Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2018, Vol. 27(7) 1979–1998

DOI: 10.1177/0962280217746444
journals.sagepub.com/home/smm
SAGE

Profile likelihood-based analyses of infectious disease models

Christian Tönsing,[1] Jens Timmer[1,2,3] and Clemens Kreutz[1,2]

Abstract

Ordinary differential equation models are frequently applied to describe the temporal evolution of epidemics. However, ordinary differential equation models are also utilized in other scientific fields. We summarize and transfer state-of-the art approaches from other fields like Systems Biology to infectious disease models. For this purpose, we use a simple *SIR* model with data from an influenza outbreak at an English boarding school in 1978 and a more complex model of a vector-borne disease with data from the Zika virus outbreak in Colombia in 2015–2016. Besides parameter estimation using a deterministic multistart optimization approach, a multitude of analyses based on the profile likelihood are presented comprising identifiability analysis and model reduction. The analyses were performed using the freely available modeling framework *Data2Dynamics* (data2dynamics.org) which has been awarded as best performing within the DREAM6 parameter estimation challenge and in the DREAM7 network reconstruction challenge.

Keywords

Modeling, dynamical systems, ordinary differential equations, parameter estimation, uncertainty analysis, profile likelihood, identifiability analysis, model reduction, vector-borne disease models, Zika virus disease

1 Introduction

Mathematical models can be used to understand mechanisms of spreading of infectious diseases. If the populations are large enough, deterministic models like ordinary differential equation models (ODEs) can be utilized to describe the dynamics of the different stages and for deriving strategies for controlling an epidemic or for predicting future outbreaks.[1]

Most models describe the number or fraction of individuals in several population subsets, usually termed as compartments. In the well-known *SIR* model

$$S \rightarrow I \rightarrow R \tag{1}$$

as an example, individuals are passing from suceptibles S to the infected subset I before becoming recovered R. The velocity of transitions between compartments is determined by rate constants, usually considered as unknown parameters. Schemes like equation (1) can be directly translated into sets of ODEs in generally yielding first-order derivatives which are nonlinear with respect to the dynamic states.

A variety of approaches has been established to estimate parameters in such models from experimental observation,[2] using Bayesian inference networks,[3–6] Markov chain Monte Carlo methods[7] in combination with Gaussian Processes[8] and with focus on robust Bayesian inference,[9] as well as agent-based methods,[10] spatio-temporal autoregression models,[11] but also stochastic maximum likelihood analyses[12] as well as multiple shooting for stochastic systems.[13] The major challenge is that the solution of ODE models is nonlinear with respect to parameters like initial conditions and transition rate constants. Therefore, many classical methods for parameter fitting and assessing uncertainties are only applicable to a strongly limited manner.[14,15]

[1]Institute of Physics, University of Freiburg, Freiburg im Breisgau, Germany
[2]Center for Biosystems Analysis (ZBSA), University of Freiburg, Freiburg im Breisgau, Germany
[3]BIOSS Centre for Biological Signalling Studies, University of Freiburg, Freiburg im Breisgau, Germany

Corresponding author:
Christian Tönsing, Institute of Physics, University of Freiburg, Hermann-Herder-Str. 3, Freiburg 79104, Germany.
Email: christian.toensing@fdm.uni-freiburg.de

The profile likelihood is a generalization of classical approaches like standard errors which are based on the Fisher-Information matrix and it has been shown that the profile likelihood provides reasonable confidence intervals for parameter estimation of ODE models. Moreover, the profile likelihood has been used to assess the identifiability of model parameters,[16] for the determination of informative observations[17] and for the calculation of uncertainties for predictions.[18] Moreover, the profile likelihood pinpoints model components which can be removed without loosing statistical agreement with observations. It thereby provides a strategy for reducing ODE models down to a completely identifiable level of detail.[19]

In this article, these well-established modeling methods are illustrated in the setting of infectious disease modeling. Moreover, insights concerning numerically efficient optimization are provided. The analyses are performed using the MATLAB-based Data2Dynamics modeling environment.[20,21] The presented methodology and its implementation has been awarded twice within the Dialogue on Reverse-Engineering Assessment and Methods (DREAM) challenges for parameter estimation[22] and network reconstruction,[23] there mentioned as "Team Crux." The software package is used for all further analyses. Setup scripts, model definitions, and data files for all presented examples are available within the Data2Dynamics software on www.data2dynamics.org.

2 Methodology

We introduce the methodology using a canonic *SIR* model with the well-known data set of an influenza outbreak at an English boarding school in 1978. In a second step, we apply the same methods and the profile likelihood analyses to a more sophisticated model of the vector-borne disease model with data from the ZIKV outbreak in Colombia in 2015–2016. Hereby the goal is illustration of profile likelihood analyses in a standard example and application study, rather than focusing on credible biological discoveries.

In both examples, we only use raw data time series although additional data, i.e. prior information for model parameters would be available. The intention is to introduce a methodology which is also applicable in applications where less or no prior knowledge is available. Nevertheless, the presented approach can also account for multiple priors but it does not necessarily require it.

2.1 ODE models

In infectious disease models, a set of coupled ordinary differential equations (ODEs)

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{k}) \tag{2}$$

is often used to describe the dynamics of the *compartments* represented by the vector $\boldsymbol{x}$. The term compartment is equivalent to the terms *species*, *compound* or *node* which are more commonly used in other fields. In general the function f can be nonlinear and depends on the compartment states $\boldsymbol{x}(t)$ at time t, on a possibly time depend input function $\boldsymbol{u}(t)$ and on the model parameters $\boldsymbol{k}$, such as transmission rate constants. The initial states of the system at time $t = 0$ are described by $\boldsymbol{x}^{init} = \boldsymbol{x}(t = 0)$. Thus, the time dependence of the system is entirely described by a set of kinetic parameters $\boldsymbol{\theta} = (\boldsymbol{k}, \boldsymbol{x}^{init})$ which contains both, rate constants $\boldsymbol{k}$ and initial conditions $\boldsymbol{x}^{init}$.

2.2 Observations

In addition to prior knowledge, empirical data is used in order to estimate the unknown parameters $\boldsymbol{\theta}$. Usually not all compartments are observed individually and linear combinations like the sum of several compartments may occur as observations. In order to compare the model dynamics simulated for a candidate set of parameter values, empirical data

$$\boldsymbol{y}_{obs}(t_i) = \boldsymbol{g}(\boldsymbol{x}(t_i)) + \boldsymbol{\varepsilon}(t_i) \tag{3}$$

is mapped to the dynamics $\boldsymbol{x}(t_i)$ at time t_i via observation functions $\boldsymbol{g}$ which are not necessarily bijective and invertible. If a single compartment x_j is directly observed, g is the identity function, i.e. $y_{obs} = x_j(t_i) + \varepsilon(t_i)$. In other cases, g may include linear combinations of compartments or contain scaling parameters in the case of relative data or offset parameters if background correction is required. In our illustration examples, independent additive Gaussian errors $\boldsymbol{\varepsilon}(t_i) \sim \mathcal{N}(0, \sigma^2)$ with unknown standard deviation σ are assumed. However, the

methodology is discussed in terms of the likelihood which comprises also other error distributions. Moreover, σ can be described by an error model, e.g. $\sigma^2 = \sigma_{\text{rel}}^2 x^2 + \sigma_{\text{abs}}^2$ as it is assumed for the Zika virus example.

2.3 Maximum likelihood estimation

Maximum likelihood estimation is a general approach for estimating model parameters. The so-called maximum likelihood estimate

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{y}|\boldsymbol{\theta}) = \arg\min -2\log(L(\boldsymbol{y}|\boldsymbol{\theta})) \tag{4}$$

is asymptotically unbiased and has several beneficial statistical properties like consistency and efficiency.[24] In practice, $-2LL(\boldsymbol{\theta}) = -2\log(L(\boldsymbol{y}|\boldsymbol{\theta}))$ is usually minimized instead of maximizing the likelihood because of numerical reasons and since this is equivalent to least-squares estimation in the case of Gaussian observation errors.

For model calibration the model's dynamics, i.e. the observables $\boldsymbol{g}$ are compared to the available data $\boldsymbol{y}_{obs}$. If prior knowledge about the model parameters is available, it can be incorporated by penalizing the log-likelihood LL. For Gaussian priors $\bar{\theta}_i$ with standard errors $\bar{\sigma}_i$, the penalized log-likelihood is given by

$$-2LL_{pen}(\boldsymbol{\theta}) = -2\log(L(\boldsymbol{y}|\boldsymbol{\theta})) + \sum_i \frac{\left(\theta_i - \bar{\theta}_i\right)^2}{\bar{\sigma}_i^2} \tag{5}$$

For ODE models, optimization of equation (4) or (5) has to be performed numerically. In the following, the unpenalized log-likelihood equation (4) is used for all methodology and analyses.

2.4 Integration and optimization

The solution of an ODE system is usually nonlinear with respect to the parameters. In consequence, also the objective function is strongly nonlinear and, thus, parameter estimation in ODE models is numerically challenging. In addition, evaluating the objective function requires numerical integration of the ODEs which is only feasible with limited accuracy. Therefore, it is not possible to calculate derivatives based on finite differences.[25] Moreover, it has been shown that it is usually more efficient to optimize the logarithms of parameters.[26] Another aspect is that even ODE integrators for stiff systems struggle with numerical instabilities for processes with extremely large differences in time scales. Therefore, the parameter space is typically restricted, e.g. to eight orders of magnitude.[25] Although analytical solutions for small ODE systems exist, numerical integration is chosen throughout the manuscript to not restrict applicability of the methodology on small special cases. Due to the nonlinearity of the ODEs and because of the restricted amount of empirical data, likelihood landscapes typically exhibit multiple local optima.

