values of unknown parameters of the model. However, a tolerance sequence is required as a selection criterion for accepting or rejecting sampling parameters in this algorithm. More seriously, the ABC-SMC algorithm must manually define an appropriate threshold sequence to ensure the accuracy of the algorithm, but choosing an appropriate threshold sequence requires us to try many times, which may be very troublesome and time-consuming. Another problem of the ABC-SMC algorithm is that the algorithm selects an appropriate model at each iteration instead of selecting an optimal model at the end of the algorithm, which results in a longer computation time and lower efficiency of the algorithm. So, is there a better way to solve these problems? Is it possible to choose the best model to fit the corresponding data?

This study is intended to investigate the effects of the proposed algorithm in model selection and parameter estimation. To overcome the drawbacks of classical ABC-SMC algorithm, we propose an IABC-SMC algorithm based on the ABC-SMC algorithm and recalibration postprocessing method. Taking dengue outbreak data and A/H1N1 outbreak data of infectious disease as examples, the IABC-SMC algorithm is used to estimate the values of unknown parameters of classic model including logistic model and Gompertz model, and the model that best matches the corresponding data is selected. The simulations show that the parameter values estimated by the IABC-SMC algorithm are very accurate, and the model that matches the data can be quickly selected from multiple candidate models. By comparing the IABC-SMC algorithm with the ABC-SMC algorithm, we can see many advantages of the IABC-SMC algorithm.

2. Methods

2.1. Background Knowledge. In this section, we reviewed the theory and some details of ABC-SMC algorithm and recalibration postprocessing method, before introducing the IABC-SMC model selection algorithm.

2.1.1. Approximate Bayesian Computation Based on Sequential Monte Carlo Algorithm. The approximate Bayesian computation (ABC) algorithm is a Bayesian inference method developed in recent years based on data simulation. When dealing with complex or computationally tricky likelihood problems, ABC is an improved Bayesian inference algorithm for the purpose of inferring the posterior distributions of parameters. Based on the ABC algorithm, many methods have been extended including the ABC rejection sampler algorithm and ABC MCMC algorithm [21, 22]. ABC rejection algorithm is one of the basic algorithms of ABC. When the prior distributions of parameters are far from the posterior distributions of parameters, it may lead to too long computation time of the ABC rejection algorithm. The potential advantage of ABC MCMC algorithm is that it saves computation time of algorithm due to the introduction of acceptance probability. However, this algorithm can cause sample values of parameters to be trapped in a low probability region for a long time and we may never get a good approximation of parameters. To solve these problems, the concept of particle filtering has been introduced. The ABC algorithm is accelerated using a large pool of candidate objects called particles instead of selecting a candidate particle. In each step of the algorithm, the particles are interfered and filtered by distance metric and weights, and eventually, the particle pool becomes closer and closer to meet the requirement of the posterior estimation of parameters. This method is named ABC-SMC.

The ABC-SMC Algorithm 1 is described as follows [23].

2.1.2. Recalibration Postprocessing Method. A large number of postprocessing methods have been mentioned to correct the deviation between posterior distributions of parameters of the ABC algorithm and the true distributions of parameters. The purpose of the recalibration postprocessing method introduced in [14] is to produce an approximation posterior distribution of parameters that is closer to the true distributions of parameters. Not only does this method improve the accuracy of posterior distributions of parameters but also it avoids some shortcomings of existing posterior processing methods. So, the recalibration method can directly sample from the

S1 Define the threshold values $\varepsilon_1, \ldots, \varepsilon_H$ (larger at the beginning and decrease gradually), start with iteration $h = 1$
S2 Set the particle indicator $j = 1$.
S3 If $h = 1$, sample θ^* from the prior distribution $\pi(\theta)$. A simulated data set $D_{(f)}(\theta^*)$ of F_1 times is generated and the value of $f_h(\theta^*)$ is calculated, where $p(D|\theta)$ is a posterior distribution and $D_{(f)} \sim p(D|\theta)$ for any deterministic parameter θ and F_1, $f_h(\theta^*) = \sum_{f=1}^{F1} I\left\{d(D_0, D_{(f)}(\theta^*)) \leq \epsilon_1\right\}$ represents the approximation degree between θ^* and the true parameter, D_0 is the experimental data set and I is an indicative function.
If $h > 1$, sample a particle from the last generation $\theta_{h-1}^{(j)}$ with weight $v_{h-1}^{(j)}$, $j = 1, 2, \ldots, N$ and use a kernel function K_h to disturb the particle to gain θ^*.
If $\pi(\theta^*) = 0$ or $f_h(\theta^*) = 0$, return to the beginning of $S3$.
S4 Set $\theta_h^j = \theta^*$ and determine the weights of the estimated particle θ_h^j, $v_h^{(j)} = \begin{cases} f_h(\theta_h^j) & \text{if } h = 1 \\ (\pi(\theta_h^j) f_h(\theta_h^j) / \sum_{j=1}^{N} K_h(\theta_{h-1}^j, \theta_h^j)) & \text{if } h > 1 \end{cases}$
If $j < N$, update $j = j + 1$ and go back to $S3$ until all the particles and their distributions are obtained.
S5. Normalize the weights $v_h^{(j)}$, If $h < H$ (number of threshold values), update $h = h + 1$ and go back to $S2$.

Algorithm 1: ABC-SMC algorithm.

approximate posterior distributions of samples or improve the efficiency of other posterior adjustment methods.

Recalibration postprocessing algorithm presents a standard parameter posterior simulation algorithm, which completes the recalibration process of parameters. $\theta^{(i)}$ $(i = 1, \ldots, N)$ sampled from the prior distribution of parameter is substituted into the model $\pi(y|\theta)$ to get the simulation value $y^{(i)}$. Then, the weight of parameters $\omega^{(i)}$ is calculated according to the kernel function; that is, $\omega^{(i)} \propto k_h(\|y^{(i)} - y_{obs}\|)$ and $K_h(u)$ is a smoothing kernel with scale parameter $h > 0$. The marginal distribution function $\widetilde{F}_{j,y_{obs}}(\cdot)$ based on y_{obs} is constructed with the weighted parameters $\left\{\theta^{(i)}, \omega^{(i)}\right\}_{i=1}^{N}$. Marginal distribution function $\widetilde{F}_{j,y_{(i)}}(\cdot)$ based on $y_{(i)}$ is constructed in the same way with $\left\{\theta^{(m)}, v^{(m)}\right\}_{m=1,m\neq i}^{N}$ as samples, where $\left\{\theta^{(m)}, v^{(m)}\right\}_{m=1,m\neq i}^{N}$ is the weighted parameters. It is assumed that $\widetilde{F}_{j,y_{obs}}(\cdot) \approx \widetilde{F}_{j,y_{(i)}}(\cdot)$, and then, the probability of $\theta_j^{(i)}$ is calculated as $p_j^{(i)} = \widetilde{F}_{j,y_i}(\theta_j^{(i)})$. The adjusted parameters can be obtained according to several results in [14], namely $\widehat{\theta}^i = \widetilde{F}_{j,y_{obs}}^{-1}(p_j^{(i)})$.

The recalibration postprocessing method proceeds as follows:

Data simulation and weighting:

For $i = 1, \ldots, N$:

M1.1 Sample $\theta^{(i)}$ from prior distribution $\pi(\theta)$.

M1.2 $y^{(i)}$ is obtained from likelihood $\pi(y|\theta)$.

M1.3 Compute the sample weight $w^{(i)} \propto K_h\|y^{(i)} - y_{obs}\|$, where $K_h(u)$ is a smoothing kernel with scale parameter $h > 0$.

Recalibration:

M2.1 For $j = 1, \ldots, d$, construct the jth marginal distribution function $\widetilde{F}_{j,y_{obs}}(\cdot)$ according to sample $\left\{\theta^{(i)}, w^{(i)}\right\}_{i=1}^{N}$.

