Computer Methods and Programs in Biomedicine Update 7 (2025) 100178

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine Update

journal homepage:
www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine-update

A comparative approach of analyzing data uncertainty in parameter estimation for a Lumpy Skin Disease model

Edwiga Renald [a,*], Miracle Amadi [c], Heikki Haario [c], Joram Buza [b], Jean M. Tchuenche [a], Verdiana G. Masanja [a]

[a] School of Computational and Communication Science and Engineering (CoCSE), The Nelson Mandela African Institution of Science and Technology, P.O. Box 447, Arusha, Tanzania
[b] School of Life Science and Biomedical Engineering (LiSBE), Nelson Mandela African Institution of Science and Technology, P.O. Box 447, Arusha, Tanzania
[c] School of Engineering Science, Lappeenranta University of Technology, 53851, Lappeenranta, Finland

ARTICLE INFO

Keywords:
Lumpy skin disease
Uncertainty assessment
Parameter identification
Markov Chain Monte Carlo

ABSTRACT

The livestock industry has been economically affected by the emergence and reemergence of infectious diseases such as Lumpy Skin Disease (LSD). This has driven the interest to research efficient mitigating measures towards controlling the transmission of LSD. Mathematical models of real-life systems inherit loss of information, and consequently, accuracy of their results is often complicated by the presence of uncertainties in data used to estimate parameter values. There is a need for models with knowledge about the confidence of their long-term predictions. This study has introduced a novel yet simple technique for analyzing data uncertainties in compartmental models which is then used to examine the reliability of a deterministic model of the transmission dynamics of LSD in cattle which involves investigating scenarios related to data quality for which the model parameters can be well identified. The assessment of the uncertainties is determined with the help of Adaptive Metropolis Hastings algorithm, a Markov Chain Monte Carlo (MCMC) standard statistical method. Simulation results with synthetic cases show that the model parameters are identifiable with a reasonable amount of synthetic noise, and enough data points spanning through the model classes. MCMC outcomes derived from synthetic data, generated to mimic the characteristics of the real dataset, significantly surpassed those obtained from actual data in terms of uncertainties in identifying parameters and making predictions. This approach could serve as a guide for obtaining informative real data, and adapted to target key interventions when using routinely collected data to investigate long-term transmission dynamic of a disease.

1. Background

The emergence and re-emergence of livestock diseases such as Lumpy Skin Disease (LSD) pose a significant economic challenge to the livestock industry [1]. Lumpy skin disease (LSD) is an economically important trans-boundary viral disease of cattle [2–4], transmitted indirectly through blood-feeding insect vectors such as flies and mosquitoes and some species of ticks [5]. The transmission route involves a susceptible vector biting an infectious cattle, and thus carrying and transmitting it to another susceptible cattle through biting. As the risk of epidemics grows, mathematical and statistical inference and simulation methods can play a crucial role in developing effective prevention and mitigation strategies [1,6].

Mathematical modeling has of late been an important tool in studying dynamics of various real-world phenomena [7–9]. These models are defined by a system of equations and their parameters that together quantify system states through a class of connected dynamic quantities [1]. However, dynamical system models are somewhat biased due to limitations in their ability to fully capture actual phenomena, randomness in the occurrence of some events. One of the reasons is that when formulating or developing deterministic models, a few assumptions which are regarded as the physical laws of the actual phenomena are taken into account to minimize complexities in order to theoretically analyze the models. Nevertheless, in order to manage uncertainty in the model's output and to ensure reliable predictions, the literature recommends the use of sufficiently informative and less noisy data when estimating parameters [6,10,11]. Various studies have done uncertainty quantification in epidemiological models but a few focused on the specific features in data uncertainty. Other studies such as [12–14] were based on the general aspect of uncertainty.

Herein, a novel yet simple technique for analyzing data uncertainties in compartmental models is introduced, then implemented in a

* Corresponding author.
E-mail address: edwiga1991@gmail.com (E. Renald).

https://doi.org/10.1016/j.cmpbup.2025.100178
Received 15 July 2024; Received in revised form 30 December 2024; Accepted 13 January 2025
Available online 20 January 2025

Susceptible–Exposed–Infectious–Recovered–Susceptible (SEIRS) mathematical model of the transmission dynamics of LSD of cattle, thoroughly described in [15]. The model in [15] considers the waning rate of LSD infection-induced immunity among other important epidemiological facts on LSD transmission which has not been considered in previous studied such as [16–18]. This is based on the fact that even after developing natural immunity, recovered cattle may still be at risk of contracting the disease again because the duration of immunity induced by LSD infection remains uncertain [19,20]. This underscores the complexity of LSD dynamics, making it essential to use a nuanced model that includes recovery–reinfection risks for infectious disease with recurring incidence such as LSD. As shown in [15], the parameter capturing the transition from recovery to susceptible plays a key role in LSD dynamics, and omitting it would weaken the credibility of the analyses and results in this study. The importance of this parameter has also been emphasized in other epidemiological modeling studies [21,22].

Markov Chain Monte Carlo (MCMC) sampling is a Bayesian parameter estimation technique that produces samples from the distribution of the parameters [23]. The Adaptive MCMC employed herein is a class of algorithms that automatically adjust their parameters based on previous samples to improve performance. While MCMC is typically used to directly estimate model parameters, our approach uniquely leverages it by comparing these three critical factors to offer new insights into data availability, evaluate its impact by varying length of data-time, noise levels, and the number of compartmental classes on model reliability. Our method establishes practical guidelines for optimal data collection in epidemiological research. Additionally, its adaptability extends its potential applications to various infectious diseases and data types, showcasing its broad utility. Such an innovative approach is expected to significantly contribute to the field of mathematical modeling by demonstrating that sufficient and high-quality data are essential for accurate model predictions, thereby informing more effective public health strategies. Also, by incorporating both synthetic and real data, we provide a more comprehensive validation of the model than the one presented in [15].

Data uncertainty (that is uncertainty in specified model parameters) stands out as the most frequently cited uncertainty in infectious diseases models [24]. Thus, the proposed study aims to examine if the model can produce promising and trustworthy predictions despite its limitations if enough, well-informed and less noisy data are available.

Markov Chain Monte Carlo (MCMC) method is renowned for its ability to check the uncertainty present in dynamical system models [10,11]. Moreover, it has been regarded as a predominant method that can give various sets of parameters (parameter space) that give similar model output. Markov Chain Monte Carlo (MCMC) is often preferred to other common traditional methods such as the Least Squares method which gives a single parameter set that produces a model's best fit to data. In addition, Adaptive Metropolis–Hastings is an improved version of Metropolis–Hastings (MH) as it corrects previously sampled points of proposal covariance during MCMC run.

The contribution of this study is as follows - there are less uncertainties when data are - collected over a longer period - available for all the model variables - with synthetic data, the model performs better than when real data is used as predictive plots exhibit significantly larger uncertainties with higher levels of noise and poorly fit the data trend in the latter. This performance is due to several uncertainties around the data, and the long interval of time between the collected data points.

The rest of this paper is organized as follows; Section 2 presents materials and methods used, Section 3 presents and discusses results from the uncertainty analysis based on different factors, and Section 4 provides a conclusion based on the study findings.

2. Materials and methods

2.1. Model equations

The model considers two populations, namely, cattle population (c) and vector population (v). The cattle population is subdivided into four classes which are Susceptible (S_c), Exposed (E) due to a considerable incubation period, Infectious (I_c), and Recovered (R). Blood-feeding vectors are grouped into two classes which are, Susceptible S_v (those blood-feeding vectors with no virus) and Infectious I_v (those blood-feeding vectors with virus). For more information see the compartmental flow diagram in Fig. 1.

Following is the system of non-linear ordinary differential equations governing the transmission dynamics between cattle and vector populations.

$$\begin{aligned}
\frac{dS_c}{dt} &= \Lambda_c + \omega R - (\mu_c + \frac{av_c I_v}{N_v})Sc,\\
\frac{dE}{dt} &= \frac{av_c I_v}{N_v} S_c - (\mu_c + \rho)E,\\
\frac{dI_c}{dt} &= \rho E - (\mu_c + \phi + \sigma)I_c,\\
\frac{dR}{dt} &= \phi I_c - (\mu_c + \omega)R,\\
\frac{dS_v}{dt} &= \Lambda_v - (\mu_v + \frac{av_v I_c}{N_c})Sv,\\
\frac{dI_v}{dt} &= \frac{av_v I_c}{N_c} S_v - \mu_v I_v,
\end{aligned} \tag{1}$$

where the forces of infections for cattle and vector populations are respectively given by

$$\frac{av_c I_v}{N_v}, \tag{2}$$

and

$$\frac{av_v I_c}{N_c} \tag{3}$$

with initial conditions; $S_c(0) > 0$, $E(0) \geq 0$, $I_c(0) \geq 0$, $R(0) \geq 0$, $S_v(0) > 0$ and $I_v(0) \geq 0$.