Although it is theoretically possible for stochastic optimization algorithms to find the global optimum (potentially requiring an infinite number of steps), it has been shown that they are not applicable without tuning in the setting of signaling pathway models.[25] Instead, the applicability of deterministic optimization in combination with a multistart approach using a sufficient number of different initial guesses has been proven in several applications.[27–30] This strategy is implemented in the Data2Dynamics modeling environment[20,21] and has been awarded in scientific benchmark challenges in the setting of parameter estimation[22] as well as for network reconstruction[23] and is applied in this article. Here we use a Gauss–Newton gradient-based trust region optimizer as implemented as function `lsqnonlin` in the optimization toolbox of MATLAB[31] in combination with the CVODES initial value problem solver[32] which is implemented in C. Gradient information of the likelihood with respect to the parameters is calculated by solving the so-called sensitivity equations

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}\boldsymbol{\theta}} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}} + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}} \tag{6}$$

Because the partial derivatives $\partial \boldsymbol{f}/\partial \boldsymbol{x}$ and $\partial \boldsymbol{f}/\partial \boldsymbol{\theta}$ can be calculated analytically, the parameter sensitivities $\partial \boldsymbol{x}/\partial \boldsymbol{\theta}$ occur as solutions of another ODE system of dimension $\dim(\boldsymbol{\theta}) \times \dim(\boldsymbol{x})$ which can be attached to the original ODE system (equation (2)) and then numerically integrated simultaneously.[33]

2.5 Profile likelihood

Traditionally, confidence intervals for point estimates of parameters are calculated from the Fisher information which is derived from second derivatives of the log-likelihood around the maximum. This approach yields exact results for linear models but does not account for higher order derivatives as they occur in the nonlinear setting. In contrast, the profile likelihood approach comprehensively evaluates the shape of the log-likelihood enabling the calculation of reliable confidence intervals of the estimated parameters also for nonlinear settings.[16,18,34–36]

To calculate the profile likelihood, each parameter θ_i is iteratively changed around its maximum likelihood estimate while reoptimizing all other parameters $\theta_{i\neq j}$. Thus, the *profile likelihood* is defined as

$$PL_i(\boldsymbol{p}) = \min_{\{\boldsymbol{\theta}|\theta_i=p\}} -2LL(\boldsymbol{\theta}) \tag{7}$$

The uncertainty of a parameter can be assessed by the profile likelihood-based confidence interval

$$CI_{PL}(\theta_i) = \{p \mid PL_i(\boldsymbol{p}) \leq -2LL(\hat{\boldsymbol{\theta}}) + \Delta_\alpha(\chi^2_1)\} \tag{8}$$

which represents all acceptable parameter values θ_i with PL_i below the threshold defined by the α quantile $\Delta_\alpha(\chi^2_1)$ of the χ^2_1 statistics.[37]

Optimization of equation (7) for each i and $\boldsymbol{p}$ yields parameter vectors which are suited as representative candidates for sampling the dynamics within the high-dimensional confidence region given by the $\Delta\chi^2_1$ threshold. It has been shown that using these representative parameters enables evaluation of prediction uncertainties[17] and the identification of informative experimental designs.[22]

2.6 Identifiability analysis

Identifiability of model parameters related to the model structure and the experimental data has been extensively discussed in the literature.[38–44] Here, we summarize and apply a numerical approach which only requires numerical implementation of parameter optimization but has no limitations in terms of model size, model structure or mathematical prerequisites.

The approach based on the profile likelihood allows for a data-based identifiability analysis of the estimated parameters,[16,17] where three classifications are available. A parameter is considered as *identifiable* if its likelihood profile exceeds the statistical threshold given by $-2LL(\hat{\boldsymbol{\theta}}) + \Delta_\alpha(\chi^2_1)$, yielding finite confidence intervals. If the amount or quality of the available data is not sufficient to indicate a confidence bound in both directions, the parameter is termed as *practically non-identifiable*. Here, a unique point estimate may be obtained, but the profile does not exceed the statistical threshold in at least one direction and thus does not have a confined confidence interval.

For *structural non-identifiability*, there are several definitions in the literature. In the context of continuous noise-free observations, a model is commonly defined as structurally non-identifiable if for $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, equality of the model equations $\boldsymbol{x}(t,\boldsymbol{\theta}) = \boldsymbol{x}(t,\boldsymbol{\theta}')$ and $\boldsymbol{g}(t,\boldsymbol{\theta}) = \boldsymbol{g}(t,\boldsymbol{\theta}')$ for all times t does not imply equality of the two parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$.[41] Instead, we use the conceptually different definition from Raue et al.,[16] which is more appropriate for a finite number of time points with observation errors. Here, parameters are already termed as structurally non-identifiable if the estimated parameters for a given data set are not unique. This means that there are several maximum likelihood estimates for given model outputs $\boldsymbol{g}(\boldsymbol{x}_j)$ and observation times t_i. Except for special cases, this means that the respective non-identifiable parameters cannot be constrained at all and the profile is constant for all values of the parameter. Such a behavior indicates redundant parametrization of the model structure which cannot be resolved by available observations. Consequently, the individual influence of a structurally non-identifiable parameter on the experimentally observed model outputs can be compensated by re-optimization of the other parameters. These redundancies may be obvious in some cases, e.g. if two parameters only appear as a product. Nontrivial structural non-identifiability can be identified by Lie group theory.[45] Identifiability analysis may also be applicable when using parameter priors via the penalized log-likelihood (equation (5)).

2.7 Model reduction

As every model is a simplification of complex natural phenomena, where some features may be over-represented while others are neglected, it can be reasonable to adjust the complexity of the model to the available information

in experimental data. Too complex models are more likely to overfit measurement errors and then providing inaccurate predictions. The model reduction strategy discussed here utilizes the profile likelihood-based identifiability analysis for indicating simplification which is not rejected statistically by available data.

For *structural* non-identifiabilties, re-parameterization based on the underlying symmetry transformation, e.g. by fixing the parameter to a reasonable, user-defined value yields the desired outcome. For *practical* non-identifiabilties, the profile likelihood can be utilized as a data-based method for reducing nonlinear models[19] since the profile likelihood calculation drives the model parameters towards their extreme values, i.e. minimal and maximal values, while keeping the model's output in agreement with the data. Furthermore, the definition of the profile likelihood PL_i (equation (7)) coincides with the test statistic of the likelihood ratio test[46] which is a well-established statistical test for assessing whether reducing the model is rejected by the data. A parameter profile which flattens out to a constant level below the threshold for the limit of small values of the parameter close to zero, points to a simplification where the parameter value is set to zero, i.e. the corresponding reaction is removed from the model. Iterative application of such a procedure enables elimination of all practically non-identifiable parameters, yielding a desirable fully identifiable model.[19] As in every step-wise model selection procedure, the order of the reduction steps might have an impact on the result. Solving such ambiguities depends on the questions which have to be answered by the model and are beyond the scope of this paper.

3 Influenza model

The applicability of the introduced methodology is demonstrated in the following. Therefore, it is applied to the frequently analyzed standard data example of an influenza outbreak at an English boarding school using a basic *SIR* model.[47–51]

3.1 Model equations

The so-called *SIR* model

$$S \xrightarrow{\beta I} I \xrightarrow{\gamma} R \tag{9}$$

was first introduced by Kermack and McKendrick[52] and consists of three compartments: susceptible individuals S which are not yet infected, infected individuals I showing symptoms and are able to transmit the disease, as well as recovered individuals R which, after the infection, are assumed to be immune to reinfection. The change of the population numbers in the compartments is described by the model equations

$$\begin{aligned} \dot{S} &= -\frac{\beta SI}{N} \\ \dot{I} &= \frac{\beta SI}{N} - \gamma I \\ \dot{R} &= \gamma I \\ N &= S + I + R \end{aligned} \tag{10}$$

where N is the total number of individuals. Susceptibles S are infected by infected I with transmission rate β and velocity of transmission from the infected I to recovered R is described by rate constant γ. Both rate parameters have the physical unit $1/time$. The initial value of recovered individuals R^{init} is set to zero, whereas the initial values $S^{init} = N_S$, which represents the total number of humans which can be infected and $I^{init} = N_I$, which is the number of initially infected humans, i.e. the source of infection, are considered as parameters to be estimated and have the unit *number of humans*.

Although analytical solutions are available for the simple *SIR* model, we use numerical methods to integrate the ODEs to not restrict applicability to analytically solvable special cases.

3.2 Data

The data set from an influenza outbreak at an English boarding school in 1978[47,50,51] contains a total of 763 boys. As the school administration kept track about daily numbers of boys staying at bed, a complete, direct, and time-resolved record of the model compartment of infected individuals I is available. One observation per day

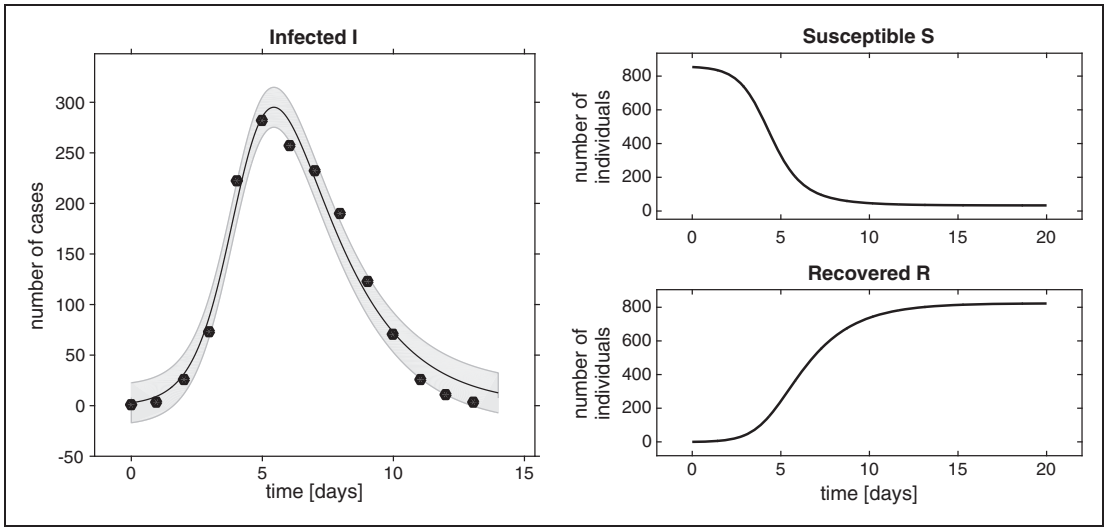

Figure 1. Model dynamics of the three compartments for the best fit parameters provided in Table 1 with data for infected individuals. Gray bands represent estimated observation error.

is available. As the number of infected boys is directly observed, the observation function is the identity and the model is linked to the data via

$$y_{obs}(t_i) = I + \varepsilon(t_i) \tag{11}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2_{\text{abs}})$ with a priori unknown standard deviation σ_{abs} which is treated as an additional parameter to be estimated.

The total number of susceptible individuals $N_S^\dagger = 762$ was published with the data and a single infected individual $N_I^\dagger = 1$ has been assumed as starting point of the epidemic. However, this information will not be included into the presented analysis, but in contrast is predicted from the model only by the time-resolved empirical data of infected individuals I.

3.3 Parameter estimation

Figure 1 shows the data as well as the best fit and the respective trajectories for the compartments of the susceptibles and the recovered.

The model was fitted by optimization of the likelihood from 100 different initial guesses leading to a clearly identifiable global optimum. The uniformly distributed initial guesses for optimization are drawn randomly from the parameter search space which spans nine orders of magnitudes from 10^{-5} to 10^4 in every dimension. Figure 2 shows the likelihood of the initial guess as well as after optimization. The optimization runs are ordered by the final likelihood value. Sixty-seven of the 100 fits converge to the same optimum which is indicated by the same final value of the objective function and, as shown in the lower panel, by the same parameter values (up to numerical tolerances). All other fits start from regions in the parameters space from which the deterministic optimizer is not able to locate the global optimum but, instead, converges to local optima with substantial inferior likelihood values. The lower panel of Figure 2 indicates that for the suboptimal local optima, some parameters values like γ and σ_{abs} exhibit patterns, while others like N_S, N_I and β seem to be rather randomly distributed. The parameter values of the best fit are provided in Table 1.