M2.2 Construct the jth marginal distribution function $\widetilde{F}_{j,y^{(i)}}(\cdot)$ with $\left\{\theta^{(m)}, v^{(m)}\right\}_{m=1,m\neq i}^{N}$ as samples in the same way of step $M1.3$ and $M2.1$.

M2.3 Calculate the probability of $\theta_j^{(i)}$ in $\widetilde{F}_{j,y^{(i)}}(\cdot)$, $p_j^{(i)} = \widetilde{F}_{j,y^{(i)}}(\theta_j^{(i)})$.

M2.4 (optional) Correct $p_j^{(i)}$ using a regression adjustment.

M2.5 Set $\widehat{\theta}^i = \widetilde{F}_{j,y_{obs}}^{-1}(p_j^{(i)})$.

2.2. Model Selection Algorithm. Mathematical models play an important role in understanding how the disease spread. There is an evidence that mathematical models have the ability to inform policymakers, in particular the feasibility of achieving the ambitious goal of keeping the prevalence of moderate and severe infections below 1% by 2020 [24]. So, model prediction based on appropriate epidemiological data is very important. The goal of ABC algorithm is to obtain approximate posterior distributions of parameters that are easy to calculate the following:

$$\pi(\theta|y_{\text{obs}}, M) \propto L(y_{\text{obs}}|\theta, M)\pi(\theta|M), \tag{1}$$

where M is a model based on a series of parameter θ, $\pi(\theta|M)$ represents the prior distribution of the parameter space, and $L(y_{obs}|\theta, M)$ is the likelihood of the observed data y_{obs} for a given series of parameter θ. To overcome the problem of intractable likelihood function, ABC algorithm compared the observed value with the simulated value and accepted the simulated value when the distance between the observed value and the simulated value is less than the artificial threshold. In the ABC-SMC algorithm, it sampled from a set of parameters and treated each parameter vector set as a particle, instead of having only one parameter vector at a time.

The disadvantage of ABC-SMC algorithm is that an appropriate threshold sequence must be selected to ensure the accuracy of the algorithm, and it is very troublesome and time-consuming to choose an appropriate threshold

sequence. Generally, if the defined tolerance sequence is too long, it will lead to too many simulations of the algorithm and take a long time. On the contrary, the posterior distributions of the estimated parameters are inaccurate if the defined tolerance sequence is too short. The principle of the ABC-SMC algorithm is to find the optimal model and parameters in each iteration of the algorithm, rather than selecting the optimal model and parameters at the end of the algorithm, which is also time-consuming. Therefore, to overcome these shortcomings, we proposed the improved approximate Bayesian computation algorithm (IABC-SMC) based on the ABC-SMC algorithm. The principle of the IABC-SMC algorithm is to calculate the values of unknown parameters of each candidate model separately, then recalibrate the values of unknown parameters in the model, and finally select the model that best matches the data. In each iteration of the algorithm, the particles are selected by the distance between simulated data and observed data to avoid setting the threshold sequence manually. Then, the posterior distributions of unknown parameters can be adjusted through the recalibration postprocessing method to make them closer to the true distributions of unknown parameters in the model. So, the IABC-SMC algorithm also improves the accuracy of estimated unknown parameters.

In the first iteration, the IABC-SMC algorithm calculated the distance between the simulated data and the observed data and selected some particles with an acceptance rate when the distance between the simulated data and the observed data is very close, which means that these particles are selected randomly to avoid setting threshold sequence artificially. The weights of these particles are all 1. In the second iteration, the particles selected randomly from the previous generation were substituted into the model to get the simulated data, and the particles with an acceptance rate were selected through the distance between the simulated data and the observed data. Finally, the particles of the second iteration were obtained by perturbation kernel. The weights of these particles were updated and normalized. After several iterations of this algorithm, preliminary posterior distributions of parameters can be obtained, but these distributions are different from the true distributions of parameters. Then, we adjusted the above distributions by the recalibration postprocessing method to make the adjusted distributions closer to the true distributions of parameters.

The IABC-SMC Algorithm 2 is described as follows.

2.3. Model Evaluation Criteria. Recently, more and more scientists are using novel model selection methods to analyze data. The AIC method is a popular method among these novel methods [25–27]. AIC provides a standard to balance the complexity of estimated model and the goodness of data. This approach allows people to compare multiple competing models and estimate which model is closest to the "real" process behind the epidemiological phenomenon being studied. Accordingly, AIC itself is meaningless, but its significance comes from the comparison of the model and AIC value. The model with the smallest AIC value is the "closest model." The calculation of AIC is not difficult, and it is counted as follows:

$$\mathrm{AIC} = -2 \cdot \ln(L) + 2k, \tag{2}$$

where L is likelihood function and k is number of parameters in the model.

Nevertheless, there are various controversies about the use of AIC [16], and many alternative methods have been proposed. In the meantime, BIC is proposed as a special alternative to AIC, which is superior to AIC in the average method of IT model [17]. BIC is denoted as follows:

$$\mathrm{BIC} = -2 \cdot \ln(L) + k\ln(n), \tag{3}$$

where L is likelihood function, k is number of parameters in the model, and n is the length of observed data.

3. Data and Results

3.1. Data and Models. We employed the data from China's Centers for Disease Control and Prevention website on the number of confirmed cases during the dengue outbreak from 2014 to 2015 and the number of confirmed cases during the A/H1N1 outbreak from 2009 to 2010. The data listed in Tables 1 and 2 represent the monthly cumulative confirmed number of dengue and A/H1N1 reported by hospitals across China, respectively. We considered these data to be observed data.

Logistic model and Gompertz model are widely used single-population models that can be easily used to fit data and estimate the values of unknown parameters. They are also two alternative models of dengue data and A/H1N1 data in this study.

The logistic model is shown as follows [28]:

$$x'(t) = rx(t)\left(1 - \frac{x(t)}{K}\right). \tag{4}$$

The Gompertz model is shown as follows [29]:

$$x'(t) = rx(t)\ln\left(\frac{K}{x(t)}\right). \tag{5}$$

For convenience, the above two models are denoted as M_1 and M_2. $x(t)$ in (4) and (5) represents the number of confirmed cases of dengue and A/H1N1 at time t, respectively. There are two distinct unknown parameters r and K. The positive parameter r represents the intrinsic growth rate, which reflects the propagation capacity of infectious diseases under ideal conditions. K denotes the maximum environmental capacity of infectious diseases.

Our purpose is to select the best model among the above two commonly used models based on the infectious disease data listed in Tables 1 and 2, which will help us to evaluate the characteristics of infectious diseases. Therefore, we need to estimate the values of unknown parameters r and K of each candidate model according to the data, to determine the best fitting model.