2.2. Relationship between model parameters and the real system

Due to the presence of LSD, cattle and vector populations are divided into various classes of LSD progression states. For the cattle population, there are four compartments, namely: susceptible, exposed, infectious and recovered. The vector population comprises the susceptible and infectious classes only. Various parameters explain the rates at which individuals in different classes progress to subsequent classes. For instance, Λ_c is the proportion of individuals in the cattle population who are at risk of contracting Lumpy Skin Disease Virus due to the presence of vectors and infectious cattle in the population. This proportion is free of LSD. When a vector bites an infectious cattle, it become infected, after which if it bites an uninfected cattle, it transmit the virus and thus the bitten cattle becomes infected. The infected individuals who have not yet developed the symptoms are called exposed individuals. A proportion $\lambda_c S_c$ of susceptible cattle progress to exposed class, E. At this early transition stage, the cattle is not yet infectious because it takes some noticeable period of time, known as the incubation period, $\frac{1}{\rho}$ to develop LSD symptoms. After the incubation period, individuals in the exposed group of cattle becomes infectious to LSD at rate ρ and move to the infectious class I_c. Since LSD is a fatal disease, an LSD-induced mortality rate σ is considered. Conversely, once infected, cattle generally develop a natural immunity to combat the pathogen. This immune response often leads to recovery at a rate of ϕ, implying that a group of size ϕI_c move to the recovered class, R. However, as natural immunity wanes over time, recovered cattle can become susceptible to LSD again, transitioning back to the

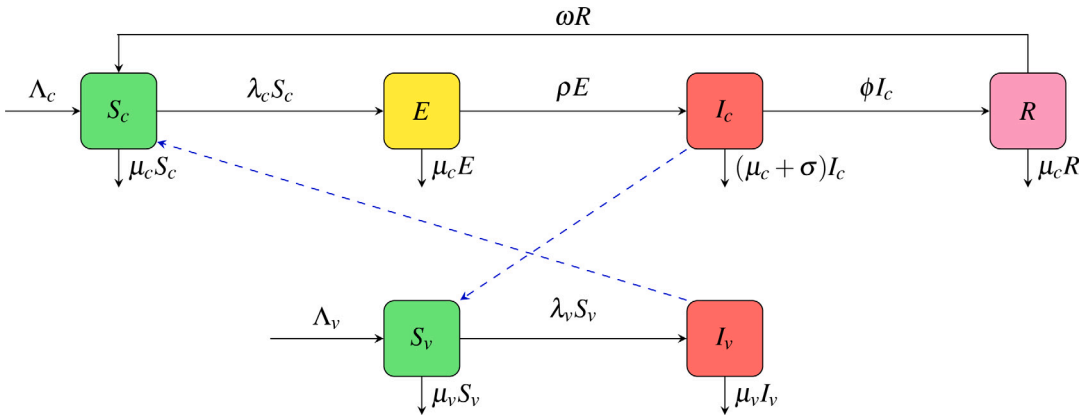


Fig. 1. Compartmental Flow Diagram.

Table 1
Model parameters' descriptions, values and sources. The assumed parameter values are hypothetical while one parameter, μ_c, has been calculated based on the cited references. These parameters were used to simulate the model to generate synthetic data for demonstrating the workability and identifiability of the model under different scenarios.

Parameter	Definition	Value	Reference
Λ_c	Cattle recruitment rate	500	Assumed
Λ_v	Vector recruitment rate	10 000	[15]
μ_c	Natural death rate for cattle population	0.0001	Calculated [26,27]
σ	LSD induced death rate	0.01	[28,29];
ρ	LSD incubation rate	1/5	[30]
ϕ	Recovery rate	0.06	Assumed
ω	Waning rate	0.01	Assumed
μ_v	Vector natural mortality	0.075	[31]
a	Vector biting rate	0.9	[31]
ν_c	Probability of a susceptible cattle to become infectious after being bitten by an infectious vector	0.64	[31]
ν_v	Probability of a susceptible vector to become infectious after biting an infectious cattle	0.46	[31]

susceptible class S_c at a rate of ω, meaning that a proportion ωR of the recovered cattle returns to susceptibility. In the vector population, the susceptible class S_v is replenished at a constant rate Λ_v through birth or immigration. Upon infection, susceptible vectors move to the infectious class I_v at a rate λ_v. Since the incubation period in the vector population is still not known [25], it is assumed that vectors become infectious immediately upon contracting the virus, and thus, no exposed class is considered. Infected vectors remain infectious for their lifetime, as they do not recover from LSD. Both cattle and vector populations experience natural mortality, with death rates of μ_c and μ_v, respectively. As LSD progresses through different stages, individuals shift from one stage to the next, decreasing the class size of the current stage and increasing the class size of the subsequent stage. These transitions collectively describe the dynamics of LSD spread and progression in both populations over time. The blue dashed lines indicate vector-host interaction via vector bites which result to virus transmission.

Description of all model parameters, and their respective values used for the model simulations are provided in Table 1. Herein, the focus is on the model parameter identifiability. For details of the model analysis, see [15].

2.2.1. Data availability

This study utilizes the dataset (see Fig. 2) that comprises biannual data collected from the Rukwa Region in Tanzania by the Ministry of Livestock and Fisheries of the United Republic of Tanzania [32]. This dataset encompasses information on both susceptible and infectious cattle classes and consists of 30 data points in total (see Fig. 2); spanning a period of 15 years from July 2005 to December 2019. Here, Rukwa region has been chosen because it represents the Southern Highlands area targeted by the government to be the leading area for livestock husbandry in Tanzania [33]. The data was collected in the districts and then compiled to come up with regional data. One of the limitations with the dataset is presence of missing values. Moreover, there is no information on how the data was processed after collection.

2.3. MCMC setup

Herein, MCMC method is applied and results of different scenarios are compared along with the discussion of parameters' uncertainties. MATLAB was chosen for its robust MCMC toolbox developed in [34] which includes the Adaptive Metropolis–Hastings algorithm. This advanced feature allows corrections in the proposal distributions of previously accepted samples, enhancing both the convergence speed and the quality of the parameter samples [35]. Additionally, MATLAB's user-friendly interface and support for parallel computations make it ideal for efficiently running multiple MCMC simulations. The use of the *fminsearch* function for least squares fitting was to refine parameter estimates to values closer to their true optimum before initiating MCMC. This step ensures that the MCMC sampling process starts from a more accurate parameter space, thereby improving convergence and efficiency (see Table 2). Detailed explanation of the algorithm and a step-by-step guideline of using it are provided respectively in [34], and [36]. Results of the implementation of the MCMC algorithm accompanied by the Adaptive Metropolis–Hastings sampling technique using the generated synthetic data are provided in the next section. To generate the figures in each case, the model ran 12,000 simulations.

2.4. Comparison with existing common methods of handling data uncertainty in parameter estimation

A brief comparison of the proposed Adaptive MCMC approach with other methods for handling data uncertainty, such as Bootstrapping and

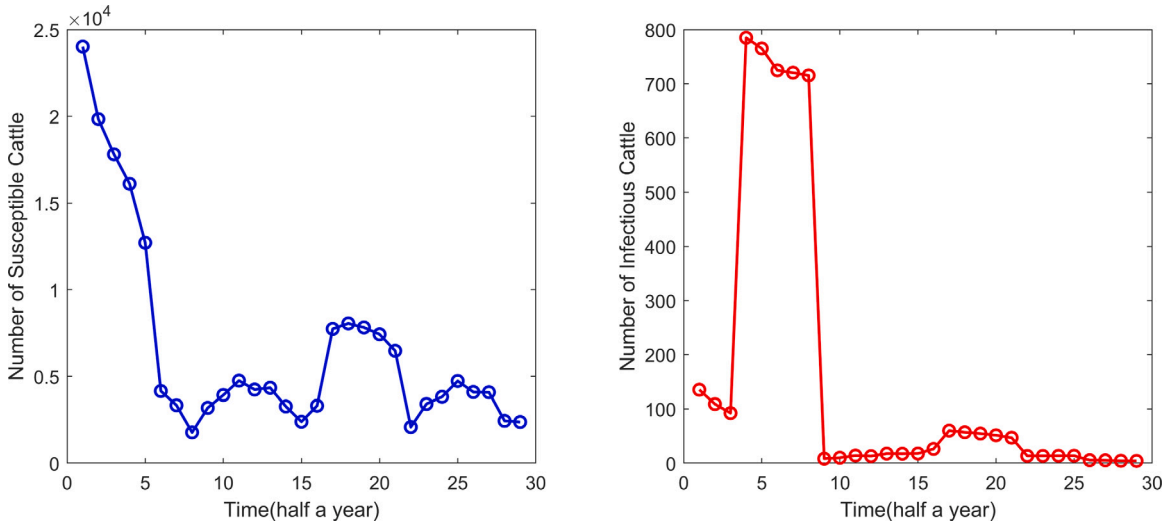


Fig. 2. Data showing the six months semester-wise incidence of LSD for Rukwa region in Tanzania from July 2005 to December 2019.

Table 2
A step-by-step guide on the use of adaptive MCMC by Haario *et al.*, [34] and implemented in [36].

Adaptive Metropolis Hastings Algorithm
Input: Initial parameter θ_1, proposal distribution $q(.\|\theta)$, chain length N
Output: Posterior parameter samples $\{\theta_k\}_{k=1}^{N}$
While *the number of iteration is less than the chain length* **do**
- Sample a new candidate point ; $\theta^* \approx (\theta^*\|\theta_k)$
- Compute $\alpha(\theta^*, \theta_k) = \min\left(1, \frac{\pi(\theta^*\|y)}{\pi(\theta_k\|y)}\right)$
- Sample a random variable r from a standard uniform distribution ie $r \in [0,1]$
if $r \leq \alpha(\theta^*, \theta_k)$ **then**
Accept: set $\theta_{k+1} = \theta^*$
else
Reject: set $\theta_{k+1} = \theta_k$
end
Update: $C_{k+1} = \mathrm{Cov}(\theta_1, \ldots, \theta_k)$
end

Jackknife provides valuable insights and better understanding of the significance of the proposed method relative to existing approaches.

Bootstrapping generates new datasets by resampling the original data with replacement or by resampling residuals, providing robust uncertainty estimates without assuming underlying data distributions [37]. However, it is computationally intensive, relies heavily on the success of optimization routines, and can struggle with small datasets where resampled subsets may inadequately represent the data [38]. Similarly, the Jackknife method estimates parameter variability by systematically leaving out portions of the data [39], making it computationally lighter and effective for small datasets. But, its results may be sensitive to the size of the omitted data points, and the method is therefore less suited for models with complex parameter spaces or highly nonlinear relationships [40,41].

MCMC algorithms used to draw samples from probability distributions, that are too complex to study using analytic techniques alone, excel in high-dimensional problems with complex correlations by providing a full posterior distribution of parameter estimates, capturing both uncertainties and parameter correlations [42,43]. While also computationally demanding, they are particularly effective with noisy or sparse datasets where traditional resampling methods might struggle [44]. For instance, Bootstrapping may fail to adequately capture complex dependencies between parameters.

Bayesian approaches offer greater flexibility but may require careful selection of priors, which may introduce biases if the prior information is not well-founded [45]. However, the Adaptive MCMC methods help to alleviate the practical issues of sampling by dynamically adapting the proposal distribution. While they do not eliminate the need for prior selection, they help the sampling to take into account the priors. In addition, the use of preliminary optimization with *fminsearch* to provide initial guesses for parameters, complements this by providing a data-driven starting point, and thereby mitigating the influence of priors.