3.4 Profile likelihood and identifiability analysis

Once the global optimum is found, likelihood profiles according to equation (7) are calculated to derive confidence intervals for the estimated parameters. The resulting likelihood profiles shown in Figure 3 exceed the threshold for 95% confidence intervals indicating identifiability of all parameters in this example.

The literature values of susceptible individuals $N_S^\dagger = 762$ and the infected individual $N_I^\dagger = 1$ meet the estimated values for parameters N_S and N_I as they lie within their 95% profile likelihood-based confidence intervals, c.f. Table 1.

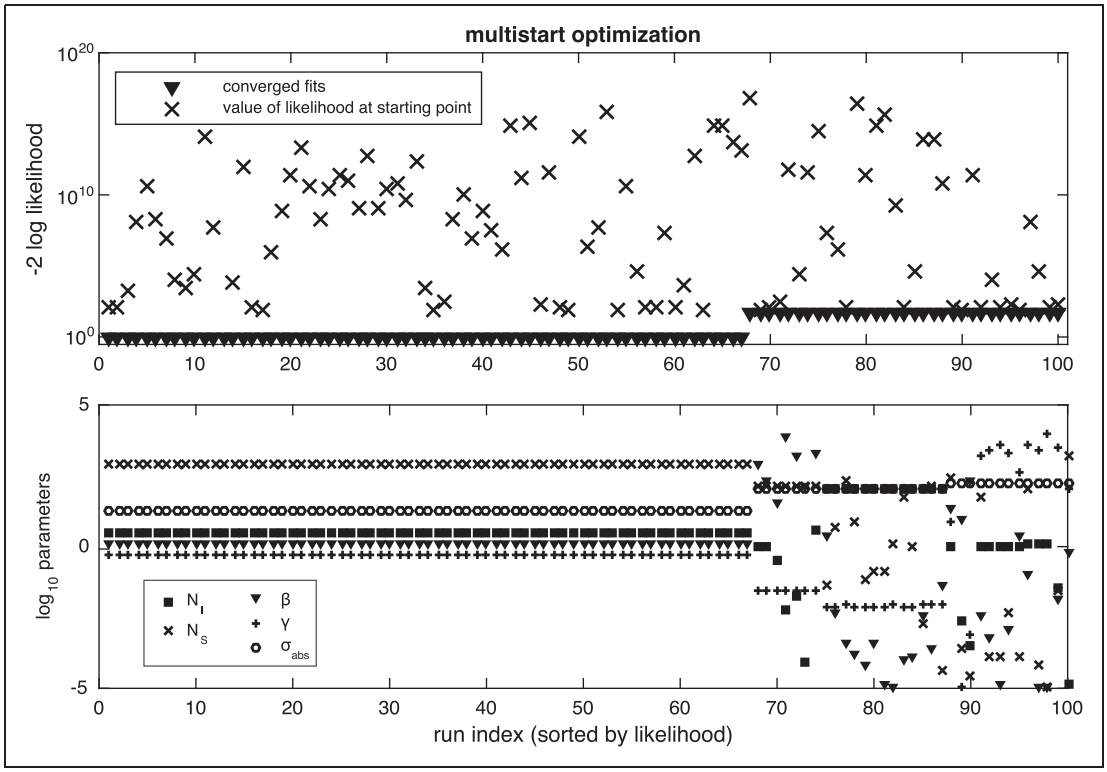

Figure 2. Results from multistart optimization with 100 initial guesses. The initial likelihood values are plotted as crosses, the end-points after the optimization are depicted as triangles. For illustration, likelihood values are shifted to the baseline by subtraction of the likelihood value for the global minimum. The lower panel shows the estimated parameter values of the respective optimization runs. The 67 best performing optimization runs converge not only to the same likelihood value, but also to the same global optimum in the parameter space, as indicated by the horizontal pattern of the parameter values. In contrast, the other optimization runs converting to the suboptimal minima share the same final likelihood value, but converge to different regions in the parameter space. Some of these regions are characterized by similar end-points for some parameters, while other parameters spread over the whole parameter search space.

Table 1. Best fit parameter values as well as confidence intervals for the *SIR* model for the English boarding school influenza data.

Parameter	Physical unit	Parameter search region	Estimated parameter value	Profile likelihood-based confidence interval
N_I	*humans*	10^{-5}–10^{4}	2.89	[0.54–8.88]
N_S	*humans*	10^{-5}–10^{4}	854	[565–2324]
β	$1/t$	10^{-5}–10^{4}	1.74	[1.57–2.06]
γ	$1/t$	10^{-5}–10^{4}	0.51	[0.34–1.10]
σ_{abs}	*humans*	10^{-5}–10^{4}	19.7	[14.2–30.1]

Each point on the profile likelihood corresponds to an individual parameter vector obtained by fixing one parameter to a specific value and reoptimizing the other parameters. Within the confidence intervals, the objective function of these parameter vectors does not exceed the threshold given by the selected significance level. Therefore, all the respective trajectories shown in Figure 4 are statistically in accordance with the examined data. For observed quantities, the trajectories differ only slightly from observations. In contrast, predictions for unobserved quantities can have large uncertainties indicated by a large spread of the trajectories. Figure 4(b) shows prediction trajectories for the number of susceptibles S and recovered R individuals as well as for the two transition fluxes (Figure 4(c)). The trajectories for the suceptibles and recovered show that the measurements for the infected individuals alone do not provide accurate information about the total number of individuals at risk N_S.

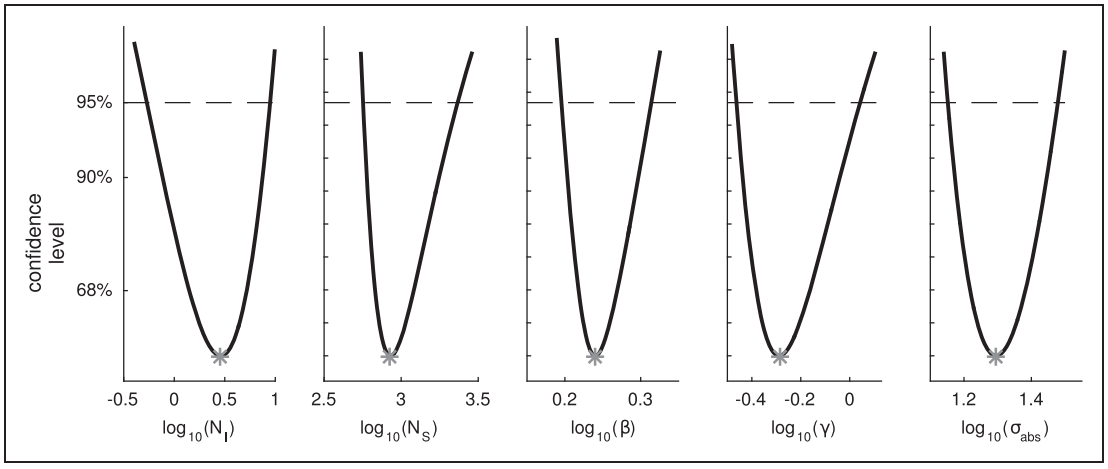

Figure 3. Likelihood profiles for all estimated parameters (solid lines). Best fit parameter values are indicated by gray asterisk. Parameter values at which the profile hits the $\Delta_\alpha(\chi^2_1)$ threshold (dashed line) are identifiable with lower and upper confidence bounds.

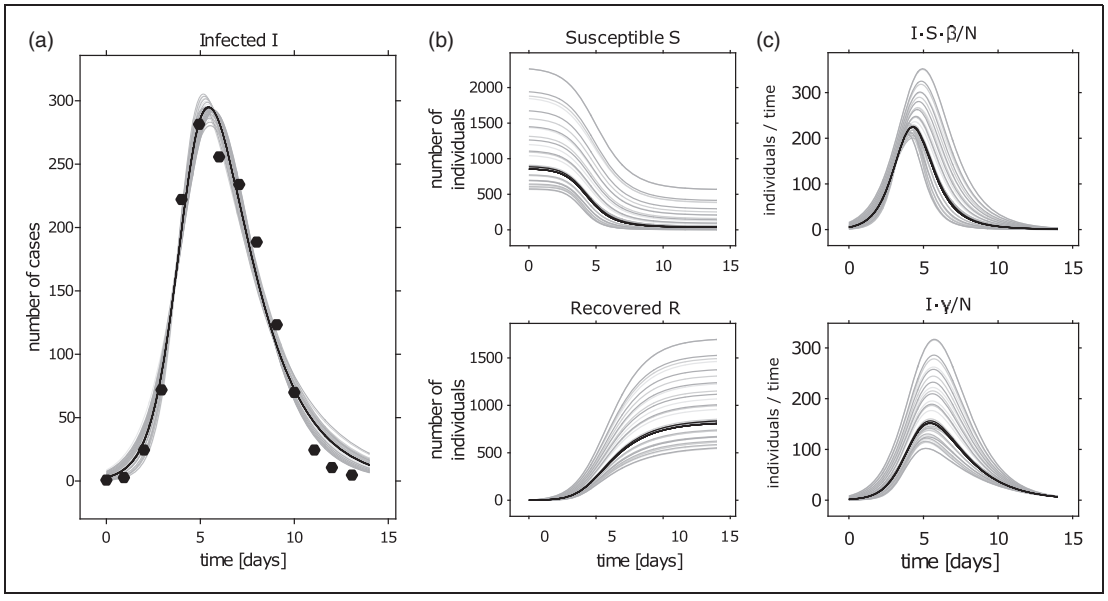

Figure 4. Model trajectories for the set of representative parameters vectors obtained by the profile likelihood calculation. (a) Trajectories for observed susceptible humans, (b) for unobserved model compartments, and (c) for transition fluxes between compartments. Gray lines indicate trajectories of parameter sets along the likelihood profile, the black circles in (a) show the data and the black solid line refers to the trajectories of the best fit parameters.

3.5 Conclusions

The SIR model can describe the data of infected school boys, using the best fit parameters from a deterministic multistart optimization sequence. Likelihood profiles indicate a fully identifiable model and yield finite confidence intervals for all estimated parameters. Furthermore, trajectories for unobserved compartments can be used to predict their dynamics.

As the used data set is a standard basic example in the literature, the estimated parameter values can be compared to earlier analyses. Our estimates parameters shown in Table 1 are close to the reference values from literature[48–51] which are in agreement with the profile likelihood-based confidence intervals.

In the case of the simple SIR model, the basic reproduction number R_0, i.e. the number of secondary cases from a primary infected individual[53] can be easily computed from the estimates as

$$R_0 = \frac{\beta}{\gamma} = 3.4 \tag{12}$$

which is close to values estimated from literature for the same data and model. The estimated values of R_0 for the English boarding school data set is slightly larger than assumed for influenza in general and the estimated infection period is slightly shorter than expected.[54,55] However, this can be explained by the ideal circumstances for an epidemic because of higher infection rates for school-aged individuals[56] and the well-mixing of the individuals in a boarding school in contrast to other observed influenza outbreaks, e.g. in larger cities or regions.