The first generation
Q1.1 Define the number of iteration $t = 1$, the number of particles $i\,(i = 1, \ldots, n)$.
Q1.2 Sample $\theta^{(i)}$ from prior distribution $\pi(\theta)$ and get simulated values $y^{(i)}$ from model $\pi(y|\theta)$.
Q1.3 An acceptance rate of particles θ^* is selected from $\pi(\theta)$ and simulated values y^* are selected from y when the simulated value y^* is close to the observed value y_{obs}.
Q1.4 Set $\theta_t^{(i)} = \theta^*$ and fix the weights $w_t^{(i)} = 1\,(i = 1, \ldots, n * \text{acceptance rate})$.
The 2. . .**T generation**
Q2.1 Define the number of iteration $t\,(t = 2, \ldots, T)$, the number of particles $i\,(i = 1, \ldots, n)$ and the particle dimension $j\,(j = 1, \ldots, m)$.
Q2.2 Select n particles with weights w_{t-1} from previous generation particles randomly and use the kernel function K to perturb those particles, an acceptance rate of particles θ^{**} is selected from $\pi(\theta)$ and simulated values y^{**} are selected from y when the simulated value y^{**} is close to the observed value y_{obs}.
Q2.3 Set $\theta_t^{(i)} = \theta^{**}$ and fix the weight of each particle $w_t^i = (\pi(\theta_t^i)/\sum_{j=1}^n w_{t-1}^j k_t(\theta_{t-1}^j, \theta_t^i))$.
Q2.4 Normalize the weights $w_t^{(i)}$.If $t < T$, update $t = t + 1$ and return to Q2.1.
Recalibration
Q3.1 According to the particles and the simulation values obtained by the Tth iteration, the weight v of each particle is calculated and $v^{(i)} \propto k_h(\|y^{(i)} - y_{obs}\|)$, where $k_h(0)$ is a smooth kernel with scale parameter $h > 0$.
Q3.2 For $j = 1, \ldots, m$, construct the jth marginal distribution function $\widetilde{F}_{j,y_{obs}}(\cdot)$ according to sample $\left\{\theta^{(i)}, v^{(i)}\right\}_{i=1}^{Q}$, where $v^{(i)} > 0$ and Q is the total number of particles in the marginal distribution function.
Q3.3 Construct the jth marginal distribution function $\widetilde{F}_{j,y_i}(\cdot)$ according to sample $\left\{\theta^{(x)}, v^{(x)}\right\}_{x=1,x\neq i}^{Q}$ in the same way of Q3.1 and Q3.2.
Q3.4 Calculate the probability of $\theta_j^{(i)}$ in $\widetilde{F}_{j,y_i}(\cdot)$, $p_j^{(i)} = \widetilde{F}_{j,y_{(i)}}(\theta_j^{(i)})$.
Q3.5 Calculate the adjusted particle, $\widehat{\theta}^{(i)} = \widetilde{F}_{j,y_{obs}}^{-1}(p_j^{(i)})$.
Model output
Q4 According to the given model, adjusted parameters and weights, the evaluation criteria of AIC and BIC are obtained, then the most fitting model can be selected.

Algorithm 2: IABC-SMC algorithm.

Table 1: Monthly cumulative confirmed data on dengue outbreaks from April 2014 to March 2015.

Month	4	5	6	7	8	9	10	11	12	1	2	3
Data	5	28	53	208	995	15754	44550	47110	47290	47309	47331	47356

Table 2: Monthly cumulative confirmed data on A/H1N1 outbreaks from May 2009 to April 2010.

Month	5	6	7	8	9	10	11	12	1	2	3	4
Data	22	566	1587	3170	19136	52079	96349	125128	131059	131893	132308	132431

3.2. Numerical Results. Our focus so far has been on the IABC-SMC algorithm, model evaluation criteria, and two alternative models with unknown parameters of the intrinsic growth rate r and maximum environmental capacity K. To simplify the interpretation of our mathematical results, we continued to discuss them by numerical simulation.

3.2.1. Results of Dengue

(1) Simulation 1: Results of IABC-SMC Algorithm. To get the results of model selection and parameter estimation of dengue by the IABC-SMC algorithm, the number of initial infections $x_0 = 5$ according to Table 1, the total number of algorithm iterations $T = 6$, and the total number of parameters $n = 6000$ is used here as the initial condition. The infection time of dengue in China was from April 2014 to March 2015 (12 months). It is proved by practice that the Monte Carlo error is minimal when the acceptance rate is 0.4, so we set the acceptance rate to 0.4.

When applying the IABC-SMC algorithm to estimate unknown parameters of model and make model selection of dengue disease, we assumed that the prior distribution of each estimated parameter is uniformly distributed, $r \sim U(0, 2)$ and $K \sim U(40000, 60000)$. The disturbance added by each sampled particle is uniform, and r and K are 0.1 and 1000, respectively. In this experiment, two particles are sampled simultaneously and the IABC-SMC algorithm ends when the most fitting model is selected for dengue data. The algorithm abstracts the parameter estimation of the above two models and the real data of dengue.

Figure 1 shows the histograms of intrinsic growth rate r (Figure 1(a)) and maximum environmental capacity K (Figure 1(b)) of the logistic model. These parameters are obtained by the IABC-SMC algorithm. X-coordinate indicates the range of estimated parameter, and Y-coordinate indicates the frequency of parameter occurrence. In Figure 1,

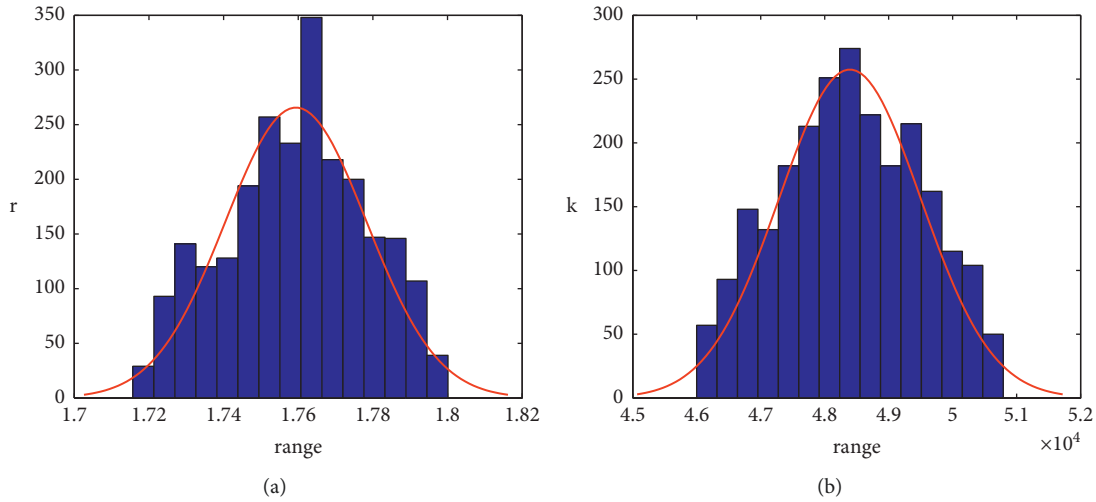

Figure 1: Parameter histograms of logistic model of the IABC-SMC algorithm. (a) Estimation of r. (b) Estimation of K. Here, the initial states are $n = 6000, x_0 = 5, a = 0.4, T = 6$, and $t = 12$.

Table 3: Parameter ranges and statistics of the logistic model.

Parameter	Lower bound	Upper bound	Mean	Std.	[2.5th, 97.5th] percentiles
r	1.7157	1.8001	1.7596	7.2548×10^{-4}	[1.7231, 1.7927]
K	45998	50790	48393	42.6146	[46351, 50413]

the range of parameter r is between 1.72 and 1.8 and the range of parameter K is between 4600 and 5100, which indicates that the range of posterior distributions of these two parameters is very small and concentrated. The distributions of both parameters are all close to normal distributions. When r is about 1.764, the cumulative number of r reaches a peak of about 350 times. When the parameter K is 48500, the cumulative number of K reaches a peak of about 270 times.

The parameter statistics associated with candidate model can be estimated. Table 3 gives the ranges of parameters and parameter statistics of approximate posterior distributions of logistic model. As can be seen from Table 3, the two parameters are estimated precisely and the results of these two parameters are excellent because the most parameter values are within the [2.5th, 97.5th] percentiles.

Figure 2 displays the histograms of intrinsic growth rate r (Figure 2(a)) and maximum environmental capacity K (Figure 2(b)) of the Gompertz model. They are available by the IABC-SMC algorithm, which is the same method as above. X-coordinate indicates the range of estimated parameters, and Y-coordinate indicates the frequency of occurrence of parameters. As can be seen from Figure 2, the distributions of these two parameters are all similar to normal distribution, but both distributions of these parameters in Figures 2(a) and 2(b) have two peaks. When r is 0.5 and 0.53, the peak of cumulative times of r all reaches about 235 times. When the parameter K is 50000 and 54000, the peak of cumulative number of K all reaches about 280 times. So, the estimated parameters are not particularly good.

Table 4 expresses the parameter ranges and the parameter statistics of the Gompertz model of approximate posterior distribution. We can see that the variances of these two parameters in Table 3 are smaller than that in Table 4, which indicates that the parameter range estimated by the logistic model is more accurate.