3. Results and discussion

3.1. Synthetic data

3.1.1. Generating synthetic data

To generate synthetic data, a relative Gaussian noise was added to the solutions of model (1) with a standard deviation (SD) of 0.05, which mimics the noise level seen in the real data. This was implemented using parameter values shown in Table 1. Data comprising 150 data points (see Fig. 3) for each of the model states, with $SD = 0.05$ were generated.

3.1.2. Assessing parameter uncertainties

As some of the model parameter values are known from the literature, we estimate only the remaining seven unknown parameters.

Uncertainties based on number of model classes whose data are available. Here, uncertainties were compared in three cases. A case where we have data for four compartments of cattle population, a case where we have data for compartments S_c and I_c only, and a case where we have data for compartment I_c only. Results show that there is much uncertainty when we have data for one compartment, I_c only, followed by when we have data for two compartments, S_c and I_c. When data for all compartments are available, we have the least uncertainty compared to the above-mentioned cases. For more information see Fig. 4. These results indicate that with sufficient information available, we anticipate encountering less uncertainty compared to situations with less available information. In all cases, 150 data points and 0.05 noise level were used.

In Fig. 4, it can be seen that as less class data is used, the sampled values significantly deviate from the true parameter values.

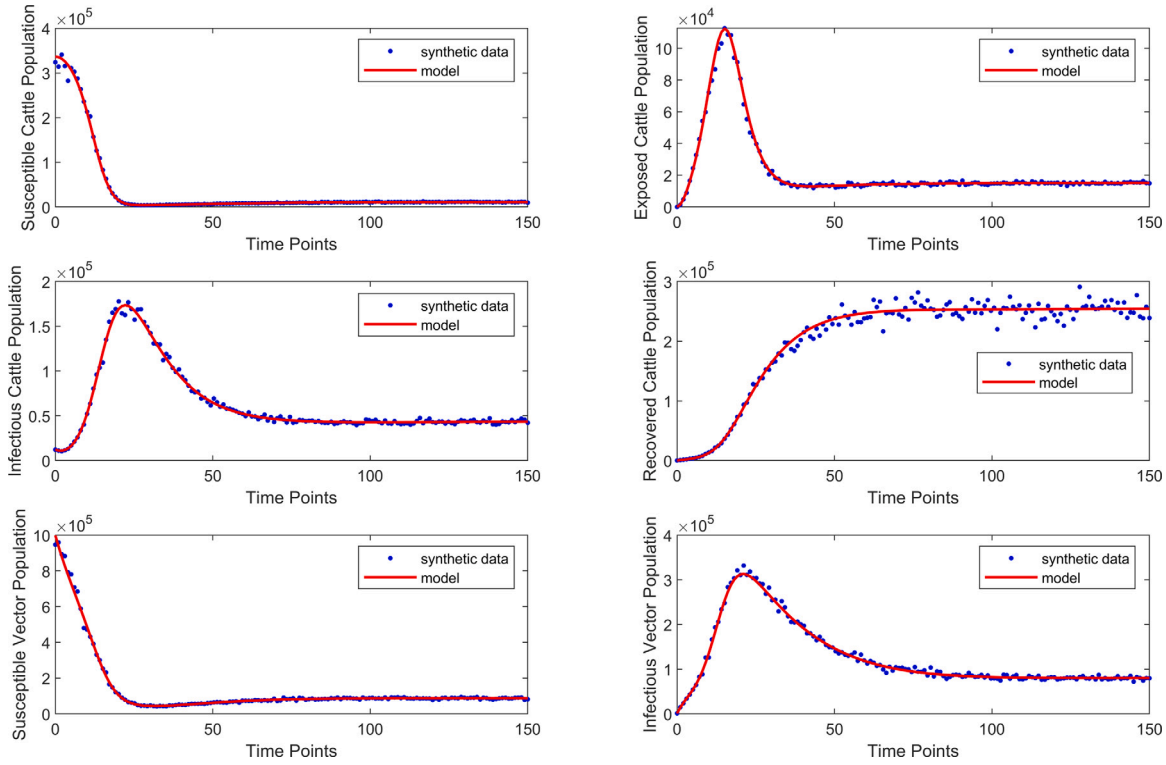


Fig. 3. Synthetic data and model solution for the infectious class in the compartmental model, displayed with 150 data points generated with a standard deviation (SD) of 0.05. The solid line represents the model solution, while the scatter points denote the generated synthetic data. The two series highlight the fit between the model prediction and the synthetic data, illustrating the variability introduced by the specified SD.

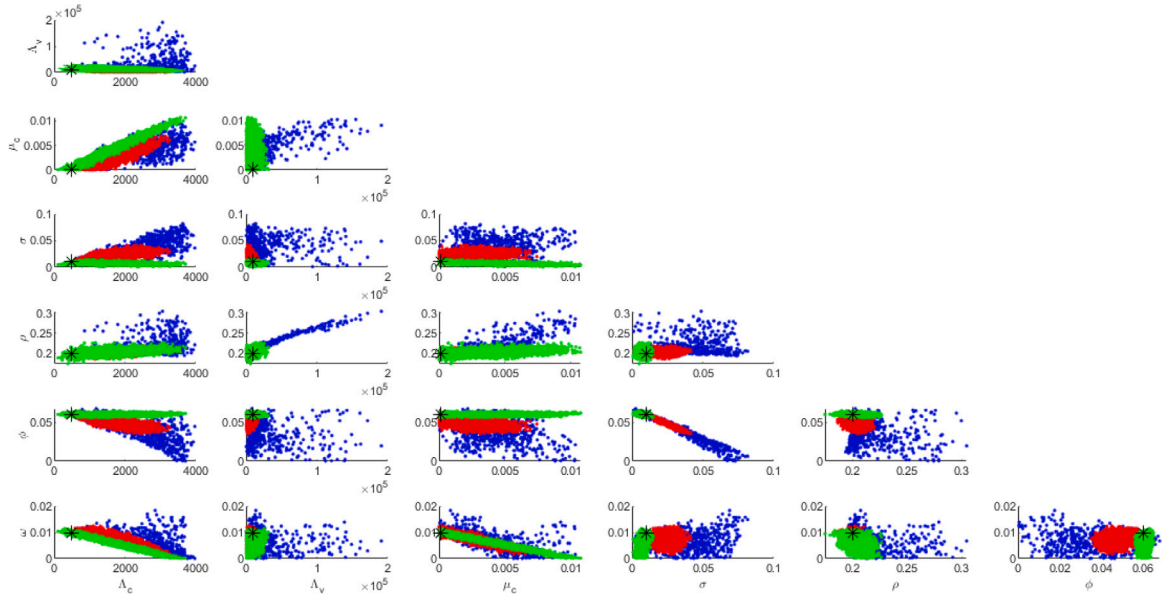


Fig. 4. Pairwise plot showing comparatively less uncertainty when data for four compartments is available versus when data for only two compartments and one compartment data is available. Blue: only data for I_c is used, red: data for S_c and I_c is used, and green: data for all cattle compartments is used. $SD = 0.05$ and 100 data points used. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Uncertainties based on the length in time for the available data. Next, we assessed how the number of data points impacts the uncertainty in parameter identification. Using the same noise level, 0.05, as previously, we specifically examined scenarios where data for two classes of cattle, Susceptible and Infectious, were available. We chose these two classes for further investigation because they are realistically measurable components, and their data are readily accessible. The results in Fig. 5 conform with the findings reported in a research by [1], demonstrating that with a longer duration of data collection, not only is uncertainty in parameter estimates reduced, but the estimates themselves tend to align very closely with their true values. However, it is important not only to take into account the number of time points but also the frequency for which the data is collected to capture the desired level of detail in modeling. For a disease like LSD whose outbreaks are less frequent, monthly data is enough to capture the dynamics of the disease. Further analysis will involve the consideration of 100 data points, covering

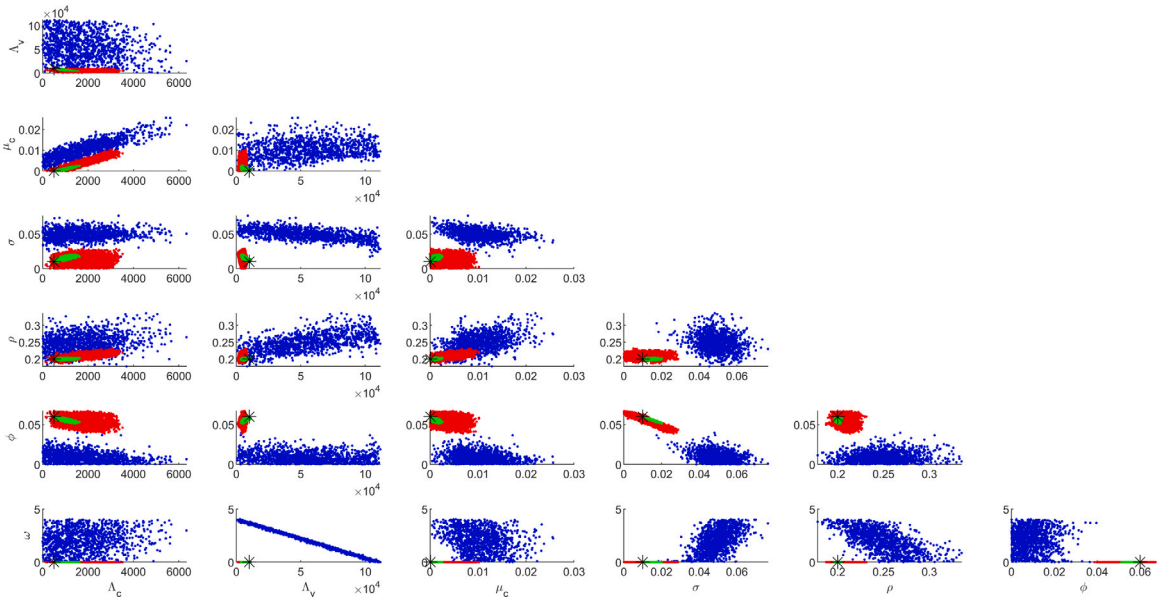


Fig. 5. Comparing uncertainty in parameter identification based on length of time of data. Blue: 30 data points, Red: 100 data points, and Green: 500 data points; $SD = 0.05$ with data for two model classes S_c and I_c used. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

approximately 8 years, a duration sufficient for capturing both seasonal fluctuations and long-term trends in disease prevalence.