Although a priori information for the estimated parameters would be available and could be integrated, it was intentionally not included into the analysis. For instance, the number of susceptibles N_S was published with the data and there are also other data sources which would, e.g., provide detailed information on infection period durations for the specific influenza type. Nevertheless, published values from earlier analyses for single parameters as well as for the basic reproduction number R_0 are in accordance with the profile likelihood-based confidence intervals.[48–51]

4 Zika virus disease model

In the following, the presented methodology is applied to empirical data from the Zika virus disease outbreak in Colombia in 2015–2016 using a complex vector-borne disease model.

The Zika virus disease is caused by the Zika flavivirus (ZIKV) and is transmitted mainly by bites of *Aedes aegypti* mosquitos.[57] Typical acute symptoms include fever, maculopapular skin rashes, conjunctivitis, retro-orbital pain and headache. Humans overcome ZIKV infection usually after days to weeks and do not have long-term effects.[58] However, a causal link to microcephaly[59] and Guillain-Barré syndrome[60] has been reported.[61] The latest Zika virus disease outbreak in the Americas in 2015 thus became of global interest and was declared as an emergency of international concern by the World Health Organisation (WHO).[62]

In order to describe and possibly intervene the spread of the disease, many different models were discussed in the literature. Mainly models adapted from other mosquito-borne diseases such as malaria,[63,64] dengue fever or chikungunya virus (CHIKV) have been adapted in order to analyze the dynamics of ZIKV outbreaks.[65–67] Amongst others, the motif of human-to-human infection is discussed and integrated into the models in order to analyze its influence on the epidemic. Presumingly, because of the global public interest and as publicly accessible empirical data of ZIKV infected humans is available from the national health institutes of most affected countries, dynamical ZIKV infectious models became a rapidly growing research interest.

4.1 Model

To describe the ZIKV transmission via humans and mosquitos, we use a model, which includes several of the discussed features of mosquito-borne infection resulting in a complex ZIKV model as shown in Figure 5(a). The core of vector-borne disease models is a so-called *SEIR* submodel for humans combined with an *SEI* submodel for mosquitos: It is commonly assumed that a human individual that is infected with the Zika virus belongs in a so-called exposed (E) compartment for the incubation time of a couple of days[68,58] before the virus can be transmitted to other individuals such as mosquitos or humans. Thus, the classical *SIR* model is extended by an intermediate compartment E, leading to the so-called *SEIR* model. Analogously, an exposed compartment is assumed within the propagation of the disease in mosquitos. However, since the lifespan of mosquitos is short compared to typical timescales of recovering from a Zika virus infection, there is no recovered compartment for mosquitos, resulting in the mentioned *SEI* model. A mosquito may progress from susceptible S_v to exposed E_v when infected by an infected human with transmission rate β_{hv} and converts to infected I_v with rate ν_v afterwards.

Additionally to the classical *SEIR-SEI* model, the following extensions are introduced into the model. Birth of susceptible mosquitos is assumed to be relative to the total number $N_v = S_v + E_v + I_v$ with rate μ_v. Since global changes of the total mosquito population should not be covered by the model, death with the same rate μ_v is assumed to all three mosquito compartments.[65,69] When biting susceptible humans S_h, infected mosquitos I_v transmit the virus with rate β_{vh}.

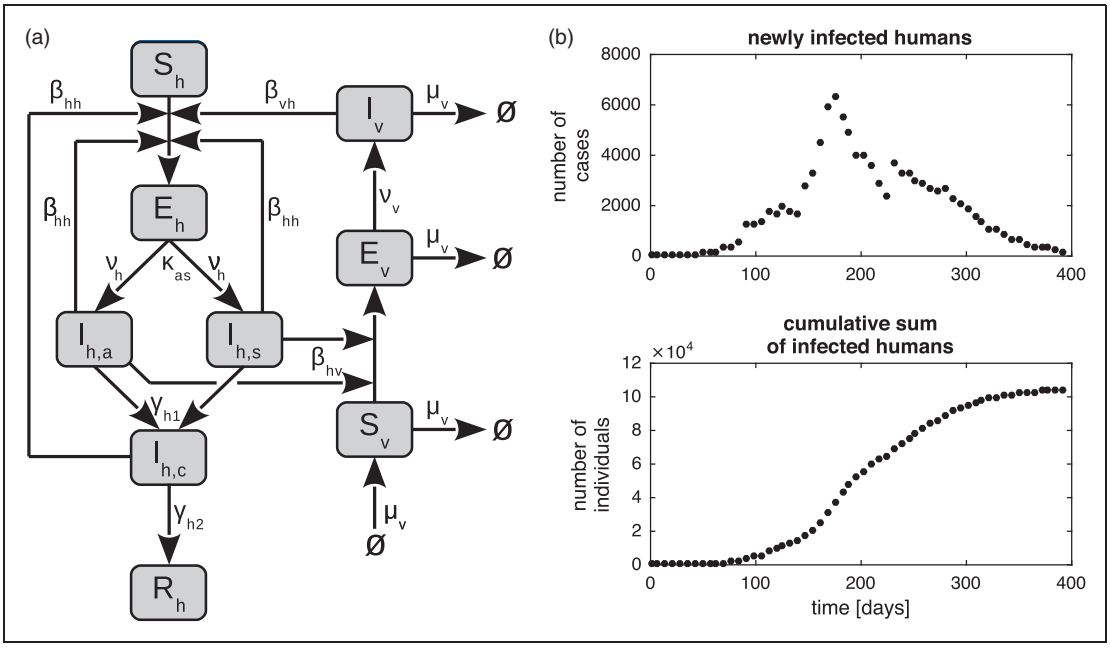

Figure 5. (a) Model structure of the full ZIKV infection model as described by equation (13). Mosquitos progress from susceptible S_v to exposed E_v when infected by a contagious human and convert to infected I_v with rate ν_v. Birth and death of mosquitos are described by rate μ_v. When bitten by an infected mosquito, susceptible humans S_h progress with rate β_{vh} either to asymptomatically infected $I_{h,a}$ or symptomatically infected $I_{h,s}$, before they convert to convalescent $I_{h,c}$ with rate γ_{h1} and finally to recovered R_h with rate γ_{h2}. The proportion between symptomatically and asymptomatically infected is controlled by κ_{as}. All infected humans can transmit the virus to susceptible humans with rate β_{hh}, whereas only symptomatically and asymptomatically infected humans infect susceptible mosquitos with transmission rate β_{hv}. Only the cumulative incidence of symptomatically infected humans is observed (c.f. equation (14)). (b) Data of newly infected humans[76] and calculated cumulative sum of infected humans.

For the ZIKV there is a considerable amount of asymptomatically infected humans of approximately 80% reported,[58] which do not show any symptoms but are able to spread the disease.[70] This is included in the model by splitting up the compartment of infected humans I_h to asymptomatically infected $I_{h,a}$ and symptomatically infected $I_{h,s}$ humans. The proportion of exposed humans becoming asymptomatically infected is represented by the parameter κ_{as}. Since this parameter is of great interest in practice and only a rough estimate is available a priori, κ_{as} is estimated simultaneously with all other model parameters.

As suggested in Gao et al.,[68] compartments $I_{h,s}$ and $I_{h,a}$ both transmit to an additional compartment of so-called convalescent humans $I_{h,c}$ with rate γ_{h1} before they proceed with rate γ_{h2} to the compartment of recovered humans R_h.[68] Sexually transmitted direct human-to-human infection has been reported[71,72] and is incorporated into the model by human-to-human transmission rate β_{hh}.[68,73] It has been shown that the virus persists longer in semen than in serum.[74,75] Thus, besides asymptomatically infected $I_{h,a}$ and symptomatically infected $I_{h,s}$ humans, in the model also convalescent humans $I_{h,c}$ are able to transmit the virus to susceptible humans, whereas mosquitos cannot be infected by biting convalescent humans.[68]

Since virus transmission rates are assumed to be orders of magnitudes faster than human birth or death rates, they are completely neglected in the model. The initial value of convalescent humans is $I_{h,c}^{init} = 0$, as well as for recovered humans $R_h^{init} = 0$. All other initial values are estimated comprehensively with the model parameters. The model equations for the described model read

$$\begin{aligned}
\dot{S}_v &= \mu_v N_v - \frac{\beta_{hv} S_v (I_{h,a} + I_{h,s})}{N_h} - \mu_v S_v \\
\dot{E}_v &= \frac{\beta_{hv} S_v (I_{h,a} + I_{h,s})}{N_h} - \nu_v E_v - \mu_v E_v \\
\dot{I}_v &= \nu_v E_v - \mu_v I_v
\end{aligned}$$

$$
\begin{aligned}
\dot{S}_h &= -\frac{\beta_{vh} S_h I_v}{N_v} - \frac{\beta_{hh} S_h (I_{h,a} + I_{h,s} + I_{h,c})}{N_h} \\
\dot{E}_h &= \frac{\beta_{vh} S_h I_v}{N_v} + \frac{\beta_{hh} S_h (I_{h,a} + I_{h,s} + I_{h,c})}{N_h} - \nu_h E_h \\
\dot{I}_{h,a} &= \kappa_{as} \nu_h E_h - \gamma_{h1} I_{h,a} \\
\dot{I}_{h,s} &= (1 - \kappa_{as}) \nu_h E_h - \gamma_{h1} I_{h,s} \\
\dot{I}_{h,c} &= \gamma_{h1} (I_{h,a} + I_{h,s}) - \gamma_{h2} I_{h,c} \\
\dot{R}_h &= \gamma_{h2} I_{h,c} \\
N_v &= S_v + E_v + I_v \\
N_h &= S_h + E_h + I_{h,a} + I_{h,s} + I_{h,c} + R_h
\end{aligned} \tag{13}
$$

The unit of the rates $\beta_{hh}, \beta_{vh}, \beta_{hv}, \nu_v, \nu_h, \mu_v, \gamma_{h1}$, and γ_{h2} is the inverse of the time t^{-1}, for the initial values $S_h^{init}, E_h^{init}, I_{h,s}^{init}, I_{h,a}^{init}$, and $I_{h,c}^{init}$ the unit is *number of humans*, for initial values S_v^{init}, E_v^{init}, and I_v^{init} it is *number of mosquitos* and κ_{as} is dimensionless.

4.2 Data

Empirical data of newly infected humans is published by Colombia's National Health Institute on a weekly basis. These numbers contain also delayed reported cases from earlier weeks. As numbers of delayed reported cases are in the same order of magnitude as the current new infections, they should not be neglected. Hence, being the only continuously updated data source, graphical data from the weekly reports had to be used.[76] For reliably extracting the approximate number of reported cases, we used the WebDigitizer tool.[77] The extracted data is depicted in Figure 5(b).