To further study which model is more credible and verify the posterior distributions of parameters, some calculations and simulations were performed. As mentioned above, the models are sorted by the value of AIC and the best approximation model is the model with the smallest AIC value. Therefore, AIC is an important element to measure the matching degree between model and data. The selection result of the two candidate models is based on the AIC values calculated in Table 5, which confirms the decisive evidence for the existence of the model. That is, they are 7831.8 and 43980 for logistic model ($M1$) and Gompertz model ($M2$), respectively. The AIC value of model $M1$ is the smallest, so the best model is logistic model. To further verify the effect of IABC-SMC algorithm, the BIC value of each model was counted. BIC values of models $M1$ and $M2$ are 7832.7 and 43981, respectively. The results also indicate that the logistic model is the best, which is consistent with the results of AIC. Finally, we verified the above results again by comparing the operation time of model selection, because the operation time of $M1$ is 151.48 seconds, which is less than 163.52 seconds of $M2$. Therefore, the logistic model saves the computation time of IABC-SMC algorithm and is more efficient than the Gompertz model.

Using the mean values of the parameters obtained above, the disease prediction figure can be made. Figure 3 manifests the comparison between the observed data and simulated data calculated from the estimated mean values of the

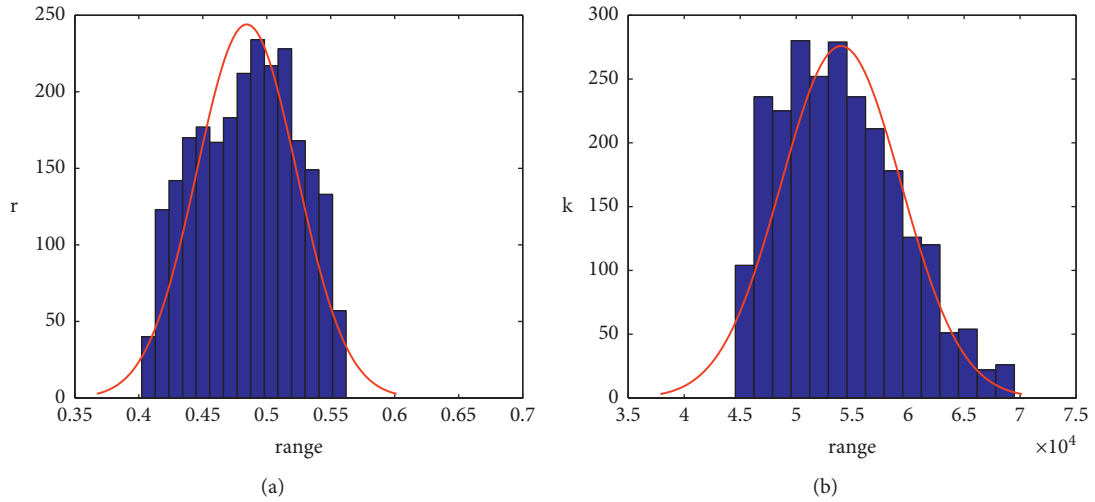

Figure 2: Parameter histograms of the Gompertz model of the IABC-SMC algorithm. (a) Estimation of r. (b) Estimation of K. Here, the initial states are $n = 6000, x_0 = 5, a = 0.4, T = 6$, and $t = 12$.

Table 4: Parameter ranges and statistics of the Gompertz model.

Parameter	Lower bound	Upper bound	Mean	Std.	[2.5th, 97.5th] percentiles
r	0.4022	0.5621	0.4861	6.0838×10^{-4}	[0.4143, 0.5507]
K	44572	69495	53852	44.5627	[45830, 65812]

Table 5: AIC, BIC, and operation time of two models.

Model	Logistic	Gompertz
AIC	7831.8	43980
BIC	7832.7	43981
Operation time (s)	151.48	163.52

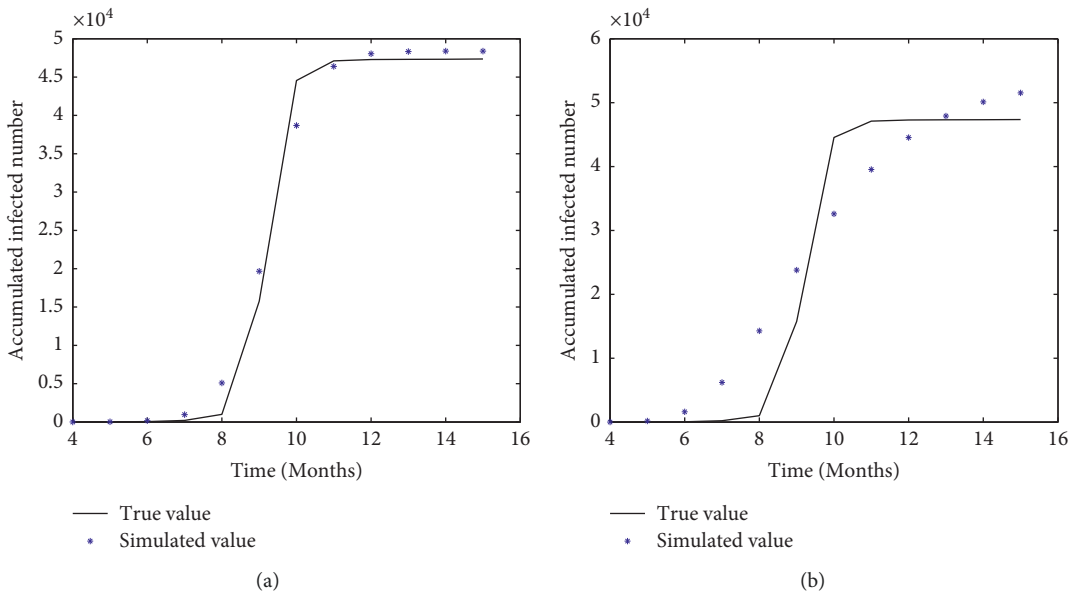

Figure 3: Comparison between the observed data and simulated data of the IABC-SMC algorithm. (a) Logistic model. (b) Gompertz model. Here, the initial states are $n = 6000, x_0 = 5, a = 0.4, T = 6$, and $t = 12$.

parameters. Figure 3(a) is the fitting effect of observed data and simulated data of the logistic model, and Figure 3(b) is the fitting effect of observed data and simulated data of the Gompertz model. The black curve represents the number of confirmed cases of dengue fever from April 2014 to March 2015, and the blue dots represent the simulated data of the model. The trend of Figure 3(a) is that the number of infections rises sharply from August and stabilizes after November. It is clear that the logistic model is very consistent with the data of the outbreak of dengue fever in 2014. The trend of Figure 3(b) is that the number of simulated infections is on the rise, but there is a certain deviation between the simulated data and the observed data. Under the same conditions, the simulated data of the logistic model match the observed data much better than those of the Gompertz model.

From the above results, it can be concluded that the IABC-SMC algorithm can improve the quality of posterior distributions of parameters greatly, obtain more accurate parameter values, and choose the most suitable model. The result of model selection is also consistent with the result of [7].

(2) Simulation 2: Results of ABC-SMC Algorithm. To verify the effects of the IABC-SMC algorithm, we compared the IABC-SMC algorithm, with the ABC-SMC algorithm in [13]. Both algorithms used the same model, experimental data, parameter initial values, and thresholds. The threshold set is selected manually and given as $\varepsilon = 65389, 41043, 33049, 30077, 28770,$ and 16433. The prior probability of each model is equal, i.e., $p(M1) = p(M2) = 1/2$. Figure 4 is the fitting effect of the observed data and simulated data obtained by the ABC-SMC algorithm. The black curve represents the number of confirmed cases of dengue fever from April 2014 to March 2015, and the blue dots represent the simulated data of the ABC-SMC algorithm. Compared with the Gompertz model, the simulated data obtained from logistic model are closer to the observed data, which indicates that the logistic model is better than the Gompertz model. The result of model selection is also consistent with the result of the IABC-SMC algorithm, and the results of model selection are all logistic model. However, from the perspective of simulation effect of algorithm, the fitting effect between the simulated data obtained from the ABC-SMC algorithm and the observed data is worse than that of the IABC-SMC algorithm. From the perspective of simulation time of algorithm, the computation time of ABC-SMC algorithm is 40272 seconds, which far exceeds the computation time of IABC-SMC algorithm.

