In Silico Modelling of Parasite Dynamics

Clement Twumasi

School of Mathematics

Cardiff University

A thesis submitted in partial fulfilment of the requirements for the degree of *Doctor of Philosophy*

2022

Abstract

Understanding host-parasite systems are challenging if biologists employ just the experimental approaches adopted, whereas mathematical models can help uncover other in-depth knowledge about host infection dynamics. Previous experimental studies have explored the infrapopulation dynamics of *Gyrodactylus turnbulli* and *G. bullatarudis* ectoparasites on their fish host, *Poecilia reticulata*. However, other important and open biological questions exist concerning parasite microhabitat preference, host survival, parasite virulence, and the transmission dynamics of different *Gyrodactylus* strains across different host populations over time. This thesis mathematically investigates these relevant biological questions to understand the gyrodactylid-fish system's complexity better using a sophisticated multi-state Markov model (MSM) and a novel individual-based stochastic simulation model. The infection dynamics of three different gyrodactylid strains are compared across three different host populations. A modified approximate Bayesian computation (ABC) with sequential Monte Carlo (SMC) and sequential importance sampling (SIS) is developed for calibrating the novel stochastic model based on existing empirical data and an auxiliary stochastic model. In addition, an extended local-linear regression (with $L2$ regularisation) for ABC post-processing analysis has been proposed. Advanced statistics and an MSM are used to assess spatial-temporal parasite dynamics. A linear birth-death process with catastrophic extinction (B-D-C process) is considered the auxiliary model for the complex simulation model to refine the modified ABC's summary statistics, with other theoretical justifications and parameter estimation techniques of the B-D-C process provided. The B-D-C process simulation using τ-leaping also provides additional insights on accelerating the complex simulation model by proposing a reasonable error threshold based on the trade-off between simulation accuracy and computational speed. The mathematical models can be extended and adapted for other host-parasite systems, and the modified ABC methodologies can also aid in efficiently calibrating other multi-parameter models with a high-dimensional set of correlating or independent summary statistics.

Acknowledgements

Many people have contributed in diverse ways to enable me to achieve success in PhD studies and the completion of this thesis. I want to express my heartfelt appreciation to my academic supervisors, Prof. Owen Jones and Prof. Joanne Cable, who read, scrutinized and provided the necessary guidance which enlightened me to complete this work. I would also like to thank Prof. Colin Riordan, Cardiff University Vice-Chancellor, for my fully-funded PhD degree as a Vice-Chancellor's scholar of Cardiff University. A special thanks also goes to Prof. Nikolai Leonenko and Dr. Andrey Pepelyshev for their enormous supervisory support.

Additionally, I appreciate the contributions of Dr. Geraint Palmer, who served as an external mentor to have a further outlet of advice unrelated to my PhD work. I would like to also thank Dr. Tracey England for her support as her postgraduate teaching assistant in Time Series Analysis and Forecasting module for MSc Mathematics students. My profound gratitude goes to the maths support coordinators (Dr. Robert Wilson and Dr. Ana Ros Camacho) as one of the postgraduate tutors for Cardiff University's maths support.

I also appreciate the enormous support of the School of Mathematics for their welcoming and supportive environment; with special thanks to Dr. Iskander Aliev, Dr. Kirstin Strokorb, Tracey Lintern and my PhD colleagues.

Furthermore, I would like to thank my research group at the School of Biosciences: Dr. Numair Masud, Rhiannon Hunt, Alice Jones and Elissavet Arapi, who assisted with data collection and made my stay in Cardiff a memorable one. Finally, my gratitude goes to my family for their prayers, support, and motivations throughout my PhD studies. I am grateful to all loved ones who contributed their quota in diverse ways to my success.

Dissemination of work

Submitted manuscript for publication (under peer review)

- 2022: Spatial and temporal parasite dynamics: microhabitat preferences and infection progression of two co-infecting gyrodactylids. Twumasi C., Jones O., and Cable J. *Parasites & Vectors Journal.*

Presentations

- 2019: **Comparative modelling of parasite population dynamics of two *Gyrodactylus* species (conference talk)**. *Stochastic Modelling in Health and Disease Conference at School of Mathematics, Leeds University, UK.* matml.github.io/smhd2019.html.

- **Comparative modelling of parasite population dynamics of two *Gyrodactylus* species (poster)**. *Cardiff School of Mathematics* (favourite presenter).

- 2020: **Three-minute thesis competition (presenter)**. *Cardiff University Doctoral Academy, 2020.* www.youtube.com/watch?v=KXt5RjKfQ_w&t=17s.

- 2020: **Comparative modelling of parasite population dynamics of three *Gyrodactylus* species (poster)**. *Virtual Heidelberg Laureate conference 2020* (selected among 200 most exceptional young mathematicians and computer scientists worldwide). www.heidelberg-laureate-forum.org/.

- 2020: **Cambridge University Plus Magazine interview (interviewee)**. *Virtual Heidelberg Laureate conference 2020.* www.youtube.com/watch?v=niXb5GwqHWc&t=2s.

- 2020: **Spatial and Temporal Parasite Dynamics: Microhabitat preferences and infection progression of *Gyrodactylus turnbulli* and *G. bullatarudis* (PGR Talk)**. *SIAM-IMA Student Chapter Cardiff.* ostlert.github.io/SSC-Cardiff/PGRtalks/.

- 2020: **Image of Research Competition (presenter)**. *Cardiff University Doctoral Academy*. twitter.com/ClementMetal/status/1336375913647394817?s=20.

- 2021: **Birth-death process with catastrophic extinction: analytical proof and model fitting via hybrid τ-leaping stochastic simulations (PGR Talk)**. *SIAM-IMA Student Chapter Cardiff, Cardiff University*. ostlert.github.io/SSC-Cardiff/PGRtalks/.

- 2021: ***In silico* modelling of parasite dynamics (3MT competition, presenter)**. *Cardiff School of Mathematics*. twitter.com/CardiffSIAM_IMA/status/1374320407625498630.

- 2021: ***In silico* modelling of parasite dynamics (Early Careers Water Talk: Fish & Furious, presenter)**. *Water Research Institute, Cardiff University*.

Other unrelated high-impact publications over the PhD period

- 2019: **Markov Chain Modelling of HIV, Tuberculosis, and Hepatitis B Transmission in Ghana**. Twumasi C., Asiedu L., and Nortey E.N. *Interdisciplinary Perspectives on Infectious Diseases* [298].

- 2019: **Statistical Modeling of HIV, Tuberculosis, and Hepatitis B Transmission in Ghana**. Twumasi C., Asiedu L., and Nortey E. N. *Canadian Journal of Infectious Diseases and Medical Microbiology* [299].

- 2020: **Modelling the Transmission Dynamics of Tuberculosis in the Ashanti Region of Ghana**. Mettle F. O., Osei Affi P., and Twumasi C. *Interdisciplinary Perspectives on Infectious Diseases* [218].

- 2021: **An Experimental Study of Lesions Observed in Bog Body Funerary Performances**. Treadway T. and Twumasi C. *Experimental Archaeology Journal* [297].

- 2021: **Machine Learning Algorithms for Forecasting and Backcasting Blood Demand Data with Missing Values and Outliers: A Study of Tema General Hospital of Ghana**. Twumasi C. and Twumasi J. *International Journal of Forecasting* [300].

Contents

List of Figures

List of Tables

List of Abbreviations

ABC: Approximate Bayesian Computation

ABC-SMC: Sequential Monte Carlo ABC

ABC-MCMC: Markov Chain Monte Carlo ABC

B-D-C: Birth-death process with catastrophic extinction

CTMC: Continuous-time Markov chain

Gb: Wild *Gyrodactylus bullatarudis* strain

GMM: Generalised method of moments

Gt: Wild *Gyrodactylus turnbulli* strain

Gt3: Laboratory-bred *Gyrodactylus turnbulli* strain

GW: Galton-Watson estimation

HR: Hazard ratio

HTL: Hybrid tau-leaping algorithm

IBM: Individual-based model

LA: Lower Aripo River fish

MCMC: Markov Chain Monte Carlo

MKW: Multivariate Kruskal-Wallis test

MLE: Maximum likelihood estimation

MSM: Multi-state Markov Model

OS: Ornamental stock

PBM: Population-based model

PPM: Predator-prey model

SIS: Sequential Importance Sampling

SMC: Sequential Monte Carlo

SSA: Stochastic simulation algorithm

UA: Upper Aripo River fish

UKW: Univariate Kruskal-Wallis test

WRID: Weighted Ridge Regression

Chapter 1

Introduction

1.1 General overview of the study

This thesis is an interdisciplinary research (between the Cardiff University School of Mathematics and School of Biosciences), primarily focusing on modelling infection dynamics of *Gyrodactylus* parasites on freshwater fish within at least a standard 17-day experimental period. The gyrodactylid-fish system, like other host-parasite systems, is widely used to investigate ecological, evolutionary, and epidemiological problems. They are particularly amenable to experimental manipulation (for instance, controlled infections on a single fish), and there is an agent-based simulation model of these dynamics (with its limitations together with other specific research problems of mathematical and biological importance presented under section 1.3). The current study demonstrates the use of novel mathematical and stochastic simulation models to add to our understanding of the gyrodactylid infection dynamics of three different parasite strains (with two strains of *Gyrodactylus turnbulli* and one strain of *G. bullatarudis*) across three different fish populations (Ornamental stock, Lower Aripo River, and Upper Aripo River fish) over time. Like other biological systems and modelling problems (as reviewed in Chapter 3), the mathematical modelling of the gyrodactylid-fish system also requires an in-depth understanding of the biology of the *Gyrodactylus* ectoparasites (as presented in sections 1.4 and 1.5) as well as expert knowledge about their infection dynamics on the fish host with the help of experimental empirical data (presented and thoroughly analysed in Chapter 2).

For the first time in this study, we have developed a multi-state Markov model (MSM), which extends the standard survival models, to investigate open biological questions concerning host survival and parasite virulence. In the MSM, we also incorporate population heterogeneity and the effect of censoring based on the entire infection history of each host until the end of the observation period. A brief motivation for adapting the MSM for the gyrodactylid-fish system compared to the traditional survival models is highlighted in section 1.3 and discussed further under Chapter 2. Additionally, the limitations of the existing agent-based model for the gyrodactylid-fish system (as discussed in section 1.3) have motivated the development of a novel individual-based stochastic simulation model in the current study (under Chapter 6). Findings from Chapter 2 (based on the MSM and a multivariate rank-based distribution-free test) have also informed some aspects of the novel stochastic simulation model concerning its model assumptions as well as other hypotheses of the study (investigated within the Bayesian setting under Chapter 6). To fit our novel stochastic simulation model, we propose a modified likelihood-free parameter estimation methodology for complex model calibration via approximate Bayesian computation (ABC) as well as a robust approach for ABC post-processing analysis (to correct for the imperfect mismatch between simulated and observed data).