In contrast to the first example, the current number of infected humans is not observed directly. Instead, the weekly number of newly infected individuals is recorded. Thus, the cumulative sum of newly infected is calculated and compared to the model, although fitting of cumulative incidence data can cause issues concerning underestimation of confidence intervals sizes.[36] However, identifiability analyses and model reduction methods should not be affected. The cumulative sum of symptomatically infected humans is assessed in the model via the integral over the influx to $I_{h,s}$ so that the observation function reads

$$
y_{obs}(t_i) = x_{obs} + \varepsilon(t_i) = \int (1 - \kappa_{as}) \nu_h E_h \mathrm{d}t + \varepsilon(t_i) \tag{14}
$$

An error model is chosen for which $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, with

$$
\sigma^2 = \sigma_{\text{rel}}^2 x_{obs}^2 + \sigma_{\text{abs}}^2 \tag{15}
$$

with relative σ_{rel}^2 and absolute σ_{abs}^2 error parameters, which are treated as additional parameters to be comprehensively estimated with the other model parameters. In the presented approach, only information contained in the data is utilized to estimate the parameters and to assess their uncertainties, so that no prior knowledge from the literature, e.g., for transmission rates or average duration times of infection states will be integrated into the analysis.

4.3 Parameter estimation

In order to obtain parameter estimates using the introduced model and data, the multistart fitting procedure with 1000 initial guesses was performed. In addition to several suboptimal local minima, the global optimum according

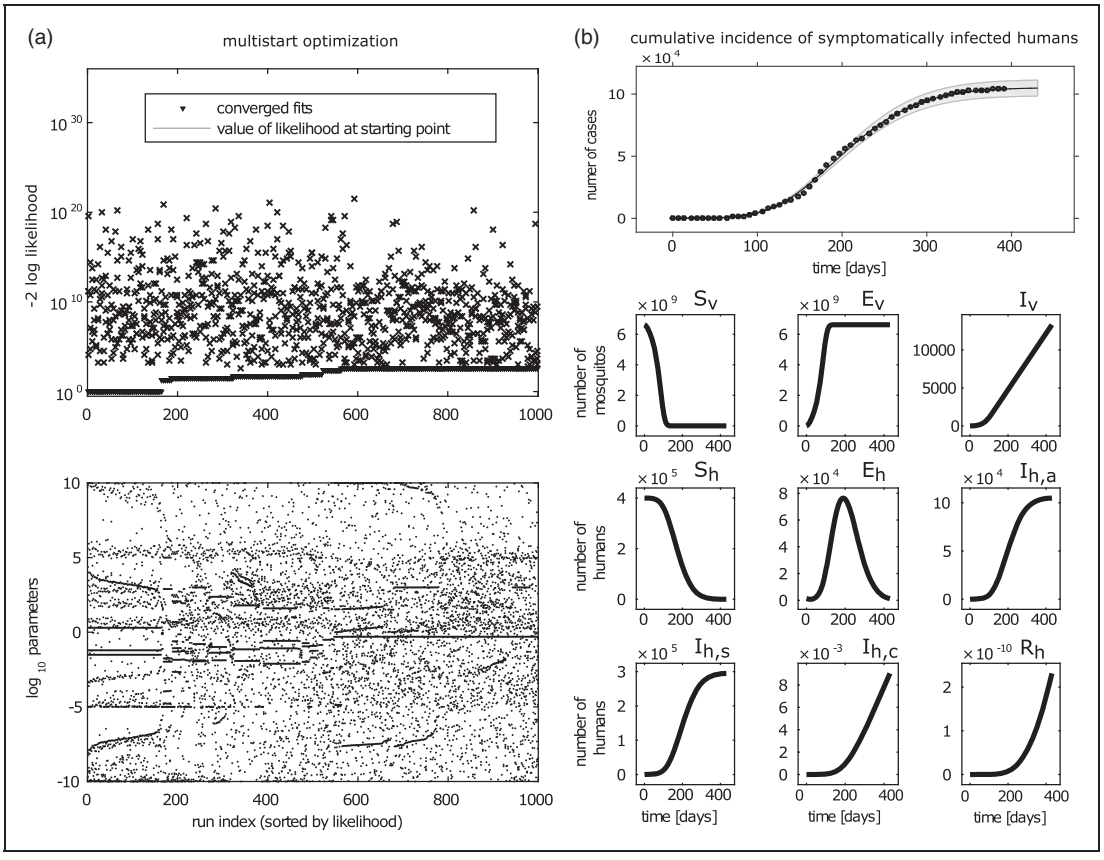

Figure 6. (a) Multistart optimization with 1000 fits for the full ZIKV infection model (equation (13)) with optimization end-point parameters. For illustration, likelihood values are shifted to the baseline by subtraction of the likelihood value for the global minimum. Search regions for the estimated parameters are: initial values $\in [10^{-5}, 10^{10}]$, kinetic parameters $\in [10^{-10}, 10^{5}]$, $\sigma_{\text{abs}} \in [10^{-5}, 10^{3}]$, $\sigma_{\text{rel}} \in [10^{-5}, 10^{-0.3}]$, $\kappa_{as} \in [0, 1]$. Computation time for 1000 fits was 56.9 min, i.e. approximately 3.4 s for an average single fit on a 3.4 GHz quad-core CPU. (b) Best fit of observation function with data and model dynamics of compartments for the estimated parameters.

to minimal objective function could be identified in 163 runs (Figure 6(a)). Using the best fit parameters, the model is able to describe the data adequately, as shown in Figure 6(b).

Inspection of the likelihood profiles indicates that the estimated parameters have large uncertainties, as many profiles are flat and reveal non-identifiabilities (Figure 7). As a consequence of structural non-identifiability of parameters γ_{h2} and κ_{as} and the practical non-identifiability of $\beta_{hh}, \beta_{vh}, \gamma_{h1}, E_v^{init} I_v^{init}, S_h^{init}, S_v^{init}, \mu_v$, and σ_{abs}, the estimated values within the global minimum are not unique as depicted in the lower panel of Figure 6(a). Even for the global likelihood optimum, there are very few horizontal structures in the parameter plot indicating many flat regions in the parameters space. Flat likelihood profiles shown in Figure 7 indicate large ranges of parameter values which are in accordance with the data. Therefore, the mere estimated values and the corresponding trajectories are of limited impact without assessing uncertainties. Since uncertainties of parameters translate to uncertainties of the compartment trajectories, the quality of the drawn conclusions is restricted.

4.4 Model reduction

In several circumstances, it is reasonable to reduce the model to a level of complexity which is identifiable for given data. In the following, model reduction based on the likelihood profiles is exemplified. Since a reduction step might have an impact on all parameters, the likelihood profiles have to be calculated and checked after each reduction step. Figure 8 shows exemplary profiles of intermediate steps. All model reduction steps are summarized in Table 2.

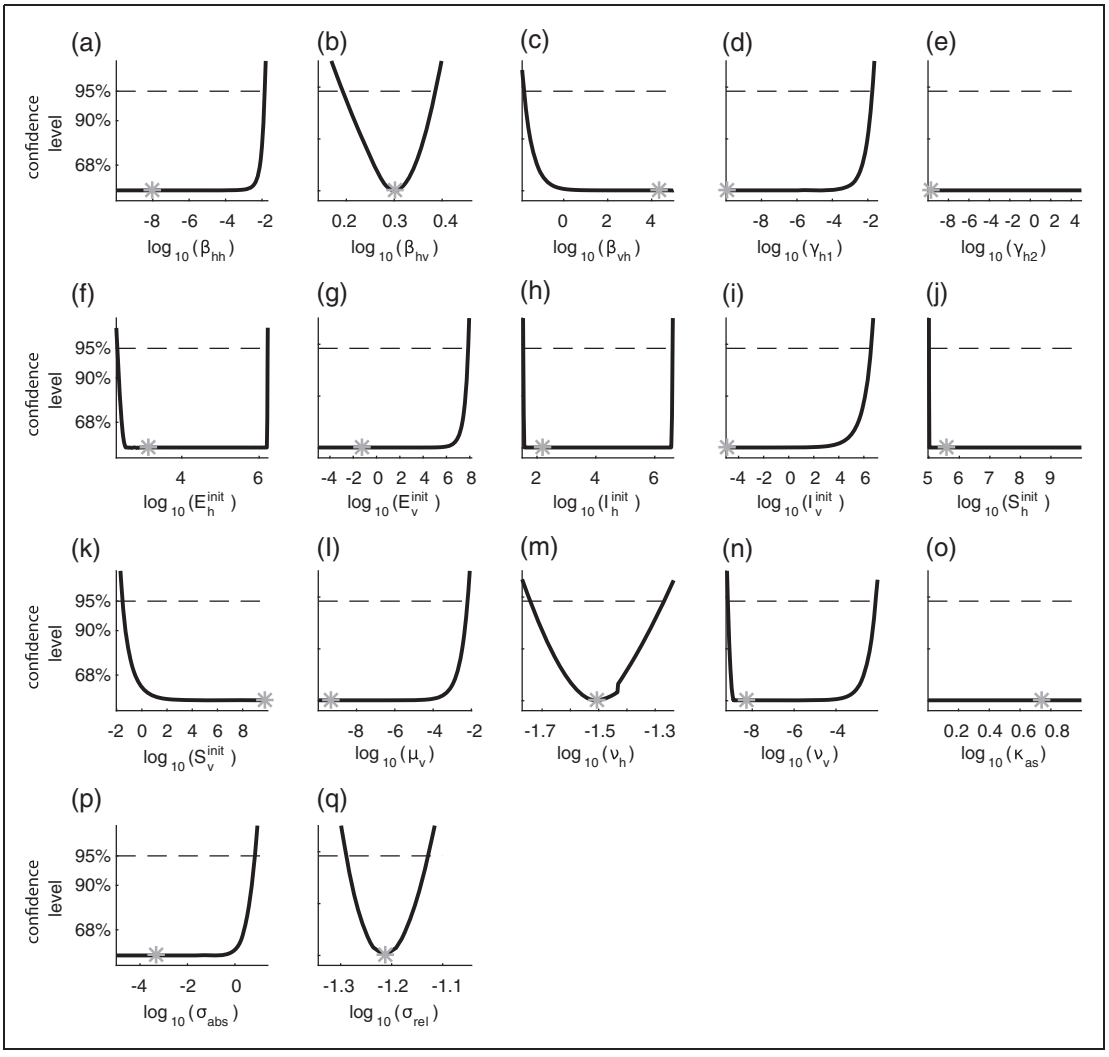

Figure 7. Likelihood profiles for the full Zika model (equation (13)). Flat profiles of parameters γ_{h2} (panel e) and κ_{as} (panel o) reveal structural non-identifiabilities, whereas the profiles of β_{hh}, β_{vh}, γ_{h1}, E_v^{init}, I_v^{init}, S_h^{init}, S_v^{init}, μ_v and σ_{abs} exhibit practical non-identifiabilities. Computation time was approximately 40.4 min on a 3.4 GHz quad-core CPU.