Figure 5 shows the histograms of parameters r (Figure 5(a)) and K (Figure 5(b)) of the model selected by the ABC-SMC algorithm. The X-coordinate is the range of the parameters, and the Y-coordinate is the frequency of parameter. When r is 0.5, parameter r appears most frequently, which is about 2700 times. However, other values of r occur between 1.5 and 2. The parameter K appears between 40000 and 60000. When K is 55000, K appears most frequently and it is about 680 times. Table 6 shows the parameter summary statistics estimated by the ABC-SMC algorithm. It can be seen from Figure 5 and Table 6 that the ABC-SMC algorithm can also estimate the posterior distributions of unknown parameters, but the parameter range is larger and more dispersed than that of the IABC-SMC algorithm.

These results confirm that the IABC-SMC algorithm has the advantages of high computational efficiency, low time complexity, and more accurate parameter values.

3.2.2. Results of A/H1N1

(1) Simulation 1: Results of IABC-SMC Algorithm. To verify the results of model selection and parameter estimation of A/H1N1 by the IABC-SMC algorithm, the number of initial infections $x_0 = 22$ according to Table 2, total number of algorithm iterations $T = 6$, and the total number of parameters $n = 6000$ are used as the initial condition. The infection time of A/H1N1 disease in China was from May 2009 to April 2010, expressed by $t = 5$ to $t = 16$ (12 months) in diagram. As above, the acceptance rate is also set to 0.4.

When applying the IABC-SMC algorithm to estimate model parameters and make model selection, we assumed that the prior distribution of each estimated parameter is uniformly distributed, $r \sim U(0, 2.5)$ and $K \sim U(100000, 150000)$. The disturbance added by each sampled particle is uniform, and r and K are 0.1 and 1000, respectively. When the IABC-SMC algorithm finishes, the model that best matches the A/H1N1 data can be obtained. The algorithm abstracts the parameter estimation of the above two models and the real data of A/H1N1.

Figure 6 shows the histograms of intrinsic growth rate r (Figure 6(a)) and maximum environmental capacity K (Figure 6(b)) of the logistic model. These parameters are available by the IABC-SMC algorithm. X-coordinate indicates the range of estimated parameters, and Y-coordinate indicates the frequency of parameters. As can be seen from Figure 6, the range of parameter r is between 1.6 and 1.68, and the range of parameter K is between 125000 and 139520, which shows that the range of posterior distributions of these two parameters is very small and concentrated. The distributions of both parameters are all close to normal distribution. When r is about 1.64, the peak of cumulative number of r reaches about 290 times. When the parameter K is 131000, the peak of cumulative number of K reaches about 235 times.

The correlation statistics of parameters r and K related to the logistic model can be obtained. Table 7 gives the range of parameters and the unknown parameters statistics of the approximate posterior distributions of logistic model. As can be seen from Table 7, the parameters all can be well estimated. The results of these two parameters are excellent because their parameter values are within the [2.5th, 97.5th] percentiles.

Figures 7(a) and 7(b) show the histograms of intrinsic growth rate r and maximum environmental capacity K of the Gompertz model, respectively. These parameters are obtained by the IABC-SMC algorithm. X-coordinate indicates the range of estimated parameters, and Y-coordinate

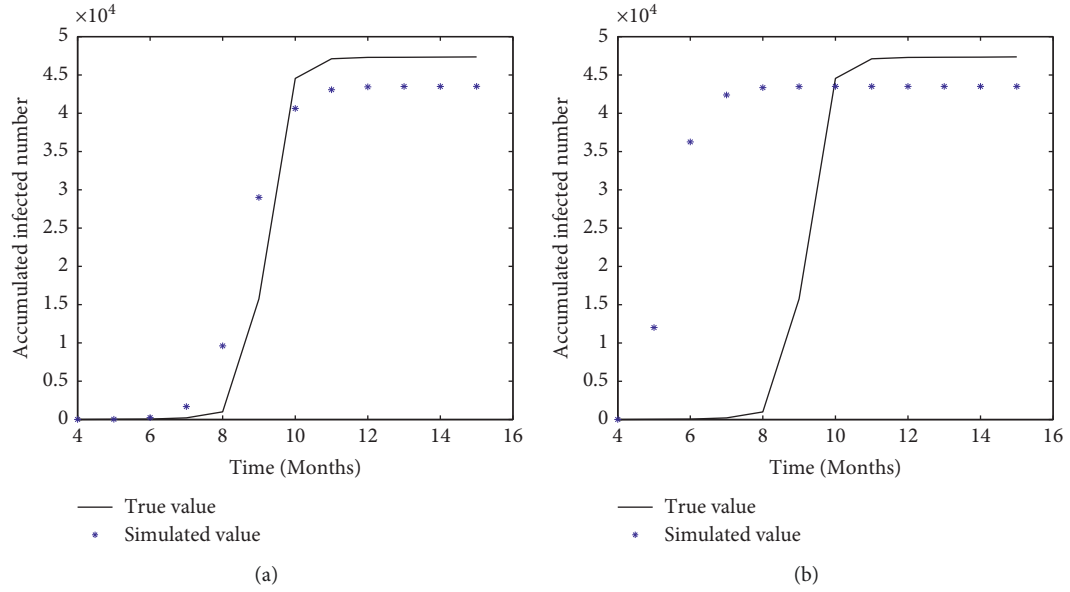

Figure 4: Fitting effect of simulated data and observed data of ABC-SMC algorithm. (a) Logistic model. (b) Gompertz model. Here, the initial states are $n = 6000, x_0 = 5, T = 6, t = 12, \varepsilon = 57453, 22335, 16047, 12872, 11540, 6720,$ and $p(M1) = p(M2) = 1/2$.

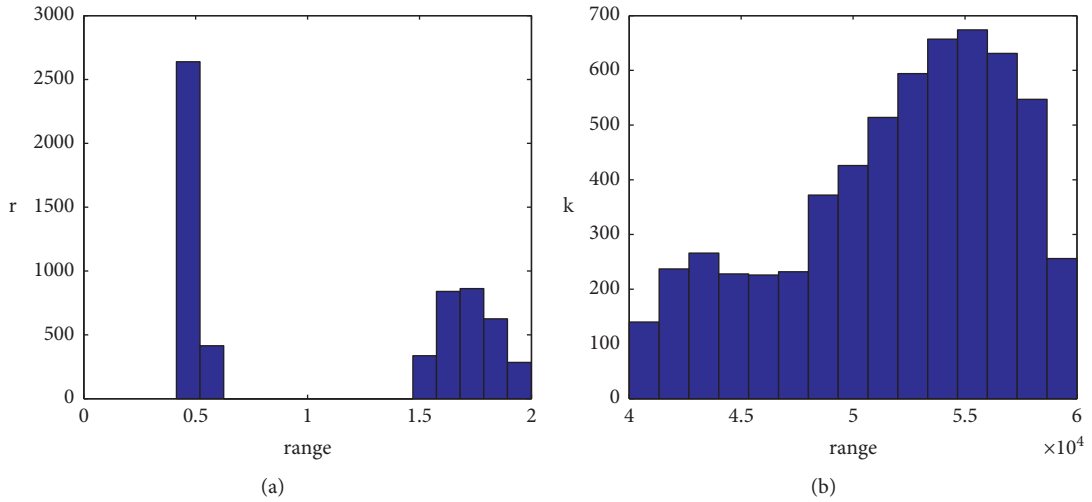

Figure 5: Model parameter histograms of ABC-SMC algorithm. (a) Estimation of r. (b) Estimation of K. Here, the initial states are $n = 6000, x_0 = 5, T = 6, t = 12,$ and $\varepsilon = 65389, 41043, 33049, 30077, 28770,$ and 16433.