The modified ABC algorithm (with its pseudo-codes outlined in Chapter 5, and dubbed in the current study as weighted-iterative ABC) is based on sequential Monte Carlo (SMC), sequential importance sampling (SIS), and ABC summary statistics weighting (adaptively computed based on the harmonic mean of previous and current summary statistics weights). Our robust ABC post-processing method (for adjusting the resulting ABC posterior and estimating its mean) is an extension of the standard ABC local-linear regression (to include an $L2$ regularisation term). It is considered an independent ABC final step after executing the weighted-iterative ABC to fit the novel simulation model. Unlike the standard ABC post-analysis methods, our proposed ABC posterior correction method is implementable even if the set of ABC summary statistics (possibly high-dimensional) is highly correlated in the neighbourhood of the observed summaries. In addition, it improves the standard ABC local-linear regression (with heteroscedastic errors) in the presence of multicollinearity, supercollinearity, outliers, or non-normal regression residuals.

A continuous-time Markov process dubbed the linear birth-death process with catastrophic extinction (B-D-C process) is further investigated and considered an auxiliary stochastic model (for the developed complex individual-based simulation model). For the first time in this study, the exact analytical results of the B-D-C transition function and its theoretical moments are derived and numerically validated, as in the setting of discretely observed processes. The motivation for the B-D-C model is to refine the set of summary statistics of the modified ABC based on parameter estimates of the B-D-C model. Before ABC fitting of the novel stochastic simulation model, three different parameter estimation methods: maximum likelihood estimation (MLE), generalised method of moments (GMM), and embedded Galton-Watson (GW) estimation methods for the B-D-C model were developed and compared by exploring the trade-off between estimation accuracy (quantified by the estimation bias, variance, and mean square error) and computational speed based on different *in silico* simulation experiments (where parasite population size is large, moderate, or low). The two zero states of the B-D-C process (due to either the natural death of parasites or parasite population extinction after host mortality) were distinguished or separately set up in the aforementioned B-D-C parameter estimators (MLE, GMM, and GW).

Furthermore, we adapted and compared two hybrid τ-leaping algorithms to simulate the B-D-C process and identify which method is cost-effective (based on their respective simulation speed and fidelity). The simulation of the B-D-C process using a hybrid τ-leaping algorithm also provided additional insights on accelerating the complex stochastic simulation model (based on its derived leap size estimator) by proposing a reasonable error threshold based on the trade-off between simulation accuracy and computational speed. Prior to fitting our novel stochastic simulation model, the fidelity of our proposed ABC methodologies was numerically assessed at different proposal draws based on a simple modelling problem (with multivariate normal likelihood and known analytical posterior distribution). Here, we investigated whether the resulting ABC approximation is mutually compatible with any Monte Carlo sample size ($N \geq 500$) or independent of N by determining the minimal number of proposal draws to achieve a good posterior estimation. The high computational cost of simulating data from the sophisticated individual-based simulation model, computing some components of the set of multidimensional ABC summary statistics (such as the B-D-C model parameters across a whole host population), and the quadratic cost of implementing ABC-SMC methods motivated this further exploration. Our ABC methodologies could also be modified and utilised in other studies to fit complex mathematical and simulation models. Section 1.2 outlines the structure of the thesis and provides a summary of each thesis chapter (including their interdependencies).

Therefore, achieving the study's overarching aims uncovered many intriguing and novel biological questions (with their underlying hypotheses tested by adapting other advanced statistical tests and the fitted mathematical models). Additionally, it has resulted in new findings and provided many directions for future studies of the gyrodactylid-fish system. Specifically, it has motivated other future research regarding between-host parasitic transmission and mixed gyrodactylid infection for this host-parasite system, amongst others (see the concluding Chapter 7 for detailed information regarding future work directions and the main thesis contributions). Briefly, the study is imperative mathematically and biologically in the following ways:

- The findings would help policymakers, disease modellers, and biologists better understand the *Gyrodactylus*-fish system using mathematical models and can be used to inform management decisions for the control of gyrodactylid infections.

- The use of the sophisticated multi-state Markov model provides a robust approach to model almost any longitudinal survival data where infection progresses through different dependent events or stages before host death may occur.

- The novel stochastic simulation model developed for the *Gyrodactylus*-fish system would help to provide a relatively realistic imitation of this biological system. Thus, it would facilitate experimental data collection and aid in investigating specific research questions and the system's complexity that may be difficult to control and implement experimentally. The simulation model can provide relevant information about parasite numbers at different body locations of fish over time for a fish group based on specific demographic characteristics (such as the parasite strain, fish type, fish sex and fish size) and underlying model parameters. The fish survival status and exact time to fish mortality are other essential outputs of the simulation model.

- The development of the modified ABC methodology (with sequential Monte Carlo and sequential importance sampling) coupled with the proposed ABC post-processing methodology (with $L2$ regularisation to deal with problems of multicollinearity, suppercollinearity and convergence issues of the standard ABC local-linear regression) would help parameter estimation of both sophisticated and simple likelihood-free models sequentially across a whole population.

- The mathematical models developed in this study can be adapted for future predictions within and beyond the standard 17-day infection period for a particular *Gyrodactylus* strain across the different fish populations. In addition, the individual-based stochastic model can also be further modified to conduct biological experiments for mixed gyrodactylid parasite populations and can be expanded further for broader host-parasite systems.

1.2 Thesis structure

This thesis is logically structured into seven main chapters, and the interdependencies (in terms of methodology development and contributions) between the thesis chapters are illustrated with arrows in an attempt to answer the underlying research questions (presented in section 1.6 after reviewing the gyrodactylid-fish system and its biology for further modelling purposes) and achieve the study's specific objectives (Figure 1.1).

Chapter 1 gives a general overview of the study (highlighting its novelty and contributions), the conceptual framework of the entire thesis (i.e., the thesis structure), other background information or research problems relevant for modelling purposes (section 1.3), the biology of *Gyrodactylus* and the gyrodactylid-fish system (sections 1.4 and 1.5), as well as formalises the research questions and the study's main limitations (presented under section 1.6).

Chapter 2 attempts to provide answers to three crucial biological questions concerning parasite microhabitat preference, host survival and parasite virulence based on existing empirical data. An introduction to the underlying research areas and other specific research gaps are discussed. The empirical data is described using two appropriate graphical summaries of mean parasite intensities (among surviving fish) across host body regions (microhabitats) and other covariates (parasite strain and fish stock) over time. Statistical tests (rank-based multivariate Kruskal-Wallis test and its post-hoc tests) are adopted to compare the spatial and temporal parasite distribution. A time-inhomogeneous multi-state Markov model is developed to explore the gyrodactylid infection progression (host survival and parasite virulence). Other analytical derivations of mean sojourn times and transition probabilities conditioned on other covariates (fish sex, fish stock, fish size, and parasite strain) are presented. The main results and further discussion sections relevant to answering research questions 1-4 (outlined in section 1.6) are summarized for subsequent modelling purposes. Findings from Chapter 2 provided epidemiological insights into the development of the novel individual-based stochastic simulation model (presented in Chapter 6).

Chapter 3 critically review mathematical models for host-parasite systems (which includes both individual-based and population-based models). Extant literature on modelling infectious diseases for microparasitic and macroparasitic infections is presented. Furthermore, an overview of population and individual-based mathematical models is given. Lastly, we present a summary of the previous modelling work carried out concerning the *Gyrodactylus* infection dynamics. This Chapter 3 thus provides the necessary information needed in developing mathematical models to study biological systems.

Chapter 4 outlines the auxiliary stochastic model, the birth-death process with catastrophic extinction (in terms of analytical proof, parameter estimation, and model fitting via two different hybrid τ-leaping algorithms). The definition of the B-D-C process, the derivation of its exact transition function, numerical validations using Monte Carlo estimation, comparison of three different parameter estimation methods (MLE, GMM, and GW), and detailed results from two different τ-leaping procedures are presented. As previously highlighted, the findings from Chapter 4 would help refine the summary statistics of the modified ABC methodologies (presented under Chapter 5) by identifying more cost-effective parameter estimation techniques for the B-D-C process (based on accuracy-speed trade-off) and further aid in accelerating the complex stochastic model (based on the results of the hybrid τ-leaping of the B-D-C process).

Chapter 5 begins with a literature review of existing ABC methodologies. Then, the weighted-iterative ABC with SMC and SIS is described with pseudocodes and other analytical or theoretical methodologies (including a proposed optimised linear function to project parasite numbers till the end of the infection period after host mortality to aid in computing ABC summaries). Finally, an extended local-linear regression with heteroscedastic errors and $L2$ regularisation (based on a weighted ridge regression) for ABC post-processing analysis is presented. The performance of the modified ABC methodology (with sequential Monte Carlo and adaptive importance sampling; described in section 5.3) and the proposed ABC posterior correction method (described in section 5.3.4) are tested using simple ABC experiments (based on multivariate normal conjugate priors with an exact form of posterior distribution), before calibrating the sophisticated individual-based simulation model based on the modified ABC methods. Finally, the robustness of the modified ABC algorithm to the choice of pre-specified decreasing tolerances and the number of proposals for adaptive importance sampling were assessed. Compared to the classical ABC posterior correction method, the modified ABC post-processing method (with $L2$ regularisation) is justified as a robust extension of the standard local-linear regression.

Chapter 6 presents the multidimensional individual-based stochastic simulation model for two age groups (using an extended hybrid τ-leaping algorithm) to include population carrying capacity (depending on the size of fish), a preference for parasites to move forwards or backwards (dependent on the parasite strain) and other specific information of host (fish sex, size and fish stock) across the external surfaces of fish over time (within a standard 17-day infection period). Hence, the individual-based model is developed to also discriminate between the behaviour of different strains of *Gyrodactylus* on the three different fish populations. Furthermore, results on the model fitting of the complex simulation model using the modified ABC and its posterior adjustment with $L2$ regularisation are given. Further multivariate analyses within the Bayesian framework are adapted to investigate the study's main hypotheses (based on adjusted posterior samples). Results from Chapter 6 are therefore used to provide answers to other research questions related to the study (i.e., research questions 5-9).

Chapter 7 finally summarises the works of the previous six chapters (Chapters 1-6), study's main contributions and future work directions.