Comparing uncertainty in parameter identification based on the level of noise in data. In order to compare uncertainty in parameter identification based on the level of noise in data, a scenario where data consisting of 100 time points for susceptible cattle (S_c) and infectious cattle (I_c) is examined. Recognizing that data noise is a significant source of uncertainty in model calibration [1], Fig. 6 shows that uncertainties escalate with higher levels of noise in the parameters. At lower noise levels, the uncertainty remains relatively contained, allowing for more precise and stable parameter estimates. However, as the noise level rises, the variability in the parameters introduces substantial fluctuations and widens the range of uncertainty, impacting the reliability of the model's predictions. This observed trend underscores the sensitivity of the model to noise in the parameters and highlights how high levels of stochasticity can complicate accurate estimation and introduce potential bias. Consequently, understanding and managing noise becomes crucial for achieving robust and reliable model outputs, especially in scenarios with naturally higher data variability.

Fig. 7 highlights the substantial impact that noise in data can have on parameter estimation. The results reveal a pronounced increase in variance of the estimated parameter values as noise levels rise. This heightened variance suggests that higher noise levels significantly reduce the reliability of parameter estimates, indicating as expected that data quality is crucial for accurate model predictions. Furthermore, Table 3 illustrates how the confidence intervals around parameter estimates widen as noise increases. These expanding intervals point to a growing uncertainty in the estimates, emphasizing that higher noise levels undermine the model's predictive certainty and stability.

In practical terms, various sources contribute to data noise. It can arise from multiple factors, ranging from severe inaccuracies such as extreme data recording errors to more subtle variations, like human error. Often, noise is assumed to follow a Gaussian distribution, representing random variations around true values, which simplifies modeling but may not fully capture the complexity of real-world noise. These insights underline the importance of minimizing noise where possible to enhance the precision and reliability of parameter estimates, ultimately improving model robustness and applicability in real-world scenarios.

Table 3
Ninety five percent (95%) confidence interval for parameter estimates at different noise levels showing wider intervals with higher noise levels.

Parameter	Noise level	Lower limit	Upper limit	Width
Λ_c	0.01	496	503	7
	0.05	488	511	23
	0.09	483	516	33
Λ_v	0.01	9954	10 045	91
	0.05	9922	10 077	155
	0.09	9628	10 371	743
μ_c	0.01	0.000092239189	0.00010776081	1.55e−5
	0.05	0.000067673625	0.00013232638	6.46e−5
	0.09	0.000040586175	0.00015941383	11.88e−5
σ	0.01	0.009953632797	0.01004636720	9.27e−5
	0.05	0.009762371896	0.0102376281	47.52e−5
	0.09	0.009692198424	0.01030780158	61.56e−5
ρ	0.01	0.199963549994	0.20003645001	7.29e−5
	0.05	0.199885595711	0.20011440429	22.81e−5
	0.09	0.199643734323	0.20035626568	71.25e−5
ϕ	0.01	0.059960141395	0.06003985861	7.97e−5
	0.05	0.059773062460	0.06022693754	45.39e−5
	0.09	0.059696583802	0.06030341620	60.68e−5
ω	0.01	0.009985914527	0.01001408547	2.82e−5
	0.05	0.009927463011	0.01007253699	14.51
	0.09	0.009902096758	0.01009790324	19.58

3.2. Synthetic versus real data with MCMC

Here, we conduct a parallel analysis of real and synthetic data (mimicking the features of the real data) to gain insights into the reliability and effectiveness of our modeling approach in capturing the underlying dynamics of the disease. Refer Table 4 for the initial parameters used for real data case. This parallel analysis also provide insights as to whether a different modeling approach is needed to effectively fit the data well. For the synthetic data, we simulated 30 data points for the classes S_c and I_c with noise level 0.05. The focus on the compartments S_c (susceptible cattle) and I_c (infectious cattle) stems from their epidemiological significance and the availability of reliable

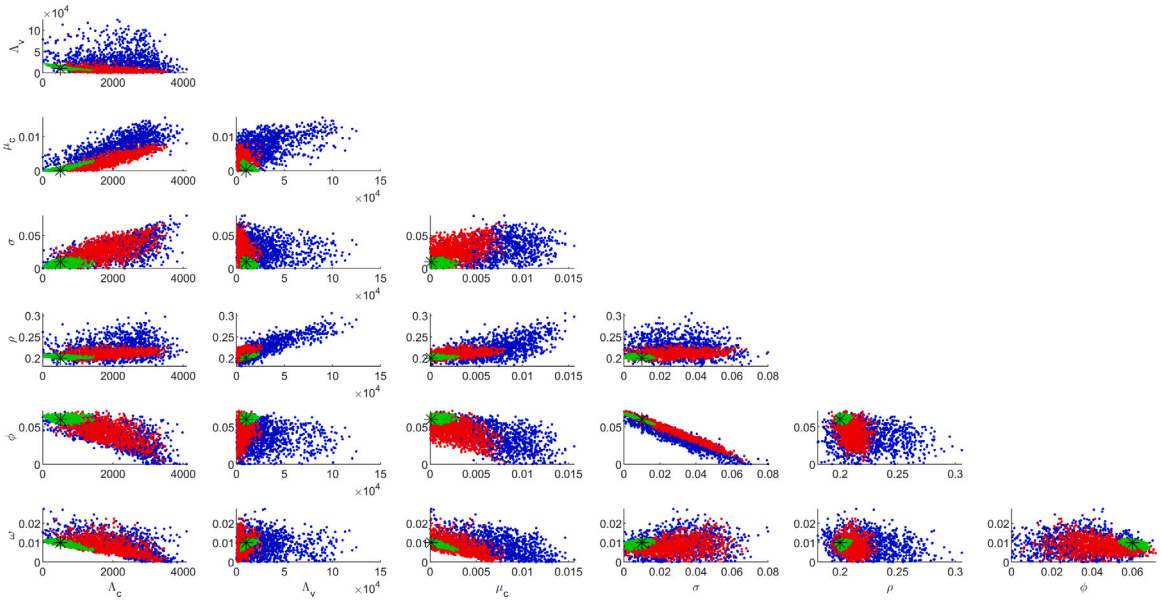


Fig. 6. Comparative uncertainty based on noise level. Blue: $\sigma = 0.09$, Red: $\sigma = 0.05$, green: $\sigma = 0.01$. A hundred (100) data points for two model classes $S_c\&I_c$ was used. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

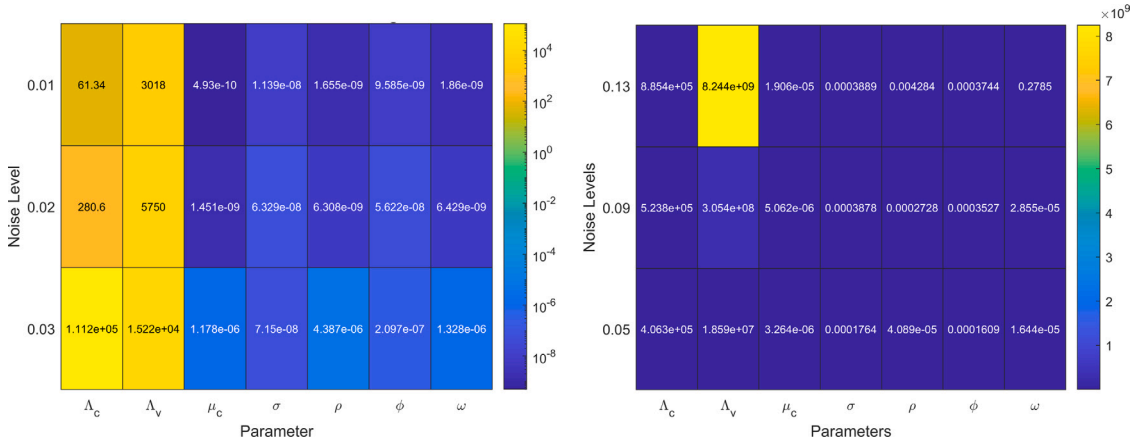


Fig. 7. Heat map showing increase in the variance between true parameter values and the estimated values with an increase in noise level. A hundred (100) data points for two model classes $S_c\&I_c$ was used.

data. These compartments are crucial for understanding the disease's transmission dynamics and are typically the most documented in field data. When mimicking realism using synthetic data, these classes were chosen to reflect the actual data availability and reliability found in field studies.

The findings offer key insights of the model when applied to real-world scenarios, as well as the role of synthetic data in evaluating model performance. Here's a detailed breakdown of the implications for real-world applications and how synthetic data contributes to the model's validation:

3.2.1. Role of synthetic data in model validation

Synthetic data, as used in this study, serves as a valuable tool for assessing the model's ability to replicate known trends under controlled conditions. With synthetic data, researchers have precise control over parameters and noise levels, allowing them to evaluate whether the model can consistently follow data trends and maintain predictive accuracy. In this case, the model using synthetic data with 12,000 samples from the MCMC results with the synthetic data (as depicted in Fig. 8, Fig. 9, and Fig. 10) samples demonstrated trend-following behavior but showed large uncertainties for certain parameters, hinting at limitations in parameter identifiability. Results using the synthetic data as shown in Fig. 5 suggest that the model could benefit from additional data to better constrain its parameters, especially for more complex and multi-parameter models. This type of validation informs model's credibility and helps ensure that, when applied to real-world data, the model is more likely to generate reliable predictions.

3.2.2. Implications for real-world data and parameter identifiability

The contrast between results using the synthetic data versus real-world data from Rukwa, Tanzania, highlights a common challenge in model-based epidemiology: real-world data often not available contain higher levels of noise and may lack the quantity and quality needed for precise parameter estimation. At the onset of an epidemic, real data are hard to come by and modeling using synthetic data is a timely option for informing quick decision-making. In the case of real data (see Fig. 11), only two parameters, Λ_c and μ_c, were well-identified, while the others remained elusive due to limited data quality or volume.