Table 6: Parameter summary statistics of ABC-SMC algorithm.

Parameter	Lower bound	Upper bound	Mean	Std.	[2.5th, 97.5th] percentiles
r	0.4138	1.9986	1.0814	0.8011	[0.4287, 1.9285]
K	40003	60000	52019	1.0596×10^{-4}	[41416, 59042]

indicates the total times of parameters. As can be seen from Figure 7, the distributions of parameters r and K do not follow normal distribution and both parameters have two peaks. When r is 0.5 and 0.53, the peak of cumulative number of r reaches about 235 times. When the parameter K is 50000 and 54000, the peak of cumulative number of K reaches about 280 times. So, the results of estimated parameters are not particularly good.

The posterior estimation results of unknown parameters and the statistics related to the Gompertz model are shown

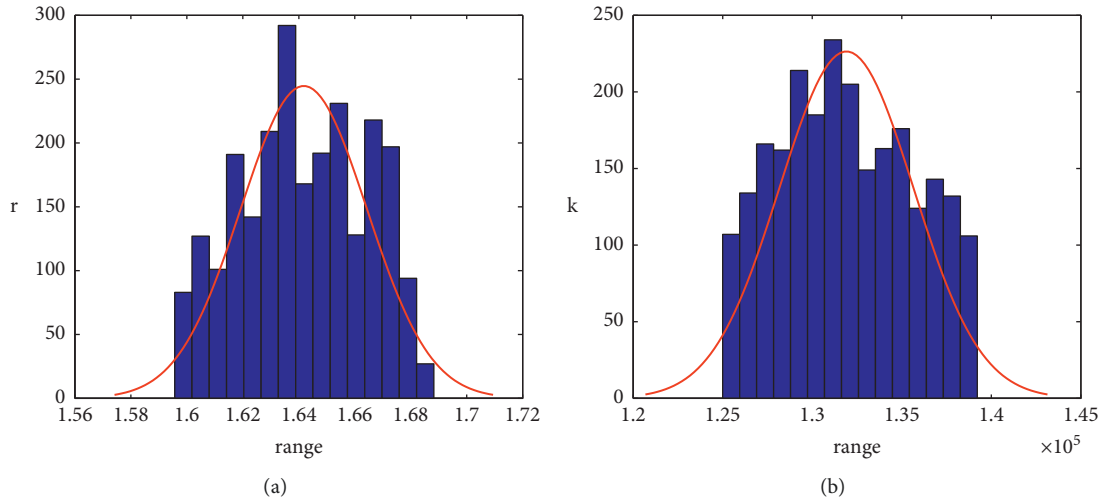

Figure 6: Parameter histograms of logistic model of the IABC-SMC algorithm. (a) Estimation of r. (b) Estimation of K. Here, the initial states are $n = 6000, x_0 = 22, a = 0.4, T = 6$, and $t = 12$.

Table 7: Parameter ranges and statistics of logistic model.

Parameter	Lower bound	Upper bound	Mean	Std.	[2.5th, 97.5th] percentiles
r	1.5983	1.6889	1.6442	6.7627×10^{-5}	[1.6022, 1.6833]
K	125130	139520	132460	93.6276	[125730, 138640]

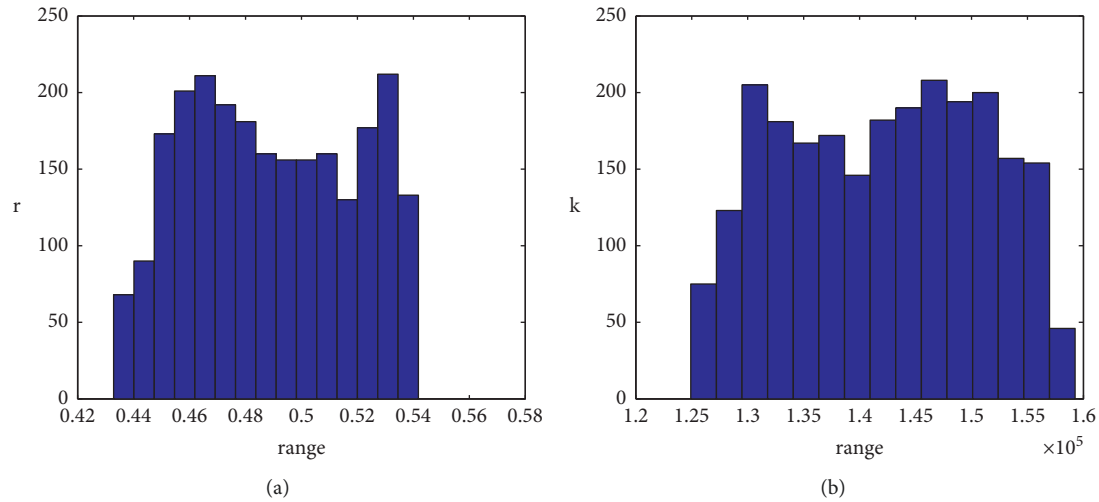

Figure 7: Parameter histograms of Gompertz model of the IABC-SMC algorithm. (a) Estimation of r. (b) Estimation of K. Here, the initial states are $n = 6000, x_0 = 22, a = 0.4, T = 6$, and $t = 12$.

Table 8: Parameter ranges and the statistics of the Gompertz model.

Parameter	Lower bound	Upper bound	Mean	Std.	[2.5th, 97.5th] percentiles
r	0.4336	0.5464	0.4971	6.4659×10^{-4}	[0.4417, 0.5421]
K	125050	161430	140430	95.3831	[127100, 158750]

in Table 8. We can see that the variances of these two parameters in Table 7 are smaller than that in Table 8, which indicates that the parameter range estimated by the logistic model is more accurate.

Similarly, to further verify the effects of the IABC-SMC algorithm, some calculations and simulations are carried out. According to the criteria presented in Table 9, the evidence for choosing logistic model is conclusive. The

Table 9: AIC, BIC, and operation time of two models.

Model	Logistic	Gompertz
AIC	4060	49816
BIC	4061	49817
Operation time (s)	209.6029	213.6122

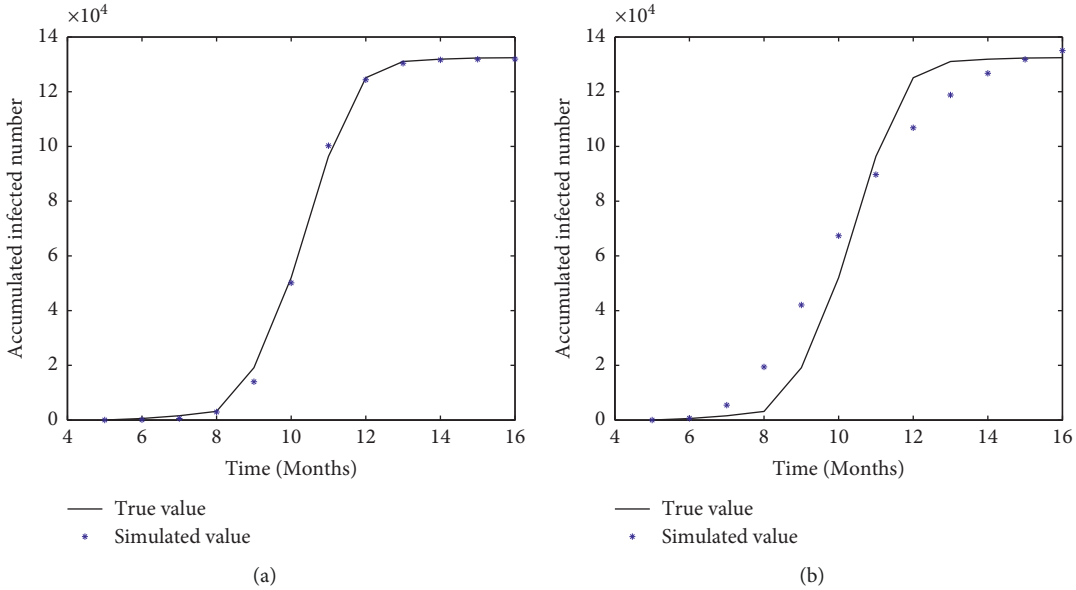

Figure 8: Comparison between the observed data and simulated data of the IABC-SMC algorithm. (a) Logistic model. (b) Gompertz model. Here, the initial states are $n = 6000, x_0 = 22, a = 0.4, T = 6,$ and $t = 12$.