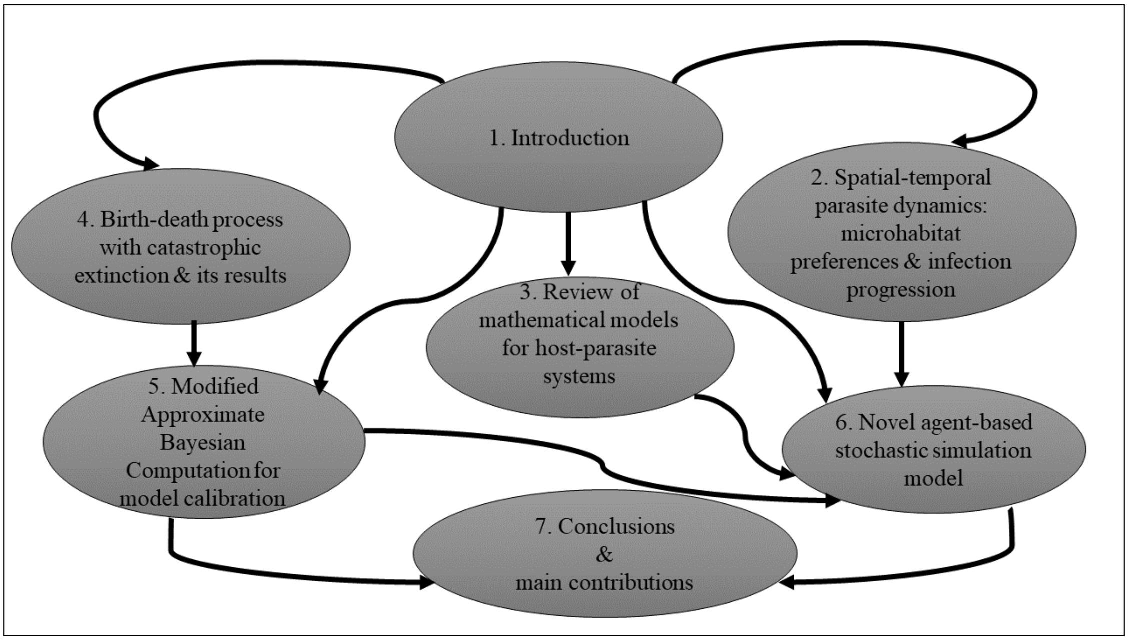
1. Introduction
2. Spatial-temporal parasite dynamics: microhabitat preferences & infection progression
3. Review of mathematical models for host-parasite systems
4. Birth-death process with catastrophic extinction & its results
5. Modified Approximate Bayesian Computation for model calibration
6. Novel agent-based stochastic simulation model
7. Conclusions & main contributions

Figure 1.1: The thesis structure.

All formulated mathematical theorems under Chapters 2, 4 and 5 of the current study are derived and proved for the first time (and thus not previously proposed in any other study). All the proposed algorithms, Algorithms 1-6 (with their respective pseudo-codes) under Chapters 4–6, are also developed for the first time in this study. All lemmas and Algorithms L1-L4 are presented in this current study based on previously published studies (with appropriate references provided accordingly). The first main work in this interdisciplinary research (corresponding to Chapter 2) has been submitted to a high-impact peer-reviewed biological journal. Results from Chapters 4–6 will be published in mathematics peer-reviewed journals. All R codes developed for statistical analyses and mathematical modelling (including their Jupyter Notebook HTML and source files) as well as the empirical data (for this study) can be found via the GitHub URL link (for reproducibility of results): github.com/twumasiclement/In-Silico-Modelling-of-Parasite-Dynamics.

1.3 Other background information of the study

Emerging infectious diseases pose a serious economic risk to freshwater fisheries, with several newly detected pathogens causing large scale disease outbreaks in fisheries and fish farms worldwide, including the UK. For example, some trout fisheries in South Wales and South West England have had to close due to *Argulus* species outbreaks [292]. Investigating the dynamics of such infectious diseases among fish is crucial since farmed fish is the major source of human protein, and aquaculture contributes significantly to the world economy. In the process of conserving wild fish stocks, the management of infectious disease dynamics has gained interest to researchers and fish-farming industries over the past years; since it is a predominant limitation to the sustainability and maintenance of farmed fish globally. In Europe, *Gyrodactylus salaris* has caused major problems in farmed salmonids; notwithstanding the lethal effects of other species, including *Gyrodactylus turnbulli* and *G. bullatarudis* [15, 187].

Experimental studies have previously explored the infrapopulation dynamics of *G. turn-*

bulli and *G. bullatarudis* ectoparasites on their fish host. Nonetheless, other important biological questions exist in relation to parasite microhabitat preference, host survival, parasite virulence, and the temporal transmission dynamics (in terms of infrapopulation, interspecies and interpopulation dynamics) of different *Gyrodactylus* strains across different host populations. Moreover, although much is known about the dynamics of gyrodactylid infections on a single fish, and an individual-based simulation model that reproduces these dynamics exists [306], the spatial information (species-specific microhabitat preference) and other relevant information were not well captured. In the existing individual-based model (IBM), information about parasite fecundity, age group (young or old parasite), parasite mortality, parasite mobility, host sex, host immune response and sources of stochasticity or random variations have not been fully incorporated for different *Gyrodactylus* strains across different fish populations.

Additionally, there exist limitations in some of the underlying model assumptions of the existing IBM. For instance, it assumed that the maximum distance that parasites can move within any time step was half the fish length, and localised immune response only occurs between 12-18 hours when at least one parasite occupies a site. However, within a realistic setting, the time to host immune response can occur at any time after infection, and localised immune response may depend on host and parasite genotype, the surface area of the body locations and host sex. The existing IBM did not distinguish between fish's major body locations (tail fin, lower body, upper body, anal fin, dorsal fin, pelvic fins, pectoral fins and head) or their respective surface area. For example, as individual host infections with *G. turnbulli* progress, parasites migrate from the caudal fin and body to the pectoral, pelvic, dorsal and anal fins; a migration to potentially facilitate transmission [129]. There exit a unique caudal-rostral preference between *G. turnbulli* and *G. bullatarudis* species [129, 131], and thus, these microhabitat preferences need to be incorporated when simulating the species-specific population dynamics. Hence, the need for a more realistic stochastic simulation model for studying these host-parasite systems' infection dynamics.

Also, through survival analysis, an individual's infection history can be modelled as a two-stage process with one possible transition from “alive" to “dead" state [1]. In such instances, we often adopt the standard logistic regression or Cox proportional-hazards regression to investigate risk factors of host mortality and to estimate hazard rates; while the non-parametric Kaplan-Meier method is used to estimate survival curves from censored data [93]. However, in most longitudinal studies, such as data collection from the gyrodactylid-fish system, the “alive" state could further be divided into two or more intermediate (transient) phases, each corresponding to a different stage of the infection [151]. Multi-state models (MSMs), also adapted in the current study, allow for time-to-event longitudinal data analysis in which surviving individuals may have varying infection outcomes over time (before host mortality may occur). A change of infection state is considered a transition or an event. Estimating progression rates, transition probabilities, the mean sojourn time in a given state, analysing the effects of individual risk factors, survival rates, and prognostic forecasting are all areas of interest under multi-state modelling [217]. Although MSMs have several applications in biomedical research, population demography, and other areas of epidemiology, this class of models is underused in studying host-parasite interactions in most parasitological studies. In the current study, for the first time, MSM is used to investigate the infection progression of two co-infecting gyrodactylids across different fish hosts. We thus develop a robust MSM for the gyrodactylid-fish system (in Chapter 2) to improve previous estimates of survival probabilities by Cable and van Oosterhout [57], and to efficiently quantify parasite virulence as a function of both host mortality and recovery (given some underlying covariates).

Therefore, this thesis mathematically investigates open biological questions to help better understand the gyrodactylid-fish system's infection dynamics based on advanced statistics and improved models. The infection dynamics of three different gyrodactylid strains across three different fish populations within a standard 17-day experimental period are investigated. Sections 1.4 and 1.5 briefly reviews the biology of gyrodactylid parasites and their host-parasite systems to provide good biological insights about the host-parasite under investigation.

1.4 Biology of *Gyrodactylus* parasites

Gyrodactylus are monogenean worms that commonly infect the gills and skin of freshwater and marine fish. There are several monogenean families with approximately 1500 species that can infect fish. Other monogenean worms belonging to the genus *Dactylogyrus* can also affect freshwater fish. The *Dactylogyrus* parasites are mostly found on the gills as opposed to *Gyrodactylus* parasites that predominantly infect the skin [67]. *Gyrodactylus* are ubiquitous on teleosts with currently over 400 described species [134]. These parasites are dubbed as "Russian-dolls" due to their rare reproductive mode in the Animal Kingdom; such that for 180 years, they have been known to produce fully matured daughters in the uterus of mothers, and each daughter contain developing embryos as represented by Figure 1.2 [15, 106]. An epidemic caused by *Gyrodactylus salaris* in Europe (Figure 1.3) has encouraged much research into gyrodactylid infections, and thus, these parasites are the best-studied monogenean. Consequently, three economically relevant species, namely: *G. salaris* on Atlantic salmon, *G. thymalli* on grayling and *G. derjavini* on brown trout have had a lot of research attention over the years. Nonetheless, other gyrodactylids have extensively been studied, including those that infect other teleosts (over 30,000 hosts). These hosts are key model systems for studying evolutionary and ecological processes and many other parasitic organisms in general [15]. This current study focuses on two distinct *Gyrodactylus* species (*G. turnbulli* and *G. bullatarudis*). Specifically, two strains of *G. turnbulli* species (laboratory-bred and wild strains) and a wild strain of *G. bullatarudis* are considered across three different fish populations.

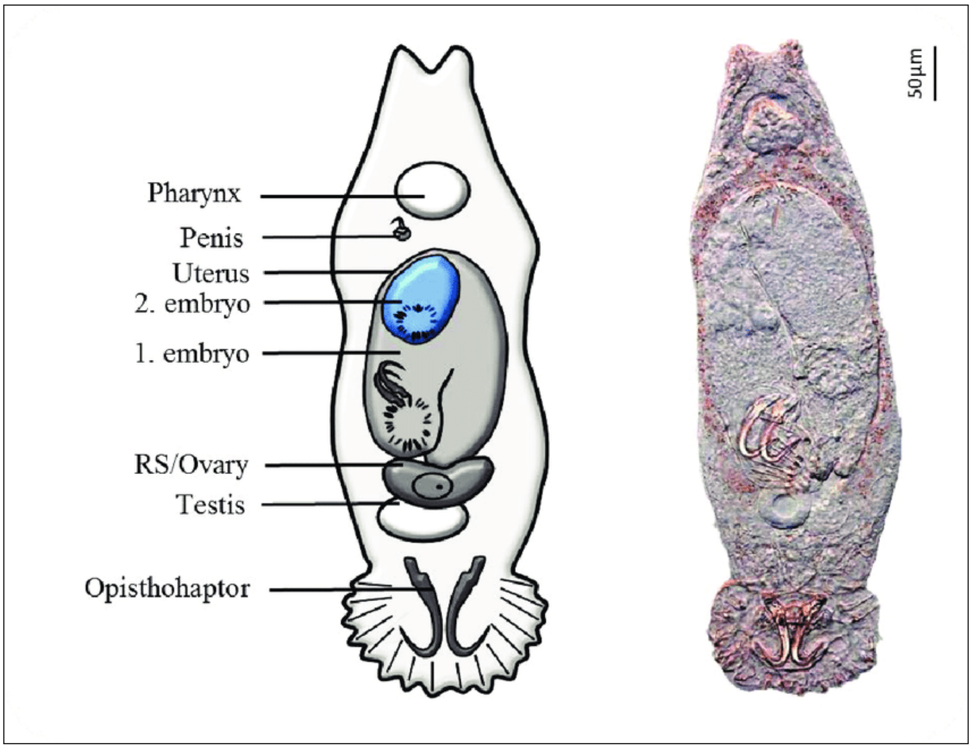

Figure 1.2: Light micrograph (interference contrast) of *Gyrodactylus salaris* with two developing daughters in *utero* like a "Russian-doll" [15, 106].