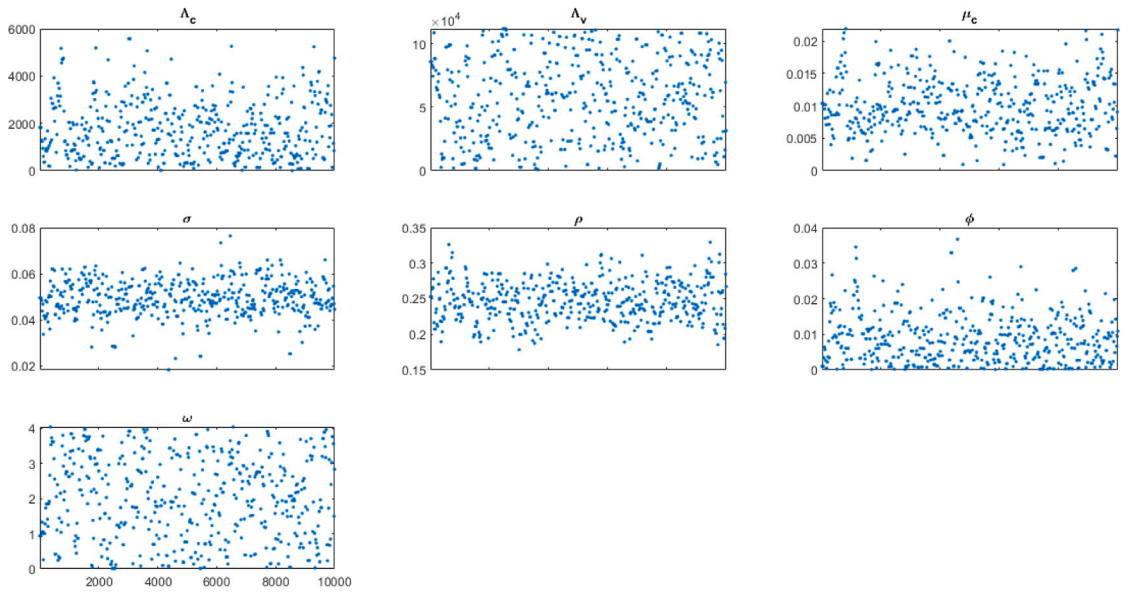


Fig. 8. Trace plot for 30 time points data.

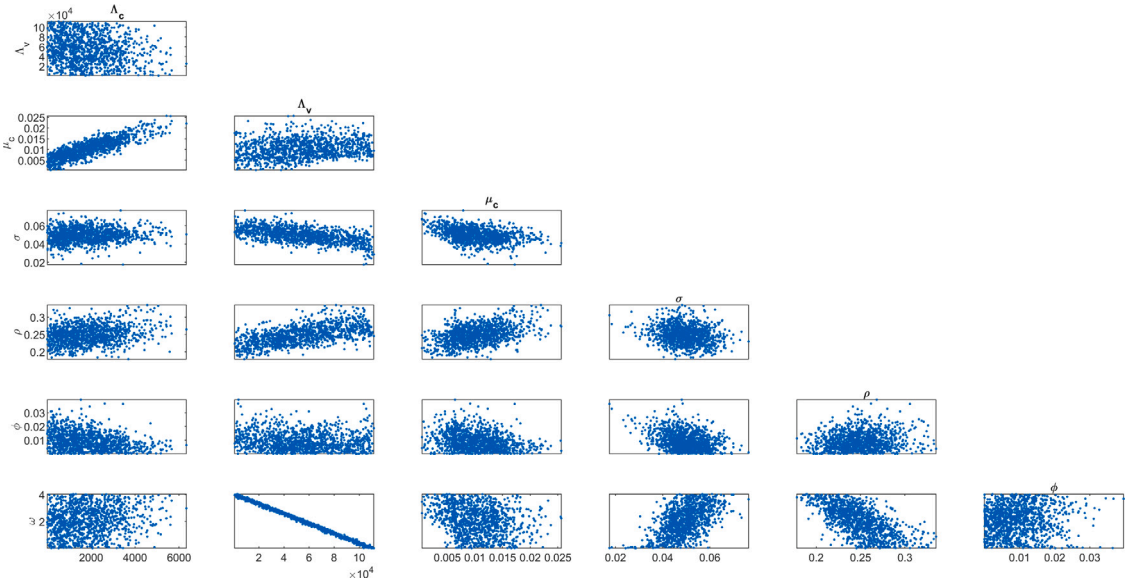


Fig. 9. Pairwise plot for 30 time points data.

The poor identifiability of parameters in real data results in broader predictive uncertainties, meaning that predictions about the spread of Lumpy Skin Disease (LSD) in Rukwa are less precise and more variable. These findings underscore the need for more comprehensive data collection, along with careful data processing, to improve model parameter estimation and enhance the model's utility in informing intervention strategies.

3.2.3. *Implications for policy and intervention*

Despite uncertainties in the model's parameters and predictions, the long-term forecasts indicate that LSD is likely to persist in Rukwa. This projection provides critical insights for public health authorities and policy-makers. Given the model's indications, there is a strong case for the urgent development and implementation of localized intervention measures aimed at reducing LSD transmission. Such strategies could include targeted vaccination campaigns, improved surveillance systems, and community education on disease prevention. Importantly, the model underscores the need for context-specific interventions, as generic or broad-scale approaches may not address localized patterns of disease transmission effectively.

Parameter identification with MCMC was done using the available data. The prediction plot indicates that the number of susceptible and infectious cattle will be declining towards a steady state (see Fig. 12). This is due to the fact that the data provided was collected after the disease had already reached an endemic stage rather than at the start of the epidemic whereby the infectious curve starts to increase exponentially. This highlights the importance of not relying on synthetic data alone when estimating model parameters. While synthetic data allow for a controlled environmental variability for model testing, real data is essential to capture the true dynamics of LSD in a specific setting. Synthetic data is often generated based on assumptions about the underlying system dynamics, which may not fully capture some of the complexities of real-world data. Its lack of real-world complexity and variability limit its ability to accurately reflect actual conditions. To understand LSD's impact accurately, parameter estimation must be grounded in real-world data, making findings more applicable and

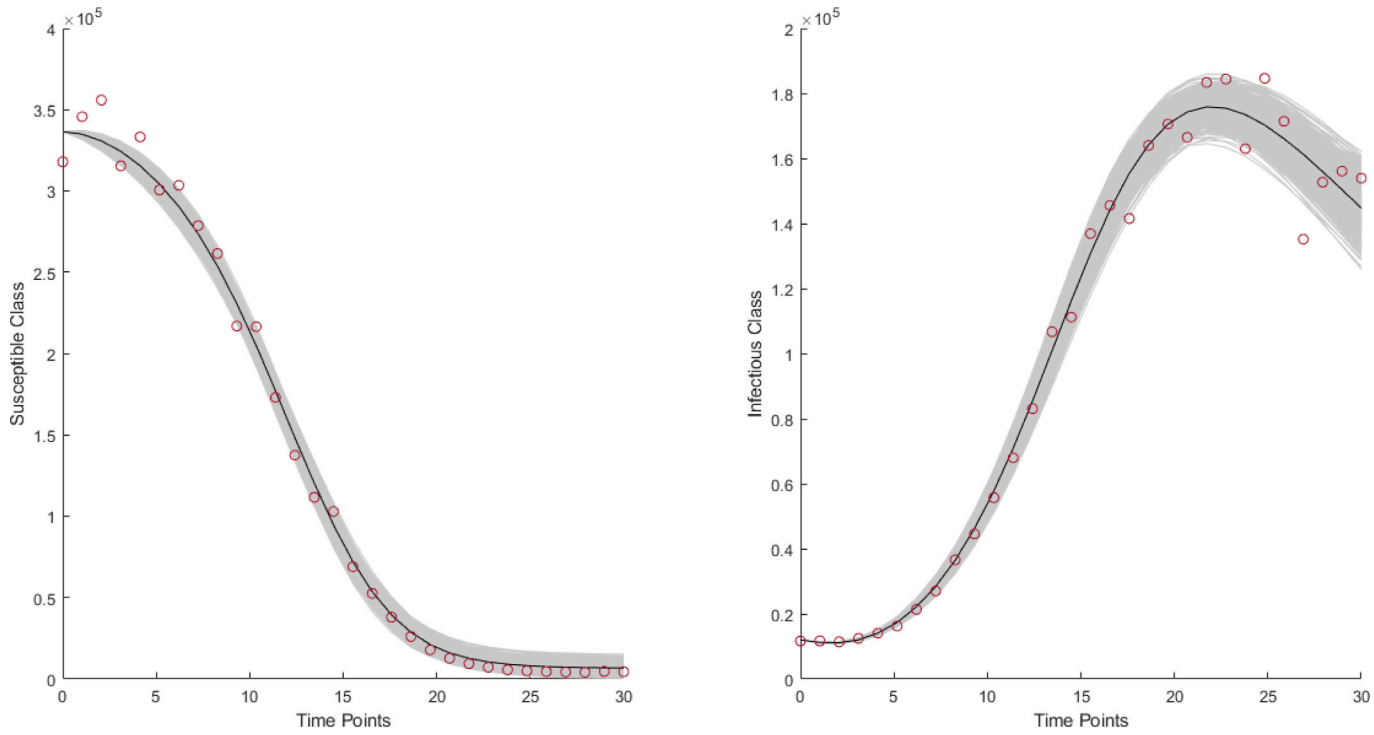
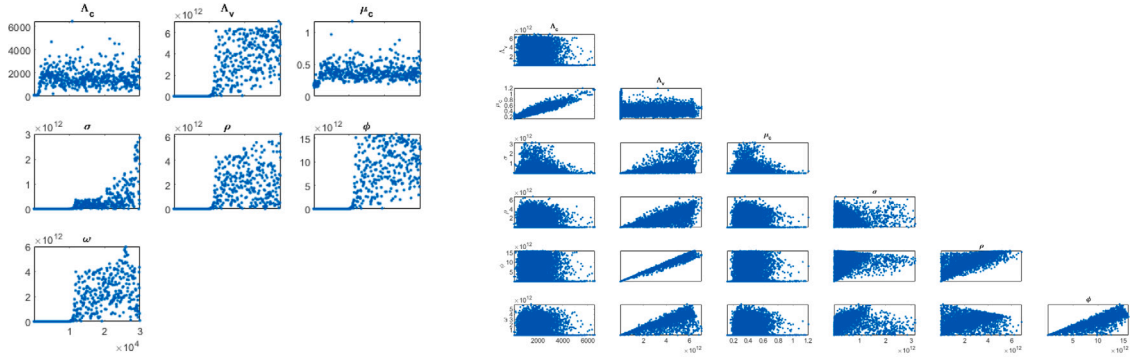


Fig. 10. Prediction plot for the model using synthetic data for a case resembling real case: Thirty (30) data time points, two variables S_c and I_c, and $SD = 0.05$.