AIC values of $M1$ and $M2$ are calculated as 4060 and 49816, respectively. It is obvious that the AIC value of $M1$ is much smaller than that of $M2$, so the best model is the logistic model. The BIC values of $M1$ and $M2$ also support us to choose logistic model. Finally, we can use the model selection time to verify the above conclusions again. It is easy to notice that $M1$ saves more time than $M2$. So, the logistic model is the "best" model for us to fit the A/H1N1 data.

The fitting results of logistic model and Gompertz model in Figure 8 are plotted based on the observed data and the simulated data calculated using the average values of the estimated parameters. The X-coordinate shows the outbreak time of A/H1N1 disease from May 2009 to April 2010, and the Y-coordinate shows the cumulative number of infections. Obviously, the simulation data obtained by the logistic model and the A/H1N1 epidemic data have the best fitting effect. The result of model selection of A/H1N1 disease is also consistent with that of dengue disease.

(2) Simulation 2: Results of ABC-SMC Algorithm. To verify the efficiency of the IABC-SMC algorithm, the IABC-SMC algorithm is compared with ABC-SMC algorithm. Both algorithms used the same model, experimental data, parameter initial values, and threshold. The threshold set is selected manually and given as $\varepsilon = 168655, 106154, 78152, 64957, 59245,$ and 41355. The prior probability of each model is equal; i.e., $p(M1) = p(M2) = 1/2$. Figure 9 is the fitting effect of the observed data and the simulated data obtained by the ABC-SMC algorithm. The black curve represents the number of confirmed cases with A/H1N1 from May 2009 to April 2010, and the blue dots represent the simulated data of the ABC-SMC algorithm. It is not difficult to see from the simulation results that the fitting effect of the observed data and the simulated data obtained by the logistic model is better than that of the Gompertz model. It is consistent with the results of the IABC-SMC algorithm and the results of model selection are all logistic model. Although the final results of model selection of the two algorithms are both logistic model, the simulation data obtained by the IABC-SMC algorithm fit the observed data better. The computation time of ABC-SMC algorithm is 3212 seconds, which is also much longer than that of the IABC-SMC algorithm.

Figure 10 shows the histograms of parameters r (Figure 10(a)) and K (Figure 10(b)) of the selected model obtained by the ABC-SMC algorithm. The X-axis is the range of the parameters, and the Y-axis is the total number of parameter. When r is 0.5, parameter r appears most frequently, which is about 2800 times. However, other values of r occur between 1.5 and 1.8. When the parameter K is between 115000 and 144000, the total number of K is about 200. When the value of K is between 144000 and 150000, the total number of K increases significantly. Table 10 represents the relevant statistical information of the unknown parameters estimated by the ABC-SMC

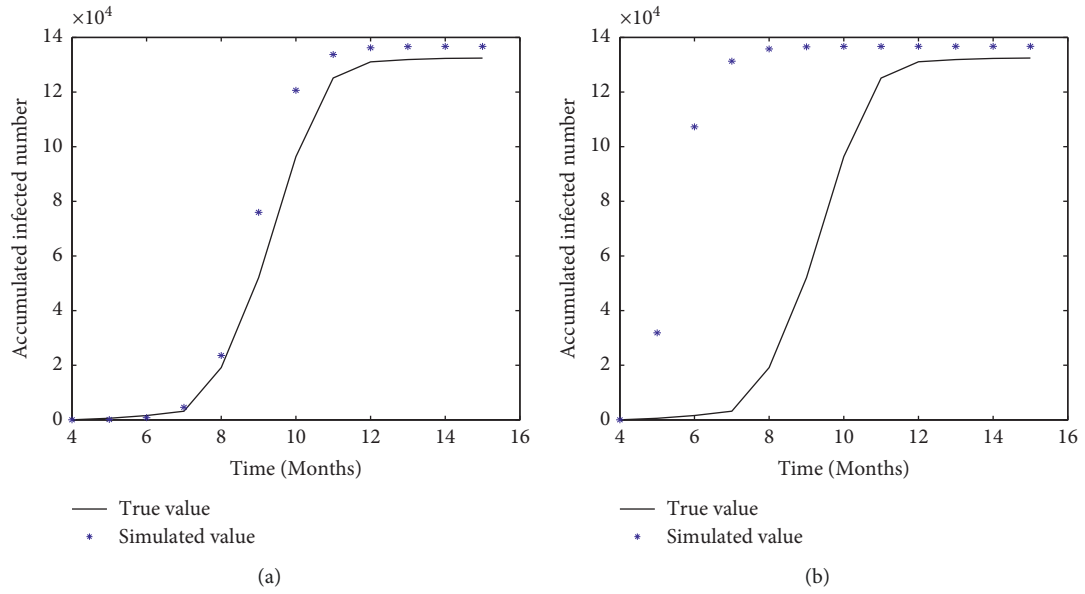

Figure 9: Fitting effect of simulated data and observed data of ABC-SMC algorithm. (a) Logistic model. (b) Gompertz model. Here, the initial states are $n = 6000, x_0 = 22, T = 6, t = 12, \varepsilon = 168655, 106154, 78152, 64957, 59245,$ and 41355, and $p(M1) = p(M2) = 1/2$.

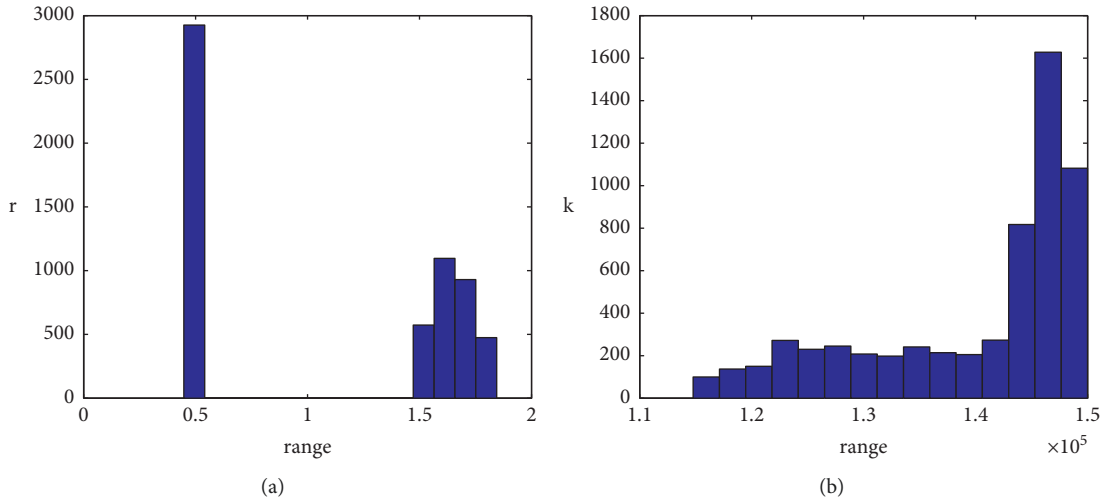

Figure 10: Model parameter histograms of ABC-SMC algorithm. (a) Estimation of r. (b) Estimation of K. Here, the initial states are $n = 6000, x_0 = 22, T = 6, t = 12,$ and $\varepsilon = 168655, 106154, 78152, 64957, 59245,$ and 41355.

Table 10: Parameter summary statistics of ABC-SMC algorithm.