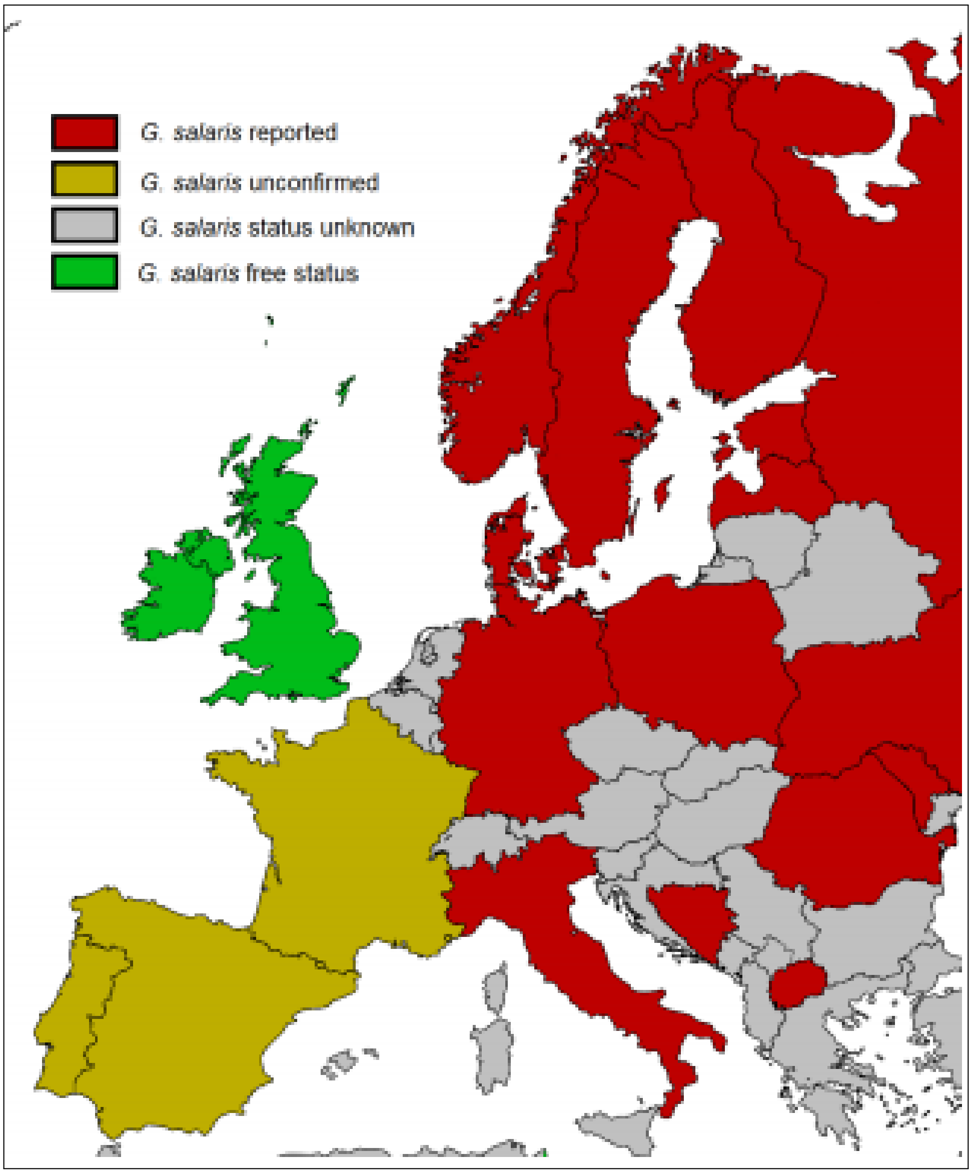

Figure 1.3: Regions of Europe reported with *Gyrodactylus salaris* infections (red), territories with unconfirmed reports (yellow), areas with unknown *G. salaris* infection status (grey) and lastly, territories free from this gyrodactylid infection (green) [81].

1.4.1 Morphology and progenesis of *Gyrodactylus*

Gyrodactylids are among the smallest monogeneans (Figure 1.4) characterized by their spindle-shaped body, posterior opisthaptor equipped with marginal hooks, hamuli and bars (Figure 1.5) [309]. In addition, there are two notable cephalic processes anteriorly, which bears adhesive glands and spike sensilla for attachment to the body of their host [see 17]. Gyrodactylids have unusual sexual maturity, with accelerated maturation of the sexual organs such that the larva can reproduce even as a juvenile. However, there is a significant difference in the progenesis of different *Gyrodactylus* parasites. A second reproductive adaptation of *Gyrodactylus* species is their viviparous nature, where parasites give birth to fully grown young parasites which are already developing embryos within their *utero*.

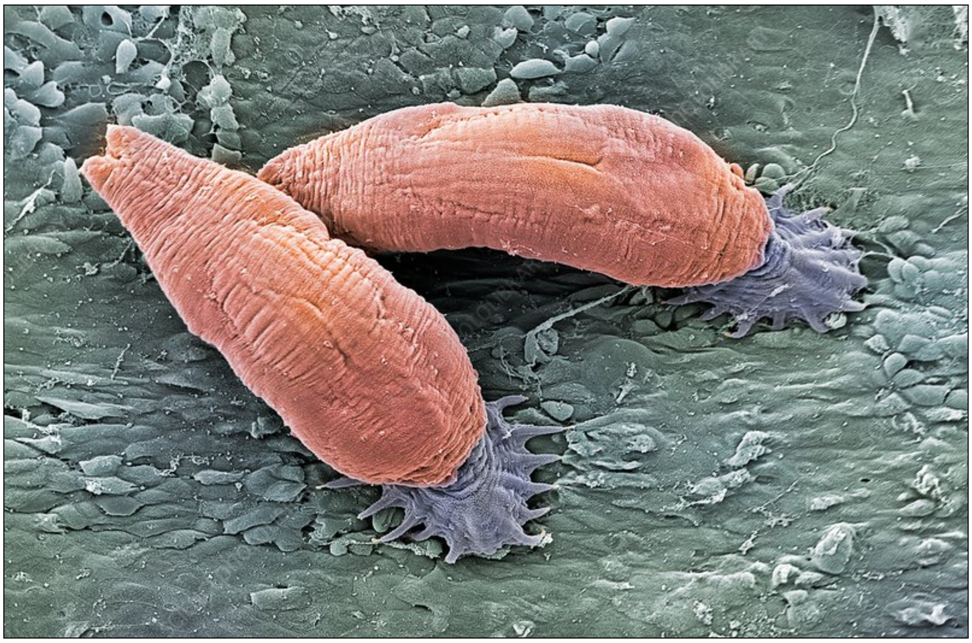

Figure 1.4: A scanned electron micrograph of *G. salaris* aquatic parasites (about 0.5 mm long)- Source: `sciencephoto.com/media/887549/view`.

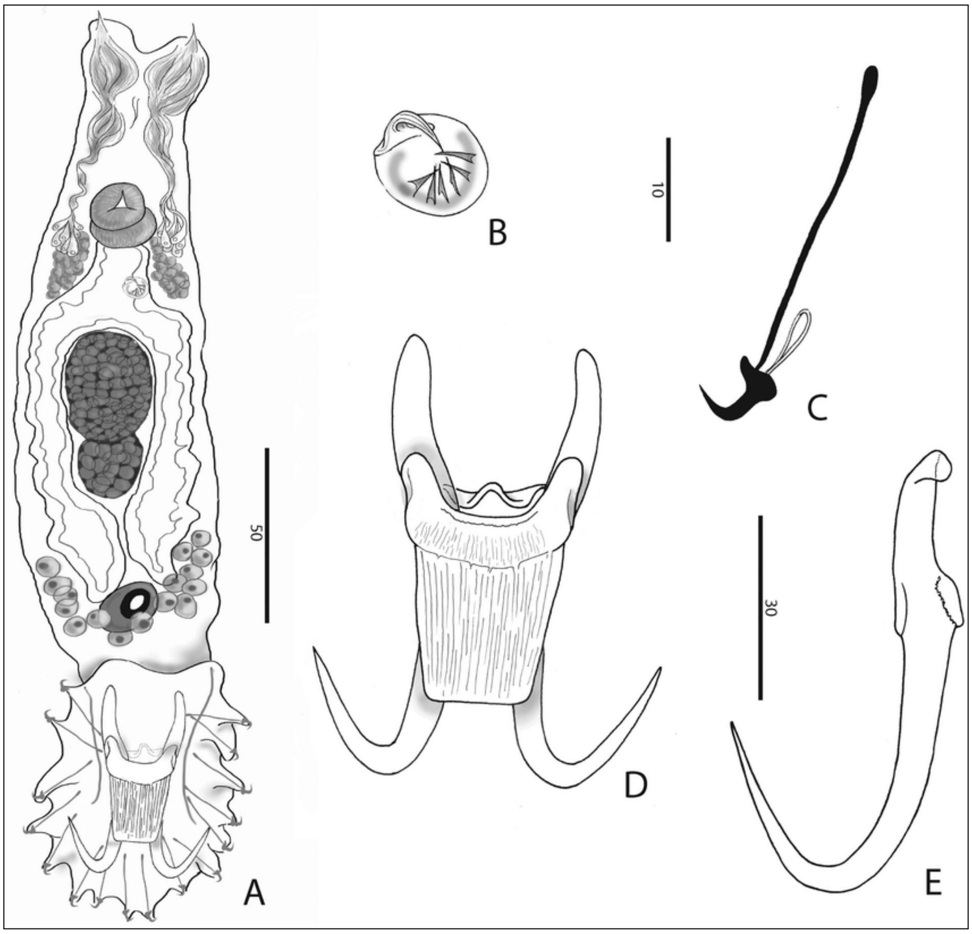

Figure 1.5: *Gyrodactylus jalalii* sp. nov. A. whole mount. B. male copulatory organ. C. marginal hook. D. anchor-bar complex. E. anchor. Scale bars represent 50 μm (whole mount), 10 μm (marginal hook, MCO) or 30 μm (anchor, anchor-bar complex) [309].

1.4.2 Ethology of gyrodactylids

The general behaviour of gyrodactylids includes locomotion, questing, feeding, reproduction and transmission. Subsequent sections summarize briefly the aforementioned behaviours of these parasites.

1.4.2.1 Locomotion

To move, the parasite uses the anterior glands (Figures 1.4 and 1.5) to cement the head momentarily to the fish while the opisthaptor is released and drawn towards the head; resulting in the release of the head eventually [15]. There could be a repetitive occurrence of this process for a period of time before the parasite settles down to a single position.

Gyrodactylids move along the skin usually, and this behaviour could be due to either avoidance of localised immune reaction from the fish or acquired immunity after second infection [57].

1.4.2.2 Questing

Questing behaviour may occur spontaneously and frequently among all gyrodactylids. This action significantly enhances their response to external stimuli such as touching substrate or host with a fibre. When questing, the parasite spreads from the substrate using its cephalic lobes. Gyrodactylids attach to the body of fish in the process or interact directly with other parasites resulting in copulation in most instances [15]. Generally, questing behaviour among these parasites allows the spike sensilla and unciliated receptors of the cephalic lobes to sample the surrounding host or other parasites (Figure 1.6).