Table 4
Model initial parameter values for Rukwa.

Parameter	Λ_c	Λ_v	ω	ρ	μ_c	μ_v	ϕ	σ	a	ν_c	ν_v
Value (Half a Year)	20	6	0.1	0	0.1	0.075	1	0.1	0.9	0.64	0.46



(a) Trace plot (b) pairwise plot

Fig. 11. MCMC results when Rukwa data is used.

valuable for actual disease management.

3.3. Convergence diagnostics of MCMC results

The convergence of the MCMC process was assessed by examining the autocorrelation plots for the parameters under investigation, both for synthetic data and real data cases. For the synthetic data case, graphical results are depicted in Fig. 13. The plots indicate high autocorrelation at short lags, which decays to zero as the lag increases. This pattern suggests that the MCMC chain is well mixed, with minimal autocorrelation between successive samples after a short period. Such behavior is indicative of a well-converged chain, ensuring that the samples are effectively independent and that the parameter estimates are reliable. The rapid decay of the autocorrelation function further supports the efficiency and convergence of the MCMC sampling process.

For the real data case as shown in Fig. 14, only two parameters, Λ_c and μ_c were well-identified and their autocorrelation is high at short lags, but it decays rapidly to zero at longer lags. This indicates that these parameters converge relatively quickly, with samples becoming independent after a short initial period. This behavior suggests that the MCMC chain for these parameters is well-mixed, and the estimates for Λ_c and μ_c can be considered reliable with minimal autocorrelation remaining after the early iterations. In contrast, the autocorrelation for the remaining parameters shows high values at short lags, but the decay to zero is much slower, indicating that these chains have not mixed as efficiently.

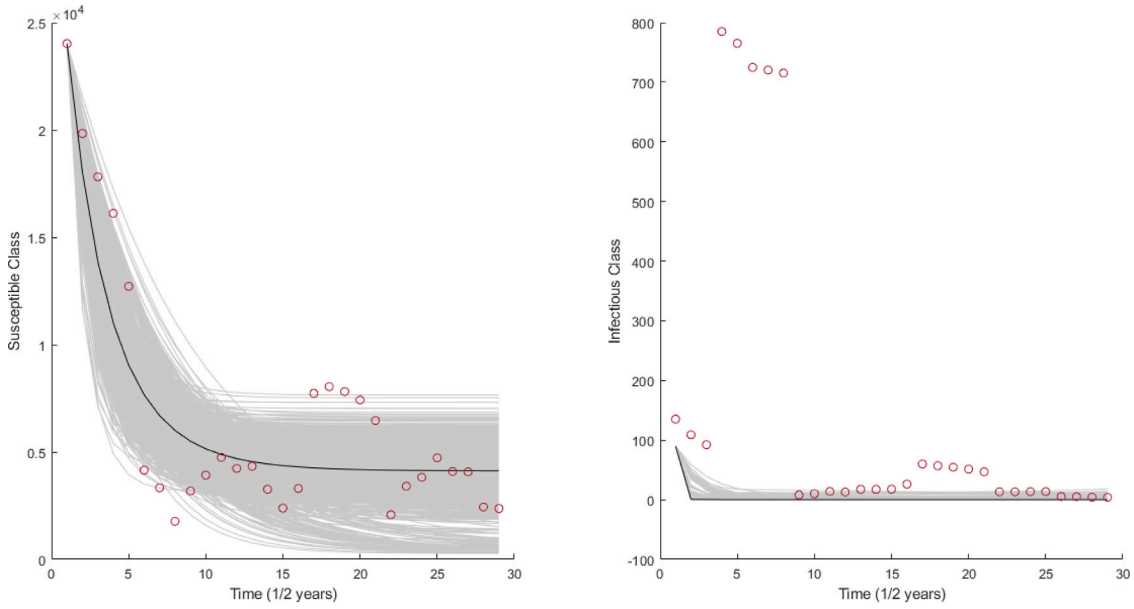


Fig. 12. Prediction plot for the model using Rukwa data.

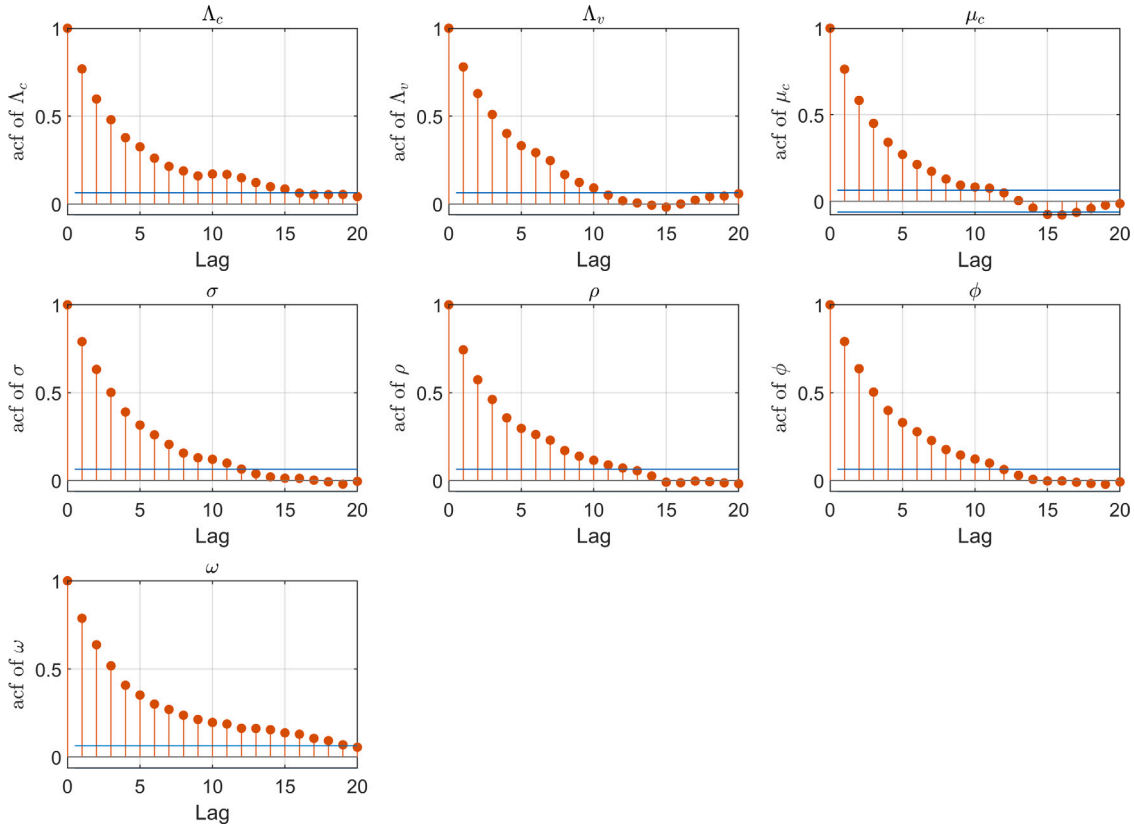


Fig. 13. Autocorrelation function of parameters in different lags showing pretty good convergence of MCMC for synthetic data case.

4. Conclusion

LSD is a viral infection with no specific antiviral treatment that affects animals, and as such only supportive/palliative care is provided to severely infected animals [46]. While LSD mortality rate is generally low, the disease has a higher morbidity rate from 3% to 85% with significant socioeconomic impact (reduction in meat consumption, leading to economic losses for farmers) in affected communities [47]. Emergence and reemergence of LSD affects animals' health, fertility, and productivity. Because by construction mathematical models which are symbolic representations of real-life situation inherit the loss of information [48], that is, there are often uncertainty around some parameter values, and investigating such uncertainties through parameter identifiability, i.e., how can the model parameters be estimated from measurements, is paramount to help build models that are usable in practice [23].