Parameter	Lower bound	Upper bound	Mean	Std.	[2.5th, 97.5th] percentiles
r	0.4477	1.8446	1.0983	1.7324	[0.4556, 1.8013]
K	114780	150000	139490	29446	[118040, 149410]

algorithm. Although the ABC-SMC algorithm can also estimate the posterior information of the unknown parameters, it has a larger and more scattered parameter range than that of the IABC-SMC algorithm. These results confirm the advantages of IABC-SMC algorithm again.

4. Discussion

Many methods have various problems in the selection of models and parameter estimation, such as low efficiency in model selection and inaccuracy in parameter estima-

tion. These problems may lead to the selection of the wrong model and the inaccurate estimation of actual data scale. Our study proposed an IABC-SMC algorithm based on the ABC-SMC algorithm and recalibration post-processing method. We took the reported data of dengue epidemic and A/H1N1 epidemic in China as examples in our study.

We used the IABC-SMC algorithm and two simplest single-population models to analyze the results of model selection of the dengue epidemic data and the A/H1N1 epidemic data and the results of parameter estimation of the selected model. The selected model in two examples is consistent, and the model selected is the logistic model. Compared with the ABC-SMC algorithm, the IABC-SMC algorithm has the advantages of higher computational efficiency, lower time complexity, more fast and accurate model selection ability, and more accurate posterior distributions of parameters. The IABC-SMC algorithm avoids the problem of setting the threshold sequence of the ABC-SMC algorithm manually and the time-consuming problem and also avoids the shortage to find the optimal model and value of unknown parameters of model in each iteration in ABC-SMC algorithm.

Although the alternative model in this study is relatively simple, it demonstrates many promising aspects of the IABC-SMC algorithm, which can be extended to complex system models to deal with model selection and parameter estimation problems effectively in the future. It can be utilized not only for deterministic models but also for stochastic models in the physical, chemical, and biological sciences.

Data Availability

The underlying data supporting the results of the study are found in Tables 1 and 2.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The work was also supported by the National Natural Science Foundation of China (grant nos. 11471243 and 11971023).

References

[1] M. Muto and J. L. Beck, "Bayesian updating and model class selection for hysteretic structural models using stochastic simulation," *Journal of Vibration and Control*, vol. 14, no. 1C2, p. 7C34, 2008.

[2] B. A. Zrate, J. M. Caicedo, J. Yu, and P. Ziehl, "Bayesian model updating and prognosis of fatigue crack growth," *Engineering Structures*, vol. 45, Article ID 53C61, 2012.

[3] Ph. Bisaillon, R. Sandhu, M. Khalil, C. Pettit, D. Poirel, and A. Sarkar, "Bayesian parameter estimation and model selection for strongly nonlinear dynamical systems," *Nonlinear Dynamics*, vol. 82, no. 3, pp. 1061–1080, 2015.

[4] J. L. Beck and K. V. Yuen, "Model selection using response measurements: bayesian probabilistic approach," *Journal of Engineering Mechanics*, vol. 130, no. 2, pp. 192–203, 2004.

[5] R. Sandhu, M. Khalil, A. Sarkar, and D. Poirel, "Bayesian model selection for nonlinear aeroelastic systems using wind-tunnel data," *Computer Methods in Applied Mechanics and Engineering*, vol. 282, pp. 161–183, 2014.

[6] T. G. Ritto and L. C. S. Nunes, "Bayesian model selection of hyperelastic models for simple and pure shear at large deformations," *Computers & Structures*, vol. 156, pp. 101–109, 2015.

[7] W. Liu, S. Tang, and Y. Xiao, "Model selection and evaluation based on emerging infectious disease data sets including A/H1N1 and ebola," *Computational and Mathematical Methods in Medicine*, vol. 2015, Article ID 207105, 14 pages, 2015.

[8] F. Cadini, C. Sbarufatti, M. Corbetta, and M. Giglio, "A particle filter-based model selection algorithm for fatigue damage identification on aeronautical structures," *Structural Control and Health Monitoring*, vol. 24, no. 11, Article ID e2002, 2017.

[9] J. Skilling, *Nested Sampling*, R. Fischer, R. Preuss, and U. V. Toussaint, Eds., American Institute of Physics Conference Series, New York, NY, USA, Article ID 395C405, 2004.

[10] J. Skilling, "Nested sampling for general Bayesian computation," *Bayesian Anal*, vol. 1, no. 4, Article ID 833C860, 2006.

[11] J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman, "Population growth of human Y chromosomes: a study of Y chromosome microsatellites," *Molecular Biology and Evolution*, vol. 16, no. 12, pp. 1791–1798, 1999.

[12] A. Ben Abdessalem, N. Dervilis, D. Wagg, and K. Worden, "Model selection and parameter estimation in structural dynamics using approximate Bayesian computation," *Mechanical Systems and Signal Processing*, vol. 99, pp. 306–325, 2018.

[13] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. Stumpf, "Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems," *Journal of The Royal Society Interface*, vol. 6, no. 31, pp. 187–202, 2009.

[14] G. S. Rodrigues, D. Prangle, and S. A. Sisson, "Recalibration: a post-processing method for approximate Bayesian computation," *Computational Statistics & Data Analysis*, vol. 126, pp. 53–66, 2018.

[15] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde, "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society: Series B*, vol. 64, no. 4, pp. 583–639, 2002.

[16] P. A. Stephens, S. W. Buskirk, G. D. Hayward, and C. M. Del Rio, "A call for statistical pluralism answered," *Journal of Applied Ecology*, vol. 44, no. 2, pp. 461–463, 2007.

[17] W. A. Link and R. J. Barker, "Model weights and the foundations of multimodel inference," *Ecology*, vol. 87, no. 10, pp. 2626–2635, 2006.

[18] P. Congdon, "Bayesian model choice based on Monte Carlo estimates of posterior model probabilities," *Computational Statistics & Data Analysis*, vol. 50, no. 2, pp. 346–357, 2006.

[19] F. Llorente, L. Martino, D. Delgado, and J. Lopez-Santiago, "Marginal Likelihood Computation for Model Selection and Hypothesis Testing: An Extensive Review," 2020, https://arxiv.org/abs/2005.08334, Article ID 08334.

[20] F. Llorente, L. Martino, E. Cuberlo, J. Lopez-Santiago, and D. Delgado, "On the safe use of prior densities for Bayesian model selection," *viXra*, vol. 2110, 2021.

[21] R. J. Boys, D. J. Wilkinson, and T. B. L. Kirkwood, "Bayesian inference for a discretely observed stochastic kinetic model," *Statistics and Computing*, vol. 18, no. 2, pp. 125–135, 2008.

[22] A. Golightly and D. J. Wilkinson, "Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo," *Interface Focus*, vol. 1, no. 6, pp. 807–820, 2011.

[23] S. A. Sisson, Y. Fan, and M. M. Tanaka, "Sequential Monte Carlo without likelihoods," *Proceedings of the National Academy of Sciences*, vol. 104, no. 6, pp. 1760–1765, 2007.

[24] World Health Organization, *Accelerating Work to Overcome the Globalimpact of Neglected Tropical Diseases C a Roadmap for Implementation*, Geneva, Switzerland, 2012.

[25] H. Akaike, "Information theory as an extension of the maximum likelihood principle," *Second International Symposium on Information Theory*, Akademiai Kiado, Hungary, Europe, Article ID 267C281, 1973.

[26] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference*, Springer, New York, NY, USA, 2002.

[27] L. Z. Garamszegi, "Information-theoretic approaches to statistical analysis in behavioural ecology: an introduction," *Behavioral Ecology and Sociobiology*, vol. 65, no. 1, pp. 1–11, 2010.

[28] P. F. Verhulst, "Notice sur la loi que la population suit dans son accroissement. correspondance mathematique et physique publiee par a quetelet, brussels," *Quetelet*, vol. 10, no. 10, Article ID 113C121, 1838.

[29] C. P. Winsor, "The gompertz curve as a growth curve," *Proceedings of the National Academy of Sciences*, vol. 18, no. 1, pp. 1–8, 1932.