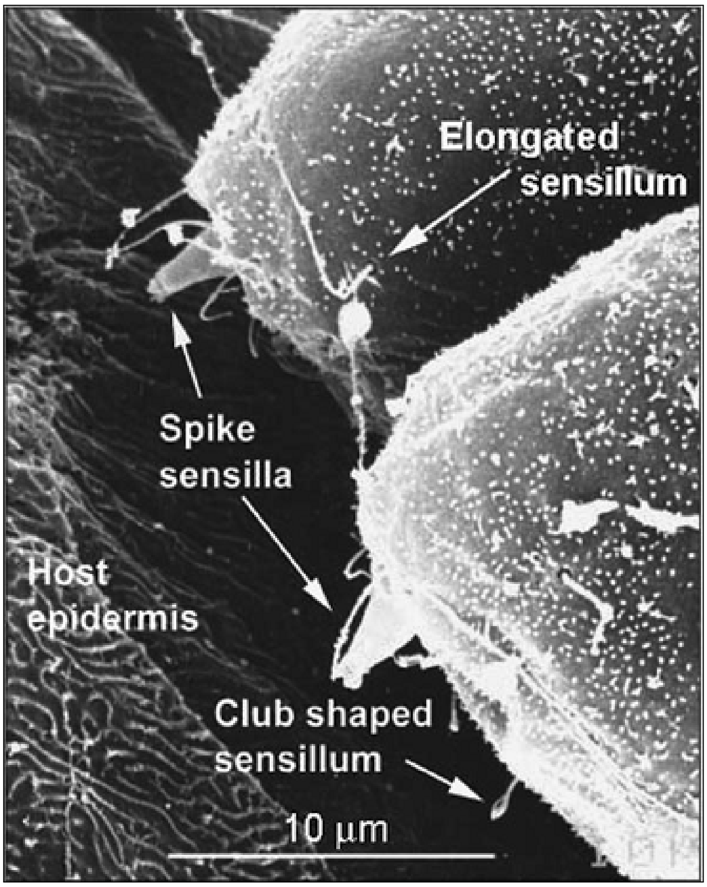

Figure 1.6: Anterior lateral view of scanned electron micrograph of the sensory apparatus at the anterior lobes of *G. salaris* actively searching the surface of its substrate [15].

1.4.2.3 Feeding

In most *Gyrodactylus* species, parasites individually lie in a distinctive pose with the anterior portion of the body stretched and flattened against the epidermis of its host (Figure 1.7); while rigidly attached by the opisthaptor, with the anterior lobes raised [53]. In the process, the pharynx is projected into close contact with the epithelium of the fish. It is often observed that the pharynx pumps, with slight waves of contraction passing along its body. Feeding usually elapses for few periods, after which the parasite uncurl but is contracted and relatively inactive for few minutes [15]. The mode of feeding can eventually cause fish mortality.

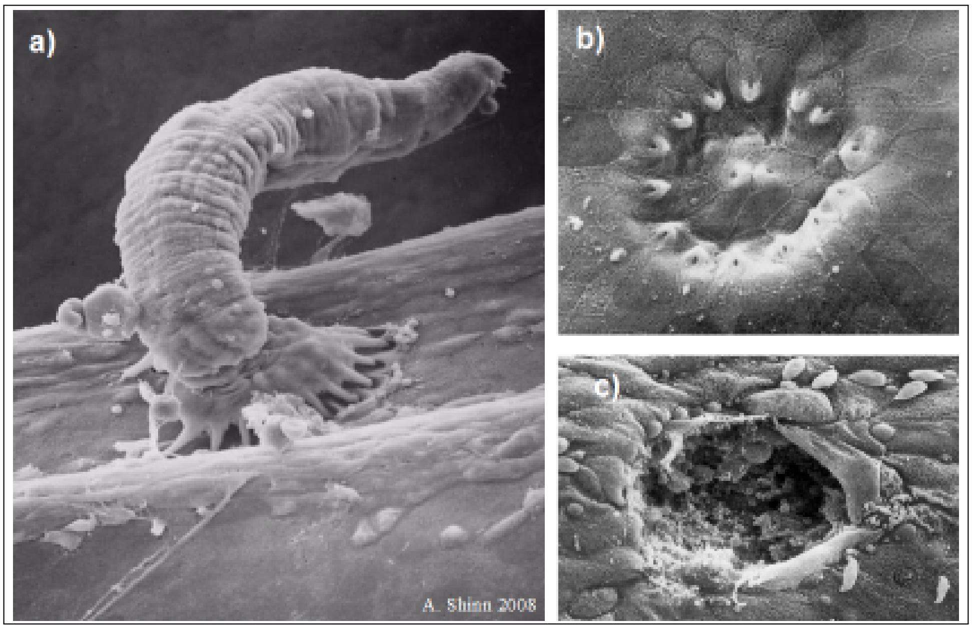

Figure 1.7: Mode of feeding of gyrodactylids which effectively kill a host through attachment, (a) - (b), and grazing activity, (c) leading to gyrodactylosis and disruption of osmotic permeability of the epidermis [81].

1.4.2.4 Reproduction

Gyrodactylus parasites are capable of both asexual and sexual reproduction. When the population density is low, these parasite species prefer asexual reproduction, while when the population density is high, they prefer sexual reproduction. Unfortunately, there is no clear information on where mates are found or how they protect themselves, although their mating behaviour is influenced by population size [53]. Gyrodactylids quest with members of the same species until they impale their penis into their preferred partner. After initiation of copulation, the partner may respond by grasping the initiator with its penis, and mutual insemination may proceed [127]. Copulation may be unilateral such as that one partner does not mutually inseminate. This behaviour can be observed in at least *G. turnbulli* [15]. Birth occurs after the copulation stage. Nevertheless, gyrodactylid birth could be viviparous (reproduction of fully grown young parasite) or oviparous (egg-laying without embryonic development inside the mother). Healthy worms attached to the body of fish can rapidly reproduce. Among viviparous *Gyrodactylus* parasites which are of interest to this study, the daughter breaks through a ventral birth pore close to the pharynx. A good way of knowing when the gyrodactylid parasite is about giving birth is through its gravid appearance and the slow waves from muscular contraction travelling along the parent's body. The first part of the daughter emerges as a bleb of tissue, which is rapidly followed by the worm's anterior portion (Figure 1.8). If the mother becomes detached during the beginning stages of birth and the daughter has no fish (host substrate) to attach to its anterior glands, both parasites may die since the daughter could not escape from the mother's uterus. However, in cases where the daughter is manually pulled with the watchmaker's forceps and fin pin, the mother may survive. Consequently, successful birth is contingent on the muscular activities of the mother and its offspring. However, abortion of young embryos could occur after prolonged detachment of both parasites [15, 53]. Hence, *Gyrodactylus* spp. reproduce with a fully grown daughter in the utero, which in turn envelopes a developing embryo, boxed inside one another embryo-like "Russian dolls" (Figures 1.2 and 1.8). A single gyrodactylid parasite on a fish's body can cause a whole population explosion due to their reproductive life cycle (Figure 1.9), thereby making them one of the most virulent monogenean parasites.

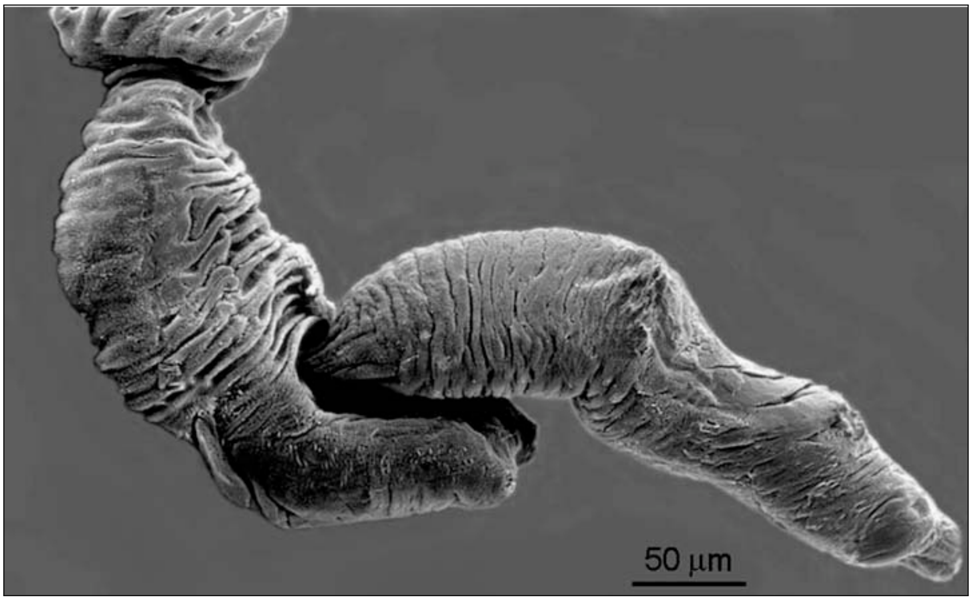

Figure 1.8: A gyrodactylid parasite individually giving birth to a pregnant daughter as large as itself [15].

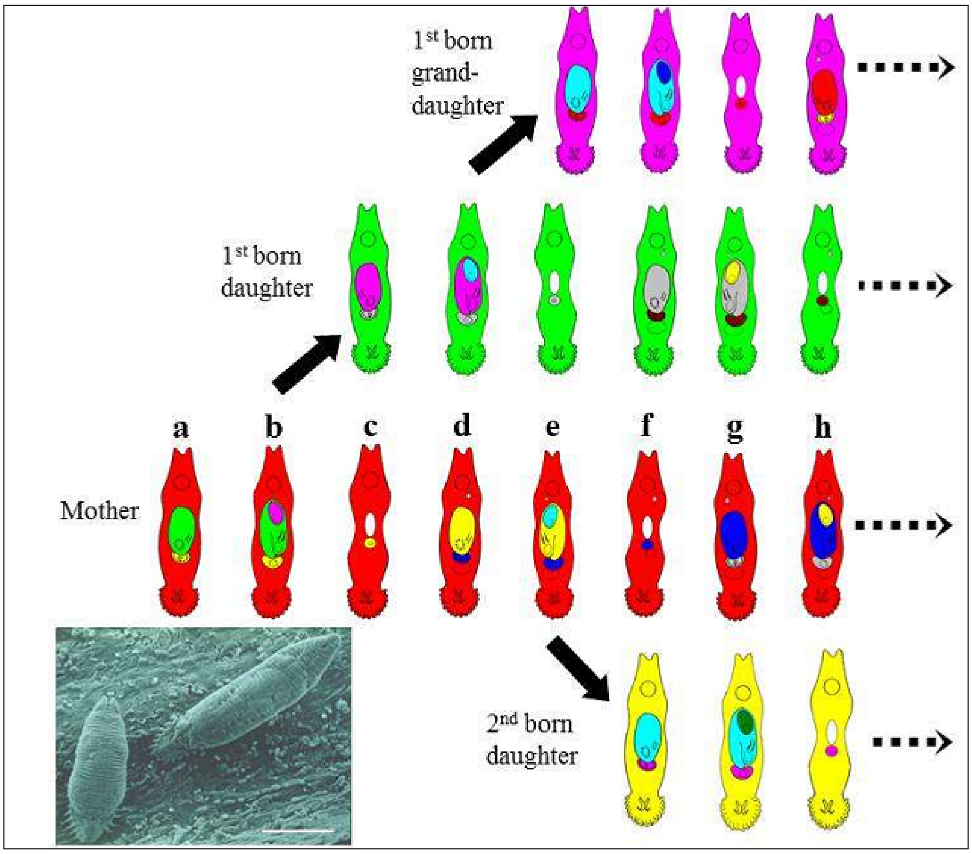

Figure 1.9: Reproductive cycle of *Gyrodactylus* parasites [53].