With the help of Adaptive Metropolis Hastings algorithm, a Markov Chain Monte Carlo (MCMC) standard statistical method is applied to

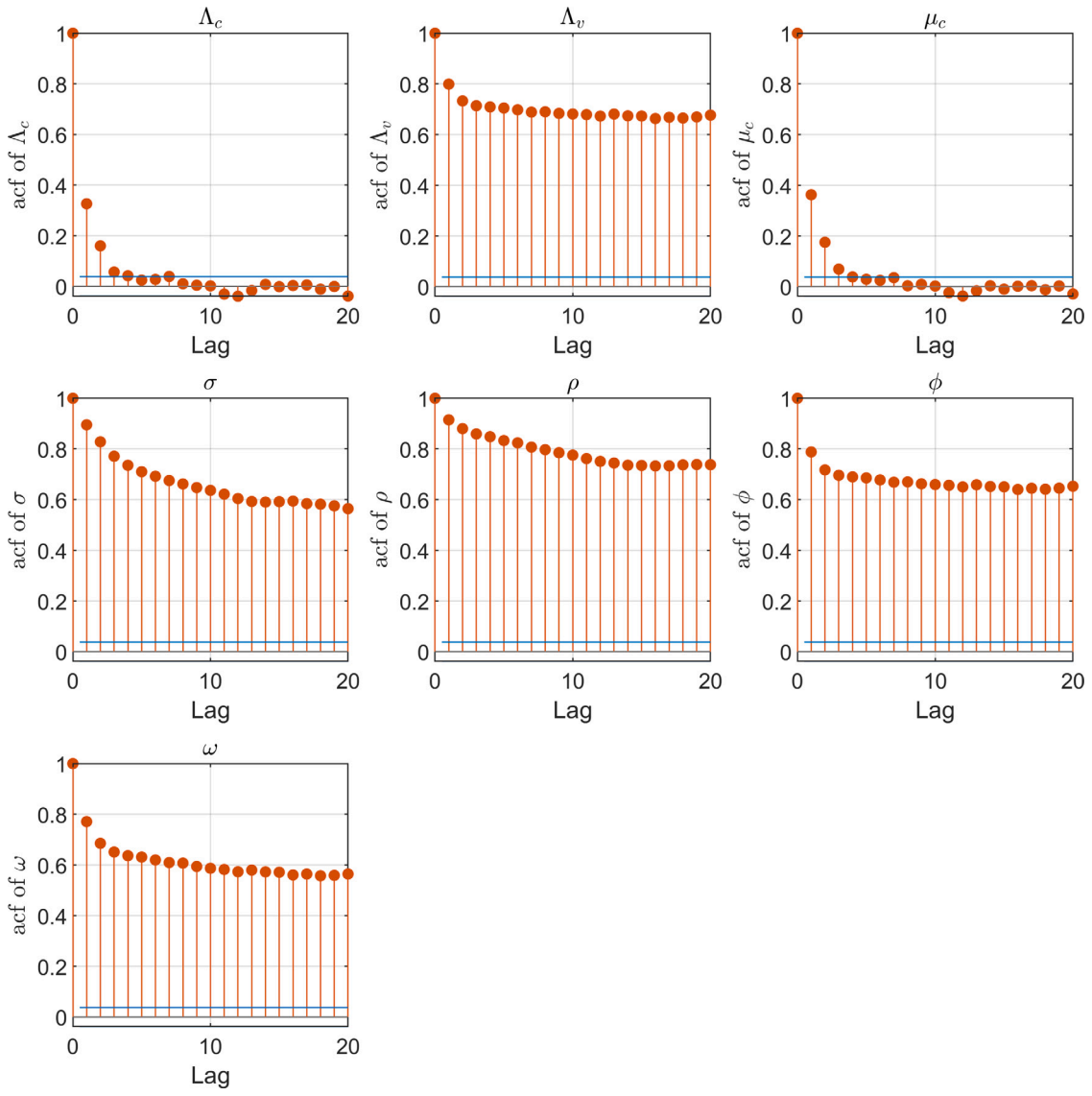


Fig. 14. Autocorrelation function of parameters in different lags showing poor convergence of MCMC for real data case. Only two parameters Λ_c and μ_c converge.

a mechanistic Susceptible–Exposed–Infectious–Recovered–Susceptible LSD model first proposed and rigorously analyzed in [15]. To this end, scenarios for which the model parameters can be well identified are investigated. The model simulations with synthetic data show that the model parameters are identifiable with a reasonable amount of synthetic noise, and enough data points spanning through the model classes. Thus, this study investigates with synthetic data how the noise and number of data points translate to the uncertainties in identifying the model parameters. A similar case with real data was also simulated. The results of the case with synthetic data largely outperformed its real data counterpart in terms of parameter identifiability and prediction. Several reasons could drive this difference (1) The biannual nature of the data may not provide sufficient frequency to accurately capture the dynamics of LSD disease transmission, and consequently monthly or even daily data collection to effectively capture the dynamics of the disease would be preferable (2) There might be issues with data accuracy, completeness, or consistency that could affect model performance. Indeed there were some missing data points that were estimated using MATLAB moving average technique for estimating missing data which may not be very effective. Also, we did not have access to information about how the data were collected in terms of consistency; whether it was from a particular farm(s) or for the whole period collected or from different farms at different time points (3) The Gaussian noise added to synthetic data presents random fluctuations with a symmetric distribution around a mean and was meant to capture measurement errors, whereas the noise in the real data had a wide range of anomalies and errors such as recording errors, outliers, systemic biases, and missing values among others (4) The model may be structurally limited to capture real dynamic features in the data, and to fit the data well. This is typical of most structural dynamic models. In this case, the use of a data-driven approach using the Kalman filter likelihood as discussed in [10,49] should be considered. Using this data-driven approach can enable long-term forecasting of the disease dynamics [50,51], even when the model is biased by improving the model fit, and distinct patterns could be identified with high accuracy. Also, incorporating regularization to manage noise and variability in real-world data could be viable.

This study is not exhaustive as it has some limitations. Only data uncertainty was considered herein. There are other categories of uncertainties that could potentially be accounted for in modeling studies such as stochastic uncertainty (derived from the method of simulation), and structural uncertainty (in the optimal model structure). We focused on Gaussian noise for simplicity and general applicability. However, we acknowledge that exploring other noise types, such as Poisson

noise, could yield different insights, especially for count data typical in epidemiological studies. Despite the fact that the availability of LSD vaccines could be limited in some settings, thereby making the control of LSD challenging in such areas, a follow-up study could incorporate a vaccinated class into the model. Also, adding both quarantine and treatment classes could provide additional insights into the model dynamics. While there is no definitive treatment, accounting for a treated class is at least from the mathematical standpoint worth considering. Finally, as more data become available, the model can first be fit to the real data, its parameter values estimated and used in lieu of the parameters values from the literature.

CRediT authorship contribution statement

Edwiga Renald: Writing – original draft, Methodology, Formal analysis, Conceptualization. **Miracle Amadi:** Writing – review & editing, Supervision. **Heikki Haario:** Writing – review & editing, Supervision. **Joram Buza:** Writing – review & editing, Supervision. **Jean M. Tchuenche:** Writing – review & editing, Supervision. **Verdiana G. Masanja:** Writing – review & editing, Supervision.

Declaration of competing interest

All authors declare no competing interests that are relevant to the content of this article.

Acknowledgments

ER acknowledges with thanks - Lappeenranta University of Technology for the invitation and hospitality, and extends special thanks to Professor Matti Heilio for organizing and coordinating her visit, which contributed to the success of this work - financial support of the International Mathematics Union (IMU) and Higher Education for Economic Transformation (HEET) project towards her PhD studies.

References

[1] G. Chowell, Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts, Infect. Dis. Model. 2 (2017) 379–398, http://dx.doi.org/10.1016/j.idm.2017.08.001.

[2] S.K. Kiplagat, P.M. Kitala, J.O. Onono, P.M. Beard, N.A. Lyons, Risk factors for outbreaks of Lumpy Skin Disease and the economic impact in cattle farms of Nakuru county, Kenya, Front. Vet. Sci. 7 (2020) 259, http://dx.doi.org/10.3389/fvets.2020.00259.

[3] A. Issimov, N. Rametov, K. Zhugunissov, L. Kutumbetov, A. Zhanabayev, N. Kazhgaliyev, B. Nurgaliyev, M. Shalmenov, G. Absatirov, L. Dushayeva, Emergence of the first Lumpy Skin Disease outbreak among livestock in the Republic of Kazakhstan in 2016, 2020, http://dx.doi.org/10.20944/preprints202011.0298.v1.

[4] C. Jiang, D. Tao, Y. Geng, H. Yang, B. Xu, Y. Chen, C. Hu, H. Chen, S. Xie, A. Guo, Sensitive and specific detection of lumpy skin disease virus in cattle by crispr-cas12a fluorescent assay coupled with recombinase polymerase amplification, Genes 13 (2022) 734, http://dx.doi.org/10.3390/genes13050734.

[5] W. Molla, K. Frankena, M. De Jong, Transmission dynamics of lumpy skin disease in Ethiopia, Epidemiol. Infect. 145 (2017) 2856–2863, http://dx.doi.org/10.1017/S0950268817001637.

[6] A.H. Briggs, M.C. Weinstein, E.A. Fenwick, J. Karnon, M.J. Sculpher, A.D. Paltiel, I.-S.M.G.R.P.T. Force, et al., Model parameter estimation and uncertainty: a report of the ispor-smdm modeling good research practices task force-6, Value Heal. 15 (2012) 835–842, http://dx.doi.org/10.1016/j.jval.2012.04.014.

[7] E.K. Renald, D. Kuznetsov, K. Kreppel, Desirable dog-rabies control methods in an urban setting in Africa-A mathematical model, Int. Math. Sci. Comput. 6 (2020) http://dx.doi.org/10.5815/ijmsc.2020.01.05.

[8] E. Renald, D. Kuznetsov, K. Kreppel, Sensitivity analysis and numerical simulation of a SEIV basic dog-rabies mathematical model with control, Int. J. Adv. Sci. Res. Eng. (2019) http://dx.doi.org/10.31695/IJASRE.2019.33526.

[9] C.I. Siettos, L. Russo, Mathematical modeling of infectious disease dynamics, Virulence 4 (2013) 295–306, http://dx.doi.org/10.4161/viru.24041.

[10] M. Amadi, H. Haario, Parameter identification and forecast with a biased model, in: European Consortium for Mathematics in Industry, Springer, 2021, pp. 227–232, http://dx.doi.org/10.1007/978-3-031-11818-0_30.

[11] J. Vanlier, C. Tiemann, P. Hilbers, N. Van Riel, Parameter uncertainty in biochemical models described by ordinary differential equations, Math. Biosci. 246 (2013) 305–314, http://dx.doi.org/10.1016/j.mbs.2013.03.006.

[12] H. Thorén, P. Gerlee, Model uncertainty, the covid-19 pandemic, and the science-policy interface, R. Soc. Open Sci. 11 (2024) 230803, http://dx.doi.org/10.1098/rsos.230803.

[13] G. Chowell, A. Bleichrodt, R. Luo, Parameter estimation and forecasting with quantified uncertainty for ordinary differential equation models using quantdiffforecast: A matlab toolbox and tutorial, Stat. Med. 43 (2024) 1826–1848, http://dx.doi.org/10.1002/sim.10036.

[14] M. Gu, Y. Lin, V.C. Lee, D.Y. Qiu, Probabilistic forecast of nonlinear dynamical systems with uncertainty quantification, Phys. D 457 (2024) 133938, http://dx.doi.org/10.1016/j.physd.2023.133938.