In a first instance, the profile of the initial value parameter S_v^{init} (Figure 7(k)) of susceptible mosquitos is investigated. It is only bounded to lower values as the likelihood profile is flat towards extreme large values. Also both remaining initial values of the mosquito population E_v^{init} and I_v^{init} (Figure 7(g) and (i)) exhibit practical non-identifiability but with profiles being flat to minus infinity. Obviously, the data of the infected humans does not provide information about the upper bound of the mosquito population size nor about the lower bound for the exposed and infected mosquitos. The only possibility to remove this redundancy is to use prior knowledge about the ratio of mosquitos to humans in the initial states at time $t=0$. The parameter κ_{hv} is therefore introduced by

$$\begin{aligned} S_v^{init} &= \kappa_{hv} S_h^{init} \\ E_v^{init} &= \kappa_{hv} E_h^{init} \\ I_v^{init} &= \kappa_{hv} I_h^{init} \end{aligned} \tag{16}$$

with prior value $\bar{\kappa}_{hv} = 5$ and prior uncertainty $\bar{\sigma}_{\kappa_{hv}} = 5$ as suggested in the literature.[78,79] This first step introduces a link between the total population size of humans and mosquitos in order to facilitate the further investigation and is the only additional prior information included into the analysis.

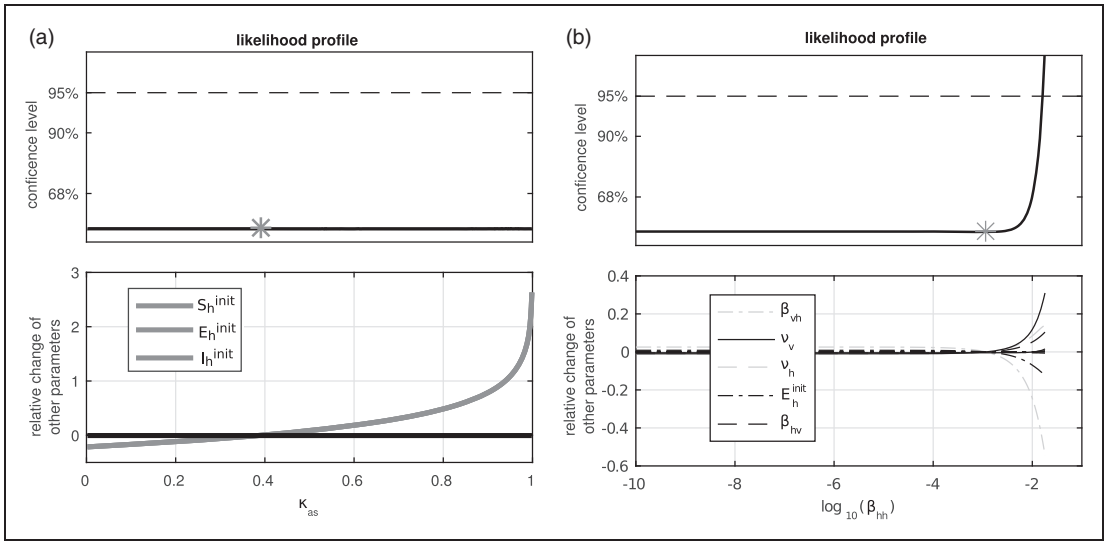

Figure 8. (a) Likelihood profile of κ_{as} at the intermediate model reduction step 6. Upper panel shows the flat likelihood profile, lower panel shows the trace of the reoptimized parameters relative to their maximum likelihood estimate. Parameters S_h^{init}, E_h^{init} and I_h^{init} are changed in parallel along the profile, as indicated by gray solid lines (all lines overlap), whereas all others parameters stay constant (black solid line). (b) Likelihood profile and relative change of reoptimized parameters for β_{hh} at model reduction step 7. Only the five parameters with the largest relative change during reoptimization are labeled in the lower panel.

Table 2. Model reduction steps from the full Zika model (equation (13)) to the reduced Zika model (equation (17)).

Step	Reduction	Reasoning
1	introduction of κ_{as}	profiles E_v^{init} and I_v^{init} flat to $-\infty$, S_v^{init} flat to $+\infty$
2	$\sigma_{\text{abs}} = 0$	profile flat to $-\infty$
3	$\mu_v = 0$	profile flat to $-\infty$
4	$\gamma_{h2} = 0$	structural non-identifiable
5	$\gamma_{h1} = 0$	profile flat to $-\infty$
6	$\kappa_{as} = 0.8$	structural non-identifiable
7	$\beta_{hh} = 0$	profile flat to $-\infty$

As described in Section 2.7, non-identifiable parameters with likelihood profiles being flat to minus infinity allow for reduction in a way that parameters are set to zero. Consequently, a purely relative error model (c.f. equation (15)) is indicated by the likelihood profile (Figure 7(p)) of σ_{abs} (reduction step 2). Likewise, μ_v (Figure 7(l)) can bet set to zero and so the mosquito's birth and death dynamics are removed from the model, as they cannot be constrained to lower values by the analyzed data (reduction step 3).

The practical non-identifiable parameter γ_{h1} and the structural non-identifiable parameter γ_{h2} (Figure 7(d) and (e)) describe the progression of infected humans $I_{h,a}$ and $I_{h,s}$ via the convalescent state $I_{h,c}$ into the recovered compartment R_h. As the human-to-human infection rate β_{hh} (Figure 7(a)) is also in accordance with zero, no interaction of the dead end $\rightarrow I_{h,c} \rightarrow R_h$ with the rest of the model is left. Thus, the distribution of individuals and their progression within this submodule cannot be resolved by the data. In order to resolve this over-complexity of the model, the structural non-identifiable parameter γ_{h2} can be set to an arbitrary value (reduction step 4) and γ_{h1} can be fixed to zero due to the flat likelihood profile towards minus infinity (reduction step 5). Setting $\gamma_{h1} = \gamma_{h2} = 0$ is equivalent with the introduction of a new state $I_{h,scR}$, which combines the three states $I_{h,s}$, $I_{h,c}$ and R_h, as final compartment for symptomatically infected humans.

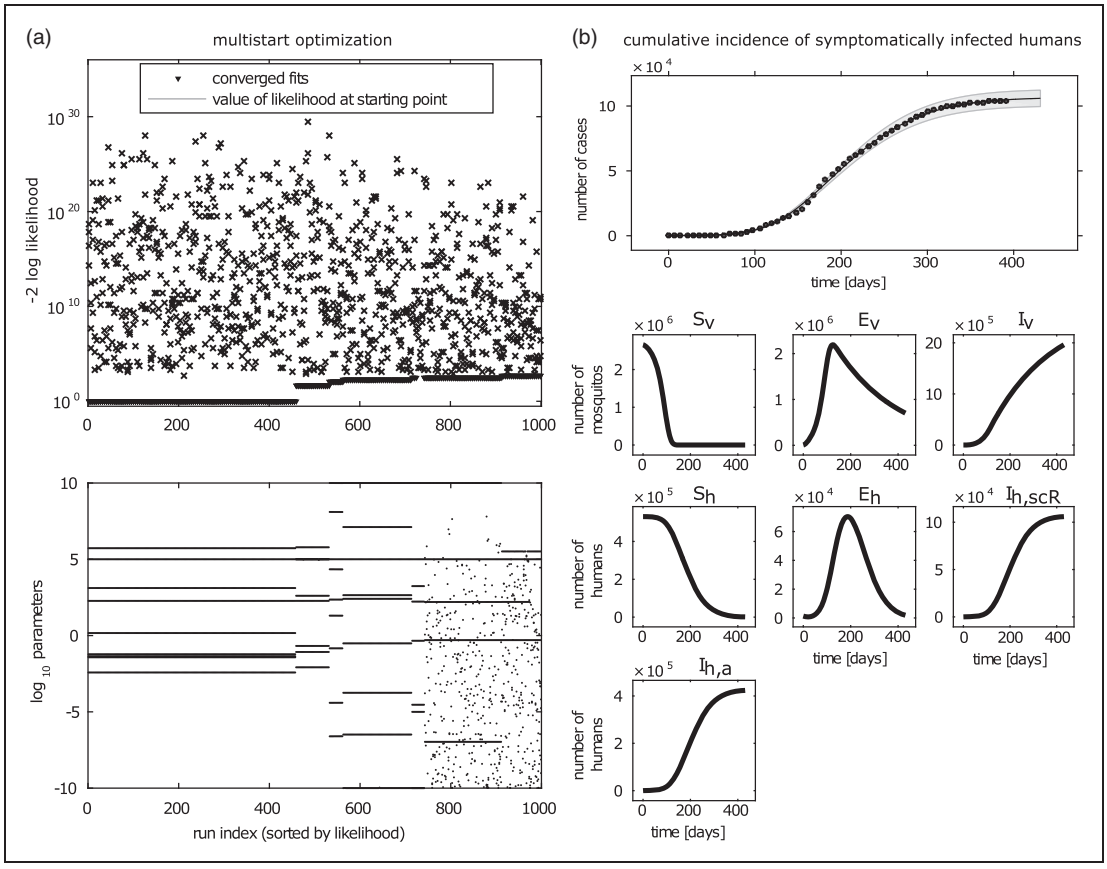

Figure 9. (a) Multistart optimization with 1000 runs for the fully reduced model. The global minimum is found in almost 50% of the runs and the parameter values within this optimum are unique. For illustration, likelihood values are shifted to the baseline by subtraction of the likelihood value for the global minimum. Computation time for 1000 fits was 44.1 min, i.e. approximately 2.6 s for an average single fit on a 3.4 GHz quad-core CPU. (b) Agreement of model and data for the best fit parameters as well as trajectories of all compartments.

A structural non-identifiability is observed for κ_{as}, as its profile is flat over the entire parameter range and moreover shows a functional relationship with parameters S_h^{init}, E_h^{init} and I_h^{init}, as illustrated in Figure 8(a). This behavior originates from the fact that κ_{as} represents the proportion of exposed humans E_h moving to the compartment of asymptomatically infected humans $I_{h,a}$ and $1-\kappa_{as}$ is the proportion moving to symptomatically infected humans $I_{h,a}$. The proportion parameter κ_{as} can be tuned to arbitrary values because the same total amount of symptomatically infected is obtained when compensating with the coupled initial value of exposed humans E_h^{init}. As long as the branch of asymptomatically infected humans cannot be compared to any observation, the parameters cannot be decoupled and κ_{as} remains structurally non-identifiable. Therefore, model reduction can be performed by either remove the module of asymptomatically infected humans $I_{h,a}$ from the model or to fix this parameter to a literature value. Here, we choose the latter and fix $\kappa_{as} = 0.8$ as reported in the literature[58] to be still able to interpret ν_h in terms of a general transmission rate from exposed to infected humans (reduction step 6).