1.4.2.5 Transmission

Among viviparous gyrodactylids (which are the focus of the current study), the relatively continuous transmission and infection of new hosts throughout the life cycle increase colonisation of new host resources as well as host shifts [41]. At favourable conditions, the viviparous gyrodactylid may remain stationary on the current host's body, and transmission can easily lead to high fish mortality. Surprisingly, transmission can cause a low death rate of the parasite in the presence of fish immune response [41]. *Gyrodactylus* parasites can show diversity in their transmission strategies when detached from fish. After prolonged detachment or fish mortality, some parasite strains (e.g., *G. salaris*) can immediately transfer to suitable hosts they contacted. In contrast, there exist very little knowledge about transmission of oviparous gyrodactylids since all stages of their life cycle may lead to transmission of infection [126], and eggs are positioned around the inactive host resulting in infra-population over time [15, 179].

1.4.2.6 Swimming

This particular life process is not common to all gyrodactylids when detached from the substrate, but reported among other species like *G. rysavyi* [94]. This monogenean parasite of the skin, fins and gills of the Nile catfish (its host) can engage in directional swimming when disconnected from the host and freed in open water as opposed to other gyrodactylid species. Other studies have revealed that forcibly detached specimens of *G. turnbulli* and *G. salaris*, for instance, may thrash back and forth until they reach a solid substrate, but this behaviour is not unidirectional. *G. turnbulli* species' particular transmission behaviour (in which detached parasites travel through the water film, see section 3.3 of Cable et al. [55] study) is not considered a swimming behaviour ([cf. 154]).

1.5 Host-parasite system: *G. turnbulli* and *G. bullatarudis* infecting guppies (*Poecilia reticulata*)

1.5.1 Introduction

Gyrodactylus turnbulli and *G. bullatarudis* are the most studied gyrodactylids after *G. salaris*. These species predominantly occur among natural guppy populations, *Poecilia reticulata*, in Trinidad (Figure 1.10); and have been adopted as a model system with their host. The parasites occur at high prevalence of over 75% [307], but with low mean intensity (approximately ~ 10 parasites per host) [305]. Experimental studies have shown that their parasitic infection can cause sudden behavioural changes among hosts such as feeding behaviour, selection of mates and courtship activities [303, 306]. These parasites can also co-infect the same host [268]. The two parasite species are relevant to the aquarium trade and have been considered a threat to the conversation of endangered fish [176]. Generally, the rate of their transmission between female fish is higher than male guppies, with transmission also affected by host personality [256]. Previous experimental studies have helped get some insights into their infection dynamics and host immune responses. There is usually a slow growth in the parasite population from the start of infection, but the parasite population extinct after host immune defence (which is temporally induced after high parasite abundance) [16]. Guppies are among the most popular fish in aquaculture due to their natural ability to reproduce, colour diversity, robustness and ease of maintenance [268]. Thus, based on these features mentioned earlier about guppies, their laboratory usage for experiments makes them preferable when studying infections from both *G. turnbulli* and *G. bullatarudis.*

A study on survival and behaviour of *Gyrodactylus turnbulli* and *G. bullatarudis* on dead fish revealed that, under certain conditions such as high parasite abundance, these parasites might survive on a dead host for a while (see Figure 1.11); however, at low parasite abundance, there is a slight chance of survival [see 53]. Moreover, the study also discovered differences in their parasitic behaviour or migration after fish mortality, such that, *G. turnbulli* may move away from the dead host into water columns as opposed to *G. bullatarudis* which stay in the bottom of the fish [53].

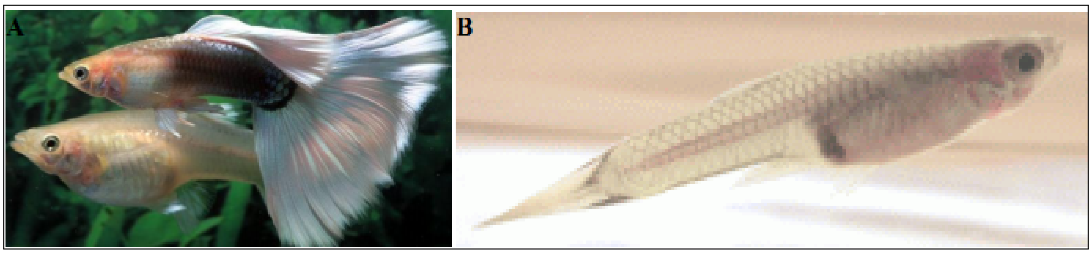

Figure 1.10: Ornamental male and pregnant female guppies (A); female guppy showing signs of fin clamping due to high parasite load (B)- Source: en.aqua-fish.net.

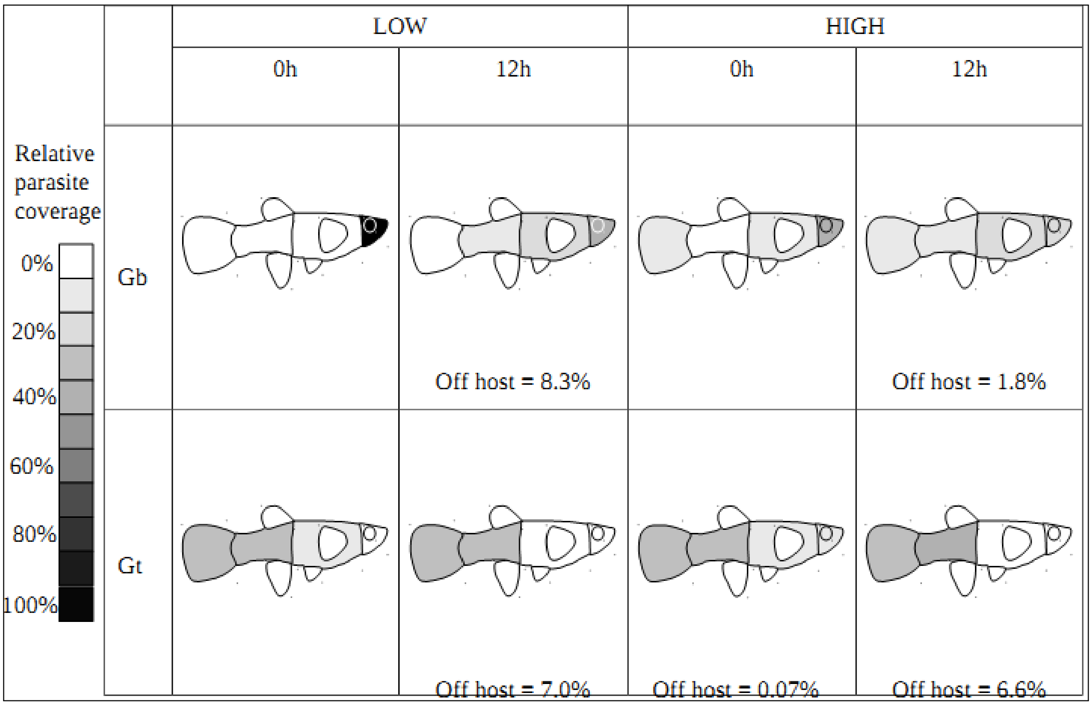

Figure 1.11: A diagram showing the mean percentage relative parasite coverage (underneath each figure) for *G. turnbulli* and *G. bullatarudis* at low and high burdens for time 0 (time of host's death) and 12h after host death [53].

1.5.2 Site specificity of *Gyrodactylus turnbulli* and *Gyrodactylus bullatarudis* on hosts

Previous studies have shown that some gyrodactylids have specific site preference, but this behaviour varies widely between species [15]. The majority of the parasites infect the skin and fins, whereas some also occur on the gills. *Gyrodactylus turnbulli* and *G. bullatarudis* unlike other gyrodactylid species of guppies, display marked site preference. The *Gyrodactylus bullatarudis* usually occur rostrally towards the head and mouth at low densities, whereas the *G. turnbulli* prefer mostly the caudal regions of fish [55, 131, 129]. Under experimental conditions, *G. turnbulli* most commonly infect guppies through their fins [129]; however, parasites individually move towards the caudal peduncle and tail fins over time when the parasite burden gradually increases. The parasites then return to the fins to conceivably cause transmission [129].

Immunological changes may influence migration behaviour in the epidermis of the host [15]. Specifically, parasites may face three scenarios that will cause them to change location or preference. The possibilities may be that: an immune-active site was encountered, and thus parasites moved, or the location was previously infected but not immune-activated and may choose to stay or finally, the site becomes immune-naive, which cannot harm parasites or force migration [15]. In the process of feeding on or infecting fish, different gyrodactylid species within a particular microhabitat exploit the same source of carbon, which in turn lead to host death after over-exploitation. Consequently, any other gyrodactylid species on such dead hosts will die out [15]. Although these species' respective caudal and rostral preferences on the host are reported, it is unknown whether this is consistent over time and across different fish stocks. The spatial and temporal dynamics concerning microhabitat preferences of three different *Gyrodactylus* strains across different host populations have been explored in details in the thesis Chapter 2, amongst other objectives.

1.6 Research questions and study's limitations

The study consists of seven chapters, together attempting to provide answers to the following nine major research questions:

1. Is the caudal and rostral preferences of the gyrodactylid strains consistent over time and across different fish stocks?

2. Does fish sex, fish size, fish stock, and parasite strain affect gyrodactylid infection progression (host recovery and host mortality over time)?

3. What is the average infection time of infected fish conditioned on the significant predictors?

4. Is the virulence (quantified by both host mortality and recovery) of the gyrodactylid strains time-varying and dependent on the covariates (fish sex, fish size, fish stock and parasite strain)?

5. Are the birth rates (for young and old parasites) and death rates (with or without immune response) of *Gyrodactylus* parasites significantly different across the three parasite strains?

6. Is the adaptive immune response from gyrodactylid infection progression, sex and host-dependent?

7. Is the mortality rate of male fish from gyrodactylid infection significantly higher than female fish?

8. Are the microhabitat preferences of *Gyrodactylus turnbulli* and *G. bullatarudis* parasite species driven by their rate of movement on their fish host?

9. What is the effective population carrying capacity of *Gyrodactylus* parasites at the major body regions of their fish host?

The research questions were motivated by several factors: the biology of the *Gyrodactylus* and previous experimental studies about its host-parasite system (reviewed under sections 1.4 and 1.5), open biological questions concerning the distinct gyrodactylid microhabitat preference (between *Gyrodactylus turnbulli* and *G. bullatarudis* species) across different host populations, key determinants of host survival and parasite virulence (conditioned on parasite mortality and host recovery), as well as underlying research hypotheses (motivated by research questions 5-8) to be investigated based on the fitted individual-based stochastic simulation model (within a Bayesian setting). Research questions 1-4 are investigated under Chapter 2. Answers to research questions 5-9 are provided through the use of Bayesian hypothesis testing after ABC fitting of the complex stochastic simulation model (presented under Chapter 6).