[15] E. Renald, V.G. Masanja, J.M. Tchuenche, J. Buza, A deterministic mathematical model with non-linear least squares method for investigating the transmission dynamics of lumpy skin disease, Heal. Anal. 5 (2024) 100343, http://dx.doi.org/10.1016/j.health.2024.100343.

[16] A. El-Mesady, A. Elsadany, A. Mahdy, A. Elsonbaty, Nonlinear dynamics and optimal control strategies of a novel fractional-order lumpy skin disease model, J. Comput. Sci. 79 (2024) 102286, http://dx.doi.org/10.1016/j.jocs.2024.102286.

[17] A. Elsonbaty, M. Alharbi, A. El-Mesady, W. Adel, Dynamical analysis of a novel discrete fractional lumpy skin disease model, Partial. Differ. Equ. Appl. Math. 9 (2024) 100604, http://dx.doi.org/10.1016/j.padiff.2023.100604.

[18] O. Falowo, J. Owolabi, O. Oludoun, R. Akingbade, Mathematical modelling of lumpy skin disease in dairy cow, in: IOP Conference Series: Earth and Environ. Sci., Vol. 1219, IOP Publishing, 2023, p. 012007, http://dx.doi.org/10.1088/1755-1315/1219/1/012007.

[19] WOAH, Frequently asked questions (FAQ) on Lumpy Skin Disease (LSD), 2022, https://www.woah.org/en/document/frequently-asked-questions-faq-on-lumpy-skin-disease-lsd/. (Accessed 13 January 2023).

[20] B. Mat, M.S. Arikan, A.C. Akin, M.B. Çvrimli, H. Yonar, M.A. Tekindal, Determination of production losses related to lumpy skin disease among cattle in Turkey and analysis using SEIR epidemic model, BMC Vet. Res. 17 (2021) 1–10, http://dx.doi.org/10.1186/s12917-021-02983-x.

[21] R. Musa, O.J. Peter, F.A. Oguntolu, A non-linear differential equation model of covid-19 and seasonal influenza co-infection dynamics under vaccination strategy and immunity waning, Heal. Anal. 4 (2023) 100240, http://dx.doi.org/10.1016/j.health.2023.100240.

[22] S. Pedro, H. Rwezaura, J. Tchuenche, Time-varying sensitivity analysis of an influenza model with interventions, Int. J. Biomath. 15 (2022) 2150098, http://dx.doi.org/10.1142/S1793524521500984.

[23] A. Solonen, H. Haario, J.M. Tchuenche, H. Rwezaura, Studying the identifiability of epidemiological models using mcmc, Int. J. Biomath. 6 (2013) 1350008, http://dx.doi.org/10.1142/S1793524513500083.

[24] L. D'Agostino, K. McGowan, H. Grantz, E. Murray, Quantifying uncertainty in mechanistic models of infectious disease, Am. J. Epidemiol. 190 (2021) 1377–1385, http://dx.doi.org/10.1093/aje/kwab013.

[25] S. Gubbins, Using the basic reproduction number to assess the risk of transmission of lumpy skin disease virus by biting insects, Transbound. Emerg. Dis. 66 (2019) 1873–1883, http://dx.doi.org/10.1111/tbed.13216.

[26] P. Ratan, L. Rubbi, M. Thompson, K. Naresh, J. Waddell, B. Jones, M. Pellegrini, Epigenetic aging in cows is accelerated by milk production, Epigenetics 18 (2023) 2240188, http://dx.doi.org/10.1080/15592294.2023.2240188.

[27] F. Brauer, C. Castillo-Chavez, C. Castillo-Chavez, Mathematical models in population biology and epidemiology, vol. 2, Springer, 2012, http://dx.doi.org/10.1007/978-1-4614-1686-9.

[28] S. Abutarbush, M. Ababneh, I. Al Zoubi, O. Al Sheyab, M. Al Zoubi, M. Alekish, R. Al Gharabat, Lumpy skin disease in Jordan: Disease emergence, clinical signs, complications and preliminary-associated economic losses, Transbound. Emerg. Dis. 62 (2015) 549–554, http://dx.doi.org/10.1111/tbed.12177.

[29] E.S. Tuppurainen, S. Babiuk, E. Klement, Lumpy Skin Disease, Springer, 2018, http://dx.doi.org/10.1007/978-3-319-92411-3.

[30] B. Datten, A.A. Chaudhary, S. Sharma, L. Singh, K.D. Rawat, M.S. Ashraf, L.M. Alneghery, M.O. Aladwani, H.A. Rudayni, D. Dayal, et al., , an extensive examination of the warning signs, symptoms, diagnosis, available therapies, and prognosis for lumpy skin disease, Viruses 15 (2023) 604, http://dx.doi.org/10.3390/v15030604.

[31] EFSA.P. Calistri, K. De Clercq, S. Gubbins, E. Klement, A. Stegeman, J. Cortiñas Abrahantes, D. Marojevic, S.-E. Antoniou, A. Broglia, Lumpy skin disease epidemiological report iv: data collection and analysis, Efsa J. 18 (2020) e06010, http://dx.doi.org/10.2903/j.efsa.2020.6010.

[32] WOAH, Quantitative data dashboard, 2024, https://wahis.woah.org/#/dashboards/qd-dashboard. (Accessed 5 January 2023).

[33] MLF, Southern highlands a strategic area for dairy cattle breeding, 2024, https://www.itv.co.tz/news/nyanda-za-juu-kusini-eneo-la-kimkakati-kwa-ufugaji-ngombe-wa-maziwa. (Accessed 25 April 2024).

[34] H. Haario, E. Saksman, J. Tamminen, An adaptive metropolis algorithm, Bernoulli 7 (2001) 223–242.

[35] G.O. Roberts, J.S. Rosenthal, Examples of adaptive mcmc, J. Comput. Gr. Stat. 18 (2009) 349–367, http://dx.doi.org/10.1198/jcgs.2009.06134.

[36] G.G. Mwanga, Mathematical modeling and optimal control of malaria, 2014, URL https://urn.fi/URN:ISBN:978-952-265-720-6.

[37] L. Wang, L. Wang, M. Georgieva, P. Machado, A. Ulagappa, S. Ahmed, Y. Lu, A. Bakshi, F. Ghassemi, Robust nonparametric distribution forecast with backtest-based bootstrap and adaptive residual selection, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2022, pp. 3903–3907, http://dx.doi.org/10.1109/ICASSP43922.2022.9747701.

[38] M. Shin, H. Cho, H. s. Min, S. Lim, Neural bootstrapper, Adv. Neural Inf. Process. Syst. 34 (2021) 16596–16609, http://dx.doi.org/10.48550/arXiv.2010.01051.

[39] G. Meinrath, C. Ekberg, A. Landgren, J. Liljenzin, Assessment of uncertainty in parameter evaluation and prediction, Talanta 51 (2000) 231–246, http://dx.doi.org/10.1016/S0039-9140(99)00259-3.

[40] R.K. Crump, N. Gospodinov, I. Lopez Gaffney, A jackknife variance estimator for panel regressions, 2024, http://dx.doi.org/10.59576/sr.1133, FRB of New York Staff Report.

[41] N. Amann, H. Leeb, L. Steinberger, Assumption-lean conditional predictive inference via the jackknife and the jackknife+, 2023, http://dx.doi.org/10.48550/arXiv.2312.14596, arXiv preprint arXiv:2312.14596.

[42] D. Luengo, L. Martino, M. Bugallo, V. Elvira, S. Särkkä, A survey of Monte Carlo methods for parameter estimation, EURASIP J. Adv. Signal Process. 2020 (2020) 1–62, http://dx.doi.org/10.1186/s13634-020-00675-6.

[43] E. Laloy, J.A. Vrugt, High-dimensional posterior exploration of hydrologic models using multiple-try dream (zs) and high-performance computing, Water Resour. Res. 48 (2012) http://dx.doi.org/10.1029/2011WR010608.

[44] A. Benson, N. Friel, Adaptive mcmc for multiple changepoint analysis with applications to large datasets, 2018, http://dx.doi.org/10.1214/18-EJS1418.

[45] K.M. Banner, K.M. Irvine, T.J. Rodhouse, The use of bayesian priors in ecology: The good, the bad and the not great, Method Ecol. Evol. 11 (2020) 882–889, http://dx.doi.org/10.1111/2041-210X.13407.

[46] E. Renald, J. Buza, J.M. Tchuenche, V.G. Masanja, The role of modeling in the epidemiology and control of lumpy skin disease: a systematic review, Bull. Natl. Res. Cent. 47 (2023) 141, http://dx.doi.org/10.1186/s42269-023-01111-z.

[47] S.E. Saqib, M. Yaseen, S. Visetnoi, Sikandar, S. Ali, Epidemiological and economic consequences of lumpy skin disease outbreaks on farm households in Khyber Pakhtunkhwa, Pakistan, Front. Vet. Sci. 10 (2023) 1238771, http://dx.doi.org/10.3389/fvets.2023.1238771.

[48] H. Rwezaura, S. Tchoumi, J. Tchuenche, Impact of environmental transmission and contact rates on covid-19 dynamics: A simulation study, Inf. Med. Unlocked 27 (2021) 100807, http://dx.doi.org/10.1016/j.imu.2021.100807.

[49] M. Amadi, H. Haario, Parameter estimation and forecasting for biased models, Handb. Vis. Exp. Comput. Math. (2023) http://dx.doi.org/10.1007/978-3-030-93954-0_36-1.

[50] J.C. Lauffenburger, M. Mahesri, N.K. Choudhry, Use of data-driven methods to predict long-term patterns of health care spending for medicare patients, JAMA Netw. Open 3 (2020) e2020291, http://dx.doi.org/10.1001/jamanetworkopen.2020.20291.

[51] T. Krivec, J. Kocijan, M. Perne, B. Grašic, M.Z. Božnar, P. Mlakar, Data-driven method for the improving forecasts of local weather dynamics, Eng. Appl. Artif. Intell. 105 (2021) 104423, http://dx.doi.org/10.1016/j.engappai.2021.104423.