The likelihood profile of β_{hh} in Figure 8(b) indicates that the frequently discussed infection via human–human transmission rate β_{hh} is not required to explain the available data as it is statistically in accordance with zero (final reduction step 7). Thus, based on this analysis, there is no significant relevance of the human-to-human infection of the Zika virus, as already suggested in the literature.[73]

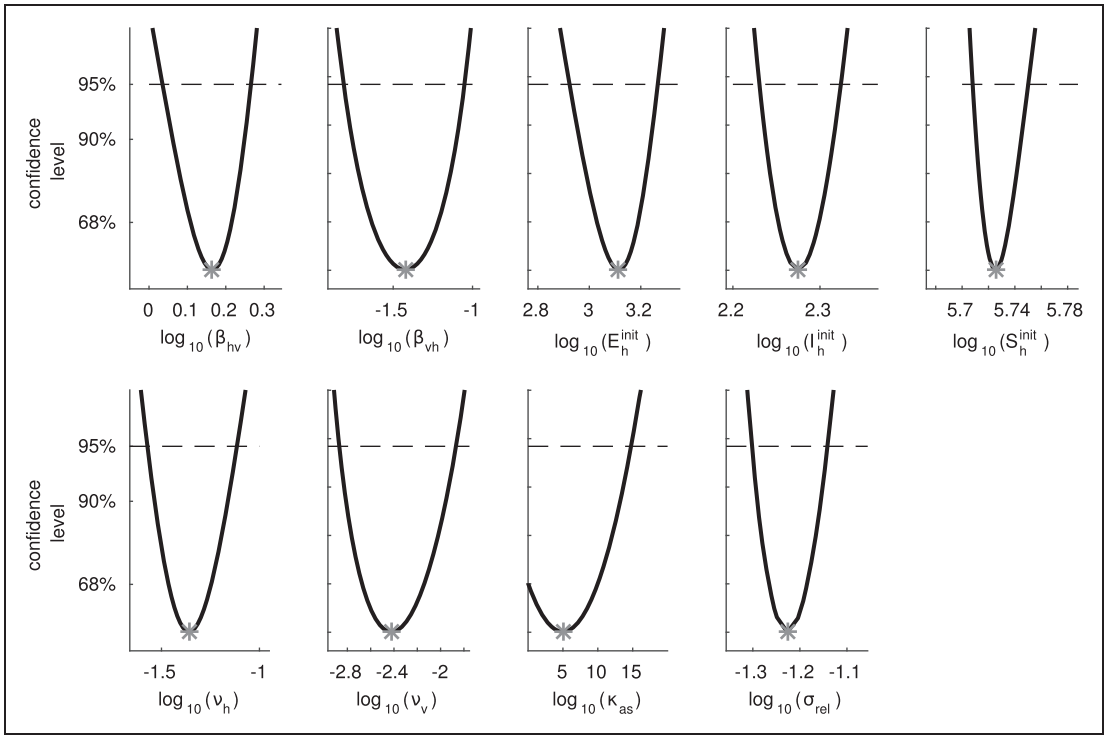

Figure 10. Likelihood profiles of the fully reduced and identifiable model (equation (17)). All parameters are identifiable, although the information about κ_{as} exclusively originates from the assumed prior. Computation time was approximately 3.4 min on a 3.4 GHz quad-core CPU.

Table 3. Estimated parameter values for reduced Zika model (equation (17)).

Parameter	Physical unit	Search region	Estimated parameter value	Profile likelihood-based confidence interval	Assumed prior
β_{hv}	$1/t$	10^{-10}–10^{5}	1.57	[1.17, 1.96]	–
β_{vh}	$1/t$	10^{-10}–10^{5}	0.087	[0.016, 0.092]	–
E_h^{init}	*humans*	10^{-5}–10^{10}	1.30×10^3	$[0.84 \times 10^3, 1.85 \times 10^3]$	–
I_h^{init}	*humans*	10^{-5}–10^{10}	189	[170, 211]	–
S_h^{init}	*humans*	10^{-5}–10^{10}	5.3×10^5	$[5.1 \times 10^5, 5.6 \times 10^5]$	–
ν_h	$1/t$	10^{-10}–10^{5}	0.04	[0.02, 0.07]	–
ν_v	$1/t$	10^{-10}–10^{5}	0.0038	[0.0013, 0.0144]	–
κ_{hv}	$\frac{mosquitos}{human}$	0–20	5	[0, 15]	$\mathcal{N}(5, 5^2)$
σ_{rel}	–	10^{-5}–$10^{-0.3}$	0.06	[0.05, 0.07]	–

4.5 Fully identifiable model after reduction

Multistart optimization of the fully reduced model

$$\begin{aligned}
\dot{S}_v &= -\frac{\beta_{hv} S_v (I_{h,a} + I_{h,scR})}{N_h} \\
\dot{E}_v &= \frac{\beta_{hv} S_v (I_{h,a} + I_{h,scR})}{N_h} - \nu_v E_v \\
\dot{I}_v &= \nu_v E_v \\
\dot{S}_h &= -\frac{\beta_{vh} S_h I_v}{N_v}
\end{aligned}$$

$$
\begin{aligned}
\dot{E}_h &= \frac{\beta_{vh} S_h I_v}{N_v} - \nu_h E_h \\
\dot{I}_{h,a} &= 0.8 \cdot \nu_h E_h \\
\dot{I}_{h,scR} &= 0.2 \cdot \nu_h E_h \\
N_v &= S_v + E_v + I_v \\
N_h &= S_h + E_h + I_{h,a} + I_{h,scR} \\
I_{h,a}^{init} &= 0.8 \cdot I_h^{init} \\
I_{h,scR}^{init} &= 0.2 \cdot I_h^{init}
\end{aligned}
$$

$$
\begin{aligned}
S_v^{init} &= \kappa_{hv} S_h^{init} \\
E_v^{init} &= \kappa_{hv} E_h^{init} \\
I_v^{init} &= \kappa_{hv} I_h^{init}
\end{aligned}
\tag{17}
$$

with the same initial guesses as for the full model in the previous section exhibits a clear global optimum with unique parameter values (Figure 9(a)). Horizontal structures of the fitted parameters in the lower panel of Figure 9(a) support the uniqueness of the global optimum in the parameter space. Consequently, Figure 10 reveals identifiability of all estimated model parameters as well as finite profile likelihood-based confidence intervals, c.f. Table 3. Figure 9(b) shows the agreement of data and model.

5 Summary

In this article, parameter estimation and profile likelihood-based analyses have been applied to two examples of infectious disease models. Keeping the usage of prior knowledge for parameter values to a minimum, solely information contained in the data of infected individuals was used in order to estimate parameters and assess their uncertainties. The first example was a basic *SIR* model where all parameters were identifiable.

As a second example, a comprehensive model of Zika virus infection via mosquitos and with data of infected humans from Colombia was analyzed. Profile likelihood-based analysis shows non-identifiability of several parameters. It can be concluded that the data does not provide information about the duration of the acute or convalescent phase of the ZIKV infection disease in humans, the proportion of asymptomatically infected humans or the population size of the mosquitos. Moreover, the data alone does not provide evidence for human–human infections. After elimination of non-identifiabilities, a minimal model with identifiable parameters could be derived.

It should be noted that here we focused on a data based model reduction scheme, which points to a result where all aspects of the model can be restricted to finite confidence intervals by the analyzed data. The validity and benefit of such a strategy as well as the number of reasonable reduction steps depend on the application setting. Therefore, model reduction always has to be augmented with plausibility checks to prevent wrong conclusions. However, the method nicely illustrates the amount and quality of information contained in the analyzed data. Furthermore, this procedure demonstrates the commonly strong dependence of infectious disease models on prior knowledge and the risk of drawing conclusions predominately based on prior information rather than on recorded data.

In both investigated examples, there is a good agreement between data and model and it was demonstrated how uncertainties about model parameters translate to compartment trajectories. Since the profile likelihood constitutes an intuitive generalization of classical approaches for quantification of uncertainties to nonlinear settings as they occur in ODE models, we enforce usage of the reviewed methodology in the infectious disease modeling field.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors acknowledge funding of the projects EA:sys (031L0080), LungSysII (0316042A) and LiSyM (031L0042) by the German Federal Ministry of Education and Research (BMBF).

References

1. Daley DJ and Gani JM. Cambridge studies in mathematical biology. *Epidemic modelling: an introduction*. Cambridge, NY: Cambridge University Press, 1999.
2. Chretien JP, George D, Shaman J, et al. Influenza forecasting in human populations: a scoping review. *Plos ONE* 2014; **9**: e94130.
3. Weidemann F, Dehnert M, Koch J, et al. Bayesian parameter inference for dynamic infectious disease modelling: rotavirus in Germany. *Stat in Med* 2014; **33**: 1580–1599.
4. Birrell PJ, Ketsetzis G, Gay NJ, et al. Bayesian modeling to unmask and predict influenza A/H1N1pdm dynamics in London. *PNAS* 2011; **108**: 18238–18243.
5. Shaman J, Karspeck A, Yang W, et al. Real-time influenza forecasts during the 2012–2013 season. *Nat Commun* 2013; **4**: 2837.
6. Farah M, Birrell P, Conti S, et al. Bayesian emulation and calibration of a dynamic epidemic model for A/H1N1 influenza. *JASA* 2014; **109**: 1398–1411.
7. Wu JT, Leung K, Perera RAPM, et al. Inferring influenza infection attack rate from seroprevalence data. *PLOS Path* 2014; **10**: e1004054.
8. Xing Z, Nicholson B, Jimenez M, et al. Bayesian modeling of temporal properties of infectious disease in a college student population. *J Appl Stat* 2014; **41**: 1358–1382.
9. Chan KP, Wong CM, Chiu SSS, et al. A robust parameter estimation method for estimating disease burden of respiratory viruses. *Plos ONE* 2014; **9**: e90126.
10. Nsoesie EO, Marathe M and Brownstein JS. Forecasting peaks of seasonal influenza epidemics. *PLoS Curr* 2013; 1.
11. Held L and Paul M. Modeling seasonality in space-time infectious disease surveillance data. *Biom J* 2012; **54**: 824–843.
12. Ming RX, Liu J, Cheung WKW, et al. Stochastic modelling of infectious diseases for heterogeneous populations. *Infect Dis Poverty* 2016; **5**: 107.
13. Zimmer C, Yaesoubi R and Cohen T. A likelihood approach for real-time calibration of stochastic compartmental epidemic models. *PLOS Comp Biol* 2017; **13**: e1005257.
14. Chitnis N, Hyman JM and Cushing JM. Determining important parameters in the spread of malaria through the sensitivity analysis of a mathematical model. *Bull Math Biol* 2008; **70**: 1272.
15. Marino S, Hogue IB, Ray CJ, et al. A methodology for performing global uncertainty and sensitivity analysis in systems biology. *J Theor Biol* 2008; **254**: 178–196.
16. Raue A, Kreutz C, Maiwald T, et al. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* 2009; **25**: 1923–1929.
17. Raue A, Becker V, Klingmüller U, et al. Identifiability and observability analysis for experimental design in nonlinear dynamical models. *Chaos* 2010; **20**: 045105.
18. Kreutz C, Raue A, Kaschek D, et al. Profile likelihood in Systems Biology. *FEBS J* 2013; **280**: 2564–2571.
19. Maiwald T, Hass H, Steiert B, et al. Driving the model to its limit: profile likelihood based model reduction. *Plos ONE* 2016; **11**: e0162366.
20. Raue A, Steiert B, Schelker M, et al. Data2Dynamics: a modeling environment tailored to parameter estimation in dynamical systems. *Bioinformatics* 2015; **31**: 3558–3560.
21. D2D development team. Data2Dynamics software, www.data2dynamics.org (2016, accessed 14 October 2016).
22. Steiert B, Raue A, Timmer J, et al. Experimental design for parameter estimation of gene regulatory networks. *Plos ONE* 2012; **7**: e40052.
23. Meyer P, Cokelaer T, Chandran D, et al. Network topology and parameter estimation: from experimental design methods to gene regulatory network kinetics using a community based approach. *BMC Sys Biol* 2014; **8**: 13.
24. Cox DR and Hinkley DV. *Theoretical statistics*. Boca Raton, FL: Chapman & Hall/CRC, 2000.
25. Raue A, Schilling M, Bachmann J, et al. Lessons learned from quantitative dynamical modeling in systems biology. *PLoS ONE* 2013; **8**: e74335.
26. Kreutz C. New concepts for evaluating the performance of computational methods. *IFAC-PapersOnLine* 2016; **49**: 63–70.
27. Bachmann J, Raue A, Schilling M, et al. Division of labor by dual feedback regulators controls JAK2/STAT5 signaling over broad ligand range. *Mol Syst Biol* 2011; **7**: 516.
28. Becker V, Schilling M, Bachmann J, et al. Covering a broad dynamic range: information processing at the erythropoietin receptor. *Science* 2010; **328**: 1404–1408.
29. Beer R, Herbst K, Ignatiadis N, et al. Creating functional engineered variants of the single-module non-ribosomal peptide synthetase IndC by T domain exchange. *Mol BioSyst* 2014; **10**: 1709–1718.