Limitations of the study

The study's limitations (associated with developed mathematical models and the scope of the study) are outlined as follows:

- The multi-state Markov model developed in this study assumed that fish could not be reinfected after infection loss due to the study's experimental design. However, this can be modified for other biological or ecological systems that allow host reinfection after parasite extinction. Thus, the multi-state model's results may change due to reinfection within a social setting where population mixing between hosts is possible.

- In addition, in the multi-state Markov model, we could not include spatial information and other relevant information about parasite fecundity, age group (young or old parasite), parasite mortality, parasite mobility and host immune response while exploring host survival and parasite infrapopulation dynamics.

- Moreover, the novel individual-based stochastic simulation model (formulated within the standard 17-day experimental period) is only developed to investigate the infrapopulation infection dynamics of gyrodactylids on a fish and cannot examine infection transmission between hosts.

- Additionally, the development of other methodologies in the current study (initially not considered in the original research plan), such as the modified ABC post-processing method with $L2$ regularisation and other additional numerical evaluation experiments of the proposed ABC methodologies, broadened the scope of the study.

- Therefore, the current study did not investigate the long-term infection dynamics of the gyrodactylid parasites beyond the standard 17-day experimental period. Hence, this study did not consider the interpopulation (or mixed-gyrodactylid) within-host infection dynamics, between-host transmission or intrapopulation infection dynamics (using a social network model) and long-term predictions beyond the standard 17-day infection period across the different host populations by adapting the novel stochastic simulation model.

Chapter 2

Spatial and temporal parasite dynamics of *Gyrodactylus*

2.1 Introduction

The fastest growing global food sector is aquaculture with parasites posing the greatest threat to economies, sustainability, and animal welfare. Management of parasitic infection and prophylactic treatment in aquaculture is costly. An estimated hatchery loss to parasitism of 20%, results in an annual loss >£100 million globally [276]. Taking into consideration the grow-out of aquatic species and associated dietary investment, the cumulative annual cost from infectious disease to fish farming is estimated between £1-10 billion [68, 276]. The current study focuses on the spatial and temporal infection dynamics of the gyrodactylid-fish system by providing new epidemiological insights with the help of a more robust multi-state Markov model, a rank-based non-parametric multivariate analysis of variance test coupled with its post-hoc tests. Gyrodactylids are common fish parasites [276]; *Gyrodactylus salaris* alone caused epidemics among farmed salmonids, which resulted in death of up to 86% of salmon in infected rivers [15, 187].

Gyrodactylids are monogenean ectoparasites that are ubiquitous on teleosts [133]. Among well studied Trinidadian guppy populations, gyrodactylids are the dominant parasites [≥42% prevalence, 3.3 mean intensity; 52]. The prevalence of *Gyrodactylus* species varies spatially across watercourses (lower, mid, and upper courses of the rivers or lakes) and temporally among Trinidadian population, and between host sexes [286]. *Gyrodactylus* prevalence is higher in female guppies, but only in lower courses [286], predominately due to fish shoaling behaviour [72, 118]. The parasites have no specific transmission stage and transfer from fish to fish occurs during host contact. Their reproductive mode is similar to that of microparasites with replication occurring directly on the host [reviewed by 15]. Their hyperviviparous nature and short generation times [< 24 h at 25°C; 270] can cause population explosions with substantial spatial and temporal variation amongst different species [e.g., 56, 130, 196, 205, 265]. Many infect the skin and fins, others occur predominantly on the gills [130, 229]. *Gyrodactylus turnbulli* and *G. bullatarudis*, which both infect the guppy (*Poecilia reticulata*), niche partition with *G. turnbulli* occurring caudally [129] and *G. bullatarudis* rostrally [131]. According to Haris [129], as individual host infections with *Gyrodactylus turnbulli* progress, parasites migrate from the caudal fin and body to the pectoral, pelvic, dorsal and anal fins; a migration to potentially facilitate transmission. Gyrodactylids may also move to optimise feeding, reduce competition and avoid localised immune reactions [15, 137, 141, 200, 257, 306]; the scorched earth hypothesis [265]. Although the respective caudal and rostral preferences of *G. turnbulli* and *G. bullatarudis* on the host are well reported [e.g,. 131, 129], consistency over time and across different fish stocks is not.

Host survival following gyrodactylid infection was previously explored by Cable and van Oosterhout [56]. They showed that mortality of guppies differed significantly between fish stocks for each parasite strain. From their experimental study, guppies were categorised according to whether they: i) fought off the infection, ii) remained infected, or iii) died while infected. The fate of these guppies was predominantly affected by fish size, such that smaller guppies were more likely to clear the infection, while larger fish either died or remained infected beyond the end of the study period [56]. Traditional survival models, including Kaplan-Meier and Cox proportional-hazards regression, are not able to effectively incorporate changes in host infection status over time [151]. For instance, individuals (hosts) are either classified as censored (alive) or uncensored (dead) at the end of the study but information on whether the censored cases remained infected or fought off the infection is excluded. This can, therefore, lead to incorrect estimation of risk parameters such as survival probabilities and hazard ratios [32]. Multi-state models (MSMs), however, are suitable for such modelling where the infection history of the host is of interest, and at any one time the host could occupy a different infection status [151].

Consequently, it is useful and more robust for analysing survival data when different treatments and intermediate events can occur in the lifetime of individuals or the population of interest [207]. MSMs are able to estimate additional relevant quantities such as the mean sojourn times from transient states and transition probabilities. Hence, MSMs are considered as a natural extension of the standard survival models [207, 211].

MSMs provide a robust approach to modelling almost any kind of longitudinal failure time data. Thus, it can be used to perform life history analysis across several areas of application, including but not limited to demography, epidemiology, actuarial science, reliability analysis, and micro-sociology [5, 151]. In the standard context of multi-state modelling, individual life histories are observed as independent sample trajectories of stochastic processes moving between states in a discrete state space [5]. In epidemiology, the states of the process could be defined as disease outcomes such as healthy, exposed, infected, diseased with complications, or dead (for instance). In other fields, the states could correspond to various statuses for an individual, an insurance policy, or a technical component (amongst others). A change of state is considered a transition or an event. The state structure (which is not unique) describes the states and the various transitions between them. For each possible transition, an MSM is specified entirely by its state structure (defined by the transition rate at which an event occurs over time) and the form of the hazard function for the respective transitions (given individuals' characteristics or covariates). This class of models can be applied to study more complicated types of longitudinal (or life history) survival data (depending on the modelling problem under study). The model can be formulated as either Markovian (if the future state of the process is dependent on only the current state and independent of previous states) or non-Markovian (if the Markov assumption fails); however, the latter is not well discussed in the literature [151]. This study also focuses on only the Markov-type MSM (as in the Markov chain setting). Andersen and Borgan [5] and Hoem et al. [147] have reviewed Markov models, whereas Cox and Miller [71] have discussed MSMs in detail.

For multi-state processes that are misclassified or can only be viewed through a noisy marker, hidden Markov models can be implemented [158]. There is more extensive literature on different classes of MSMs and Markov extension models with specific applications to the modelling of fertility, competing risks, disability, recurrent events, twin survival, and alternating events [reviewed by 151]. Additionally, the MSM can either be time-homogeneous (if the transition intensities or rates of the Markov chain are invariant over time) or time-inhomogeneous (in the case of time-varying transition rates). Also, the time-space can be either discrete or continuous. Other studies have employed the discrete-time MSM in the study of infectious diseases [e.g., 61, 192]. There exists a comprehensive and flexible software packages (e.g., *msm* and *msSurv* R packages) to help modellers fit any proposed continuous-time MSM (for a given biological system) based on a panel or longitudinal data and well-defined transition intensities [158]. Although MSMs have been applied in several fields, they are rarely used for studying host-parasite interactions and host survival in most parasitological studies. For the first time in this study, MSM is used to investigate the infection progression of two co-infecting gyrodactylids across different fish hosts.

Parasites and their hosts compete for survival. Such co-evolutionary interactions drive virulence originating from parasite pathogenesis and host defence [206]. Together, measures of host mortality, host resistance, host recovery, mutation, superinfection, host heterogeneity, and mode of transmission all contribute to explain parasite virulence [206]. There exists significant heterogeneity in virulence between *Gyrodactylus turnbulli* and *Gyrodactylus bullatarudis* strains [56]; for example, the laboratory bred *G. turnbulli* strain (*Gt3*) inflicts a higher proportion of causalities, followed by *G. bullatarudis* (*Gb*), and then the wild-type *G. turnbulli* (*Gt*). However, *G. bullatarudis* reach a higher load over time compared to the two *G. turnbulli* strains, where the wild-type reach a higher maximum load than the inbred strain [56]. Previous studies of these host-parasite systems explored parasite virulence predominantly based on host mortality, host resistance and host heterogeneity, with less emphasis on host recovery as a measure of virulence

[56, 287]. Thus, virulence of the three gyrodactylid strains on different fish stocks has not been quantified over time while accounting for possible changes in host infection status before host mortality may occur.

The current study focuses on the spatial and temporal infection dynamics of the gyrodactylid-fish system by providing new epidemiological insights with the help of a robust MSM, a multivariate rank-based distribution-free test coupled with its post-hoc tests. In this study, we focus on a continuous-time MSM that is time-inhomogeneous. This time-inhomogeneous MSM is used to analyse our longitudinal survival data (instead of its time- homogeneous version) since transition intensities may naturally differ across individuals or time-varying covariates. Here, we examine gyrodactylid microhabitat preference of three parasite strains (two strains of *Gyrodactylus turnbulli* and one strain of *G. bullatarudis*), and how these preferences vary across three different fish stocks over time based on existing experimental data. We also develop an MSM to improve on previous estimates of survival probabilities given fish sex, fish size, fish stock and parasite strain. We further quantify and compare the virulence (measured by rate of host mortality and recovery) over time, and estimate other relevant epidemiological quantities (mean time of host to remain infected and probability of infected host to either recover or die across the covariates).

2.2 Materials and description of empirical data

2.2.1 Experimental data

The data used here is from the experimental study of Cable and van Oosterhout [56], subsequently used as the basis of an agent-based model in van Oosterhout et al. [306]. Briefly, cultures of three different *Gyrodactylus* strains were used to infect three different fish stocks: Ornamental Stock (OS), Lower Aripo River fish (LA) and Upper Aripo River fish (UA); 157 guppies in total, in a full factorial design to give nine different host-parasite combinations, with $13-22$ replicates per combination. Two out of the three parasite strains were *Gyrodactylus turnbulli*, a laboratory-bred strain (*Gt3*) and a wild *G. turnbulli* strain obtained from guppies caught in the Lower Aripo River, Trinidad (*Gt*); whereas the third strain was *G. bullatarudis*, also a wild type obtained from hosts in the Lower Aripo River. Both male (68) and female (89) individually isolated guppies were used for the experiment and maintained under constant environmental conditions (25 ± 0.5°C; 12h light/12h dark regime). All tanks and containers were kept in a randomised block design to reduce common environmental effects. The fish considered in the experiment were naive and, thus, bred under parasite-free conditions. Each fish was then infected with two parasites at time 0, and parasites were counted every 48 h over a 17-day infection period. For each fish, the number of parasites was recorded across eight different body regions (tail fin, lower body, upper body, anal fin, dorsal fin, pelvic fins, pectoral fins and head). Survival data describing the various host infection status (remain infected, recovered from the infection or died) over time were extracted from the empirical data for the multi-state modelling. The number of surviving fish (with or without host infection loss) and dead fish across the nine different host-parasite groups over time from days 1 to 17 is given by Table 2.1.