30. Raia V, Schilling M, Böhm M, et al. Dynamic mathematical modeling of IL13-induced signaling in Hodgkin and primary mediastinal B-Cell lymphoma allows prediction of therapeutic targets. *Cancer Res* 2011; **71**: 693–704.
31. The MathWorks Inc. MATLAB (R2016b), 2016.
32. Hindmarsh AC, Brown PN, Grant KE, et al. SUNDIALS: suite of nonlinear and differential/algebraic equation solvers. *ACM Trans Math Softw* 2005; **31**: 363–396.
33. Leis JR and Kramer MA. The simultaneous solution and sensitivity analysis of systems described by ordinary differential equations. *ACM Trans Math Softw* 1988; **14**: 45–60.
34. Venzon DJ and Moolgavkar SH. Method for computing profile-likelihood-based confidence intervals. *J R Stat Soc Ser C Appl Stat* 1988; **37**: 87–94.
35. Murphy SA and van der Vaart AW. On profile likelihood. *JASA* 2000; **95**: 449–465.
36. King AA, Domenech de Cellès M, Magpantay FMG, et al. Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to ebola. *Proc R Soc B* 2015; **282**: 20150347.
37. Meeker WQ and Escobar LA. Teaching about approximate confidence regions based on maximum likelihood estimation. *Am Stat* 1995; **49**: 48–53.
38. Cobelli C and DiStefano JJ. Parameter and structural identifiability concepts and ambiguities: a critical review and analysis. *Am J Physiol* 1980; **239**: R7–R24.
39. Margaria G, Riccomagno E, Chappell MJ, et al. Differential algebra methods for the study of the structural identifiability of rational function state-space models in the biosciences. *Math Biosci* 2001; **174**: 1–26.
40. White LJ, Evans ND, Lam TJ, et al. The structural identifiability and parameter estimation of a multispecies model for the transmission of mastitis in dairy cows. *Math Biosci* 2001; **174**: 77–90.
41. Chis OT, Banga JR and Balsa-Canto E. Structural identifiability of systems biology models: a critical comparison of methods. *Plos ONE* 2011; **6**: e27755.
42. Chis O, Banga JR and Balsa-Canto E. GenSSI: a software toolbox for structural identifiability analysis of biological models. *Bioinformatics* 2011; **27**: 2610–2611.
43. Rodríguez-Fernández M, Balsa-Canto E, Egea JA, et al. Identifiability and robust parameter estimation in food process modeling: application to a drying model. *J Food Eng* 2007; **83**: 374–383.
44. Ljung L and Glad T. On global identifiability for arbitrary model parametrizations. *Automatica* 1994; **30**: 265–276.
45. Merkt B, Timmer J and Kaschek D. Higher-order Lie symmetries in identifiability and predictability analysis of dynamic models. *Phys Rev E* 2015; **92**: 12–20.
46. Neyman J and Pearson ES. On the problem of the most efficient tests of statistical hypotheses. *Phil Trans R Soc A* 1933; **231**: 289–337.
47. Anonymous. Influenza in a boarding school. *BMJ* 1978; **1**: 587.
48. Martcheva M. *An introduction to mathematical epidemiology*. Springer: New York, USA, 2015.
49. Keeling MJ and Rohani P. *Modeling infectious diseases in humans and animals*. Princeton, NJ, United States: Princeton University Press, 2011.
50. Murray JD (ed.) *Mathematical biology: I. An introduction, Interdisciplinary applied mathematics*. volume 17, New York, United States: Springer New York, 2004.
51. Barnes B and Fulford G. *Mathematical modelling with case studies: a differential equations approach using Maple and MATLAB*, 2nd ed. Boca Raton, FL, United States: CRC Press, 2002.
52. Kermack WO and McKendrick AG. A contribution to the mathematical theory of epidemics. *Proc R Soc A* 1927; **115**: 700–721.
53. Diekmann O, Heesterbeek J and Metz J. On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations. *J Math Biol* 1990; **28**: 365–382.
54. Mills CE, Robins JM and Lipsitch M. Transmissibility of 1918 pandemic influenza. *Nature* 2004; **432**: 904–906.
55. Earn DJD, Dushoff J and Levin SA. Ecology and evolution of the flu. *Trends Ecol Evol* 2002; **17**: 334–340.
56. Fiore AE, Shay DK, Broder K, et al. Prevention and control of influenza: recommendations of the advisory committee on immunization practices (ACIP), 2008. *MMWR* 2008; **57**: 1–60.
57. Malone RW, Homan J, Callahan MV, et al. Zika virus: Medical countermeasure development challenges. *PLOS Negl Trop Dis* 2016; **10**: e0004530.
58. European Centre for Disease Prevention and Control (ECDC). Factsheet for health professionals, http://ecdc.europa.eu/en/healthtopics/zika_virus_infection/factsheet-health-professionals/Pages/factsheet_health_professionals.aspx (2016, accessed 14 October 2016).
59. Rasmussen SA, Jamieson DJ, Honein MA, et al. Zika virus and birth defects reviewing the evidence for causality. *N Engl J Med* 2016; **374**: 1981–1987.
60. Brasil P, Sequeira PC, Freitas AD, et al. Guillain-Barr syndrome associated with Zika virus infection. *Lancet* 2016; **387**: 1482.
61. Cauchemez S, Besnard M, Bompard P, et al. Association between Zika virus and microcephaly in French Polynesia, 2013–2015: a retrospective study. *Lancet* 2016; **387**: 2125–2132.
62. World Health Organisation (WHO). WHO statement on the first meeting of the International Health Regulations (2005) (IHR 2005) Emergency Committee on Zika virus and observed increase in neurological disorders and neonatal

malformations, http://www.who.int/mediacentre/news/statements/2016/1st-emergency-committee-zika/en/ (2016, accessed 14 October 2016).
63. Ross R. Some a priori pathometric equations. *BMJ* 1915; **1**: 546–547.
64. Macdonald G. The analysis of infection rates in diseases in which superinfection occurs. *Trop Dis Bull* 1950; **47**: 907–915.
65. Kucharski AJ, Funk S, Eggo RM, et al. Transmission dynamics of Zika virus in island populations: a modelling analysis of the 201314 French Polynesia outbreak. *PLOS Negl Trop Dis* 2016; **10**: e0004726.
66. Ferguson NM, Cucunub ZM, Dorigatti I, et al. Countering the Zika epidemic in Latin America. *Science* 2016; **353**: 353–354.
67. Funk S, Kucharski AJ, Camacho A, et al. Comparative analysis of dengue and Zika outbreaks reveals differences by setting and virus. *PLoS Negl Trop Dis* 2016; **10**: e0005173.
68. Gao D, Lou Y, He D, et al. Prevention and control of Zika as a mosquito-borne and sexually transmitted disease: a mathematical modeling analysis. *Sci Rep* 2016; **6**: 28070.
69. Zhang Q, Sun K, Chinazzi M, et al. Spread of Zika virus in the Americas. *PNAS* 2017; **114**: E4334–E4343.
70. Ndeffo-Mbah ML, Parpia AS and Galvani AP. Mitigating prenatal Zika virus infection in the Americas. *Ann Intern Med* 2016; **165**: 551–559.
71. Mansuy JM, Dutertre M, Mengelle C, et al. Zika virus: high infectious viral load in semen, a new sexually transmitted pathogen? *Lancet Infect Dis* 2016; **16**: 405.
72. Hills SL, Russell K, Hennessey M, et al. Transmission of Zika virus through sexual contact with travelers to areas of ongoing transmission – continental United States. *MMWR* 2016; **65**: 215–216.
73. Towers S, Brauer F, Castillo-Chavez C, et al. Estimate of the reproduction number of the 2015 Zika virus outbreak in Barranquilla, Colombia, and estimation of the relative role of sexual transmission. *Epidemics* 2016; **17**: 50–55.
74. Atkinson B, Hearn P, Afrough B, et al. Detection of Zika virus in semen. *Emerg Infect Dis* 2016; **22**: 940.
75. Musso D, Roche C, Robin E, et al. Potential sexual transmission of Zika virus. *Emerg Infect Dis* 2015; **21**: 359–361.
76. Instituto Nacional de Salud Colombia. Boletín epidemiológico semanal, número 36 de 2016, p. 106, gráfica 98, http://www.ins.gov.co/boletin-epidemiologico/Boletn%20Epidemiolgico/2016%20Boletin%20epidemiologico%20semana%2036.pdf (2016, accessed 14 October 2016).
77. Rohatgi A. WebPlotDigitizer, http://arohatgi.info/WebPlotDigitizer (2016, accessed 14 October 2016).
78. Manore CA, Hickmann KS, Xu S, et al. Comparing dengue and chikungunya emergence and endemic transmission in *Aaegypti* and *A. albopictus*. *J Theor Biol* 2014; **356**: 174–191.
79. Manore C, Ostfeld R, Agusto F, et al. Defining the risk of Zika and chikungunya virus transmission in human population centers of the eastern United States. *PLoS Negl Trop Dis* 2016; **11**(1): e0005255.