Table 2.1: Number of surviving fish (with or without infection loss) and dead fish for the nine different host-parasite groups over time from days 1 to 17.

Days	***Gt3***			***Gt***			***Gb***			**Total** (n)
	OS	LA	UA	OS	LA	UA	OS	LA	UA	
Fish alive with infection										
Day 1	14	22	17	13	17	19	17	19	19	157
Day 3	13	20	13	11	16	16	16	19	16	140
Day 5	13	19	8	11	16	13	15	18	15	128
Day 7	13	18	4	11	14	11	14	14	10	109
Day 9	12	17	3	11	13	10	13	12	6	97
Day 11	12	15	3	10	13	6	11	10	3	83
Day 13	11	12	2	10	11	5	10	6	3	70
Day 15	9	10	0	7	10	5	7	4	2	54
Day 17	0	0	0	3	3	2	1	2	0	11
Fish alive with loss of infection										
Day 1	0	0	0	0	0	0	0	0	0	0
Day 3	1	2	0	0	0	0	1	0	1	5
Day 5	1	2	3	0	0	1	1	1	1	10
Day 7	1	1	3	0	0	1	1	1	2	10
Day 9	2	1	3	0	0	1	2	1	2	12
Day 11	2	1	3	1	0	1	4	1	2	15
Day 13	2	2	3	1	0	1	4	1	2	16
Day 15	3	2	3	2	0	1	5	2	2	20
Day 17	5	4	3	3	3	1	7	3	2	31
Fish dead										
Day 1	0	0	0	0	0	0	0	0	0	0
Day 3	0	0	4	2	1	3	0	0	2	12
Day 5	0	1	6	2	1	5	1	0	3	19
Day 7	0	3	10	2	3	7	2	4	7	38
Day 9	0	4	11	2	4	8	2	6	11	48
Day 11	0	6	11	2	4	12	2	8	14	59
Day 13	1	8	12	2	6	13	3	12	14	71
Day 15	2	10	14	4	7	13	5	13	15	83
Day 17	9	18	14	7	11	16	9	14	17	115

2.2.2 Data preprocessing and visualisation

All analyses were carried out using R version 3.6.3 [247]. Images of fish were produced in Gimp software version 2.10.12 [293] and outlined in R. Two graphical summaries of the data were produced. These are available in full in Appendices A and B, with examples given in Figures 2.1 and 2.2. For the first summary (Figure 2.1 and Appendix A), the shading represents the log mean intensity of parasites over surviving fish. The number of surviving fish for the nine different host-parasite groups (obtained from the fully crossed design of the three parasite strains and three different host populations) generally decreased (slowly) over time from days 1 to 17 (refer to Table 2.1). For the second graphical summary (Figure 2.2 and Appendix B), the eight body regions of the fish were recategorised into four: tail, lower region (comprising of the lower body, anal fin, pelvic fins, and dorsal fin), upper region (made up of the upper body and pectoral fins) and the head. This re-categorisation allowed us to visually and statistically assess any caudal-rostral preference of the three parasite strains on the three fish stocks more effectively over the study period due to low parasite numbers observed on the fish fins (anal fin, pelvic fins, dorsal fin, and pectoral fin).

2.3 Methods and results

2.3.1 Multivariate Kruskal-Wallis test for parasite distribution comparison across host body regions

The multivariate Kruskal-Wallis test (MKW) is a multivariate extension of the distribution-free univariate Kruskal-Wallis test [143]. We used it to test the null hypothesis that distribution of parasite number at the four body regions (tail, lower region, upper region and head) is equal for the different host-parasite combinations at each observed time point. Let Y_{ij} be a vector of the number of parasites at the four body regions for the jth fish from the ith group (host-parasite combination), where $i = 1, 2, 3, \cdots, 9$ and $j = 1, 2, 3, \cdots, n_i$. Let R_{ij} be the rank corresponding to Y_{ij} calculated element-wise (ties are assigned a mean rank) and $\bar{R}_i = \sum\limits_{j=1}^{n_i} \frac{R_{ij}}{n_i}$ then $E(\bar{R}_i) = m = \frac{n+1}{2}$ under H_0; where $n = \sum\limits_{i=1}^{9} n_i$ is the total number of fish ($n = 157$), $\bar{R}_i$ is the mean rank for each ith group and n_i is the number of fish in group i. The vector $U_i = (\bar{R}_{i1} - m, \bar{R}_{i2} - m, \bar{R}_{i3} - m, \bar{R}_{i4} - m)^\top$ denotes the average ranks for the ith group corrected for m for each variate (body regions). The pooled within-group covariance matrix is estimated as

$$V = \frac{1}{n-1}\sum_{i=1}^{9}\sum_{j=1}^{n_i}(R_{ij} - m\mathbb{1})(R_{ij} - m\mathbb{1})^\top, \tag{2.1}$$

where $R_{ij} = (R_{ij1}, R_{ij2}, R_{ij3}, R_{ij4})^\top$ and $\mathbb{1} = (1,1,1,1)^\top$. The MKW test statistic ($\mathcal{W}$), given as

$$\mathcal{W} = \sum_{i=1}^{9} n_i {U_i}^\top V^{-1} U_i \sim \chi^2_{k(g-1)}, \tag{2.2}$$

is approximately (asymptotically) chi-squared with $k(g-1)$ degrees of freedom, where $k = 4$ and $g = 9$ [143]. After performing the MKW, the univariate Kruskal-Wallis test (UKW) was used to further compare the distribution of parasites at each of the four body regions for each parasite strain (*Gt3*, *Gt* and *Gb*) across the fish stocks (OS, LA and UA) at each time point (days 1 to 17). A Bonferroni-Dunn's post-hoc test was finally applied for pairwise comparisons of the parasite distribution between the different parasite-fish combinations over time. The caudal-rostral preference of the three parasite strains on the three fish stocks was statistically inferred from these tests (testing the niche partition hypothesis of *G. turnbulli* and *G. bullatarudis* for preferences at the caudal and head regions, respectively).

2.3.2 Results on parasite microhabitat preferences

Fish heatmaps (Figure 2.1 and Appendix A) depict variations in parasite distribution across eight body regions (caudal fin, lower body, upper body, anal fin, pelvic fin, dorsal fin, pectoral fin and head) over time for each gyrodactylid strain (*Gt3*, *Gt* and *Gb*) on the different fish stocks (OS, LA and UA). *Gt3* showed a clear preference for the caudal fin and lower body, with higher mean intensities on OS and LA fish than on the UA stock from day 7 until the end of the infection period. By day 15, all the UA fish had lost the *Gt3* infection. Similarly, *Gt* was more abundant on the tail and lower body until day 13; but switched to a head preference among only OS and LA populations on day 15. In contrast, *Gb* showed a clear rostral preference from day 7 onwards; a preference strongest in OS$\geq$LA$\geq$UA fish stocks until the end of the infection period.

When comparing just four body regions of the fish (tail, lower region, upper region and head), the peak time of infection varied spatially across parasite strains and fish stocks (Table 2.2; Figure 2.2). On day 15, higher mean intensities were recorded on the head for both *Gt* and *Gb* on OS fish stock. Also for *Gb* on the same fish stock, a higher number of parasites occurred on the head between days 9 and 17 compared to any other body region or host-parasite combinations (Figure 2.2 and Appendix B). Parasite distributions varied at the four body regions across the nine host-parasite combinations (Figure 2.2) from days 1 to 15 (MKW: $71.25 \leq W \leq 168.57$, df=32, $p<0.001$), but not on day 17 ($W=38.12$, df=32, $p=0.211$). Only the parasite distribution at the tail and head respectively differed significantly across the nine host-parasite combinations from days 1 to 5 and on day 9 ($p \leq 0.001$). However, parasite distribution differed significantly among groups on the lower body region on days 7 and 11 (UKW: $17.12 \leq H \leq 17.74$, df=8, $0.023 \leq p \leq 0.029$), tail on days 7 and 15 ($19.49 \leq H \leq 24.93$, df=8, $0.002 \leq p \leq 0.012$) and head on days 7, 11 and 13 ($21.22 \leq H \leq 47.36$, df=8, $0.001 < p \leq 0.007$).

From the Bonferroni-Dunn's tests, there were significant pairwise differences in parasite distribution at the tail between all *Gb* groups (*Gb*-OS, *Gb*-LA and *Gb*-UA) and *G. turnbulli* strains on the fish stocks (with the exception of *Gt3* on OS) during day 1 of infection ($0.001 < p \leq 0.016$). However, there was no significant difference in parasite distribution of the *G. turnbulli* strains at the tail across the 3 fish stocks over time; with the exception of days 3 and 15, between *Gt3*-OS and *Gt*-LA groups ($p=0.019$) as well as between *Gt3*-LA and *Gt3*-UA groups ($p=0.037$). On days 3 and 5, parasite distribution at the tail was significantly different ($0.001 < p \leq 0.036$) between all *Gb* groups and *Gt* groups with the exception *Gt*-OS for day 3 and *Gt*-UA for day 5. Parasite distribution at the tail on day 7 was significantly different between *Gb*-UA and *Gt* groups (*Gt*-OS and *Gt*-UA); whilst a significant difference was found between *Gb-UA* and *Gt* groups (*Gt*-LA and *Gt*-UA). Nevertheless, there was no significant difference between groups of the *G. turnbulli* strains and *G. bullatarudis* from day 15 till the end of the infection period. Significant difference in parasite distribution on the lower region only occurred on day 7 between *Gt*-OS and *Gb*-UA groups (p=0.039); and on day 11 between *Gb*-UA and *Gt*-LA groups (p=0.014). Nonetheless, parasite distribution on the head was significantly different ($0.001<p\leq0.013$) between each of the *G. bullatarudis* groups (*Gb*-OS and *Gb*-UA) and all the *G. turnbulli* groups on day 1. But from days 3 to 5, significant pairwise difference ($0.001<p\leq0.046$) was found between all *Gb* groups and *turnbulli* strains for all fish stocks respectively at the head. However, apart from *Gb*-OS group that still showed significant difference with all *G. turnbulli* groups on day 7 ($0.001<p\leq0.016$), *Gb*-LA and *Gb*-UA rather showed significance difference ($0.001<p\leq0.037$) with *Gt3*-OS and *Gt3*-LA. Nevertheless, *Gb* on Ornamental fish showed difference significantly ($0.001<p\leq 0.013$) on the head with *Gt3* on OS and LA stocks as well as *Gt* on LA population during day 9; whereas, two groups of *Gb* (on OS and LA stocks) had significant difference with only *Gt3* on OS fish population during day 11 of the infection period. On day 13, there was significant difference in parasite distribution on the head between *Gb* and *Gt* on OS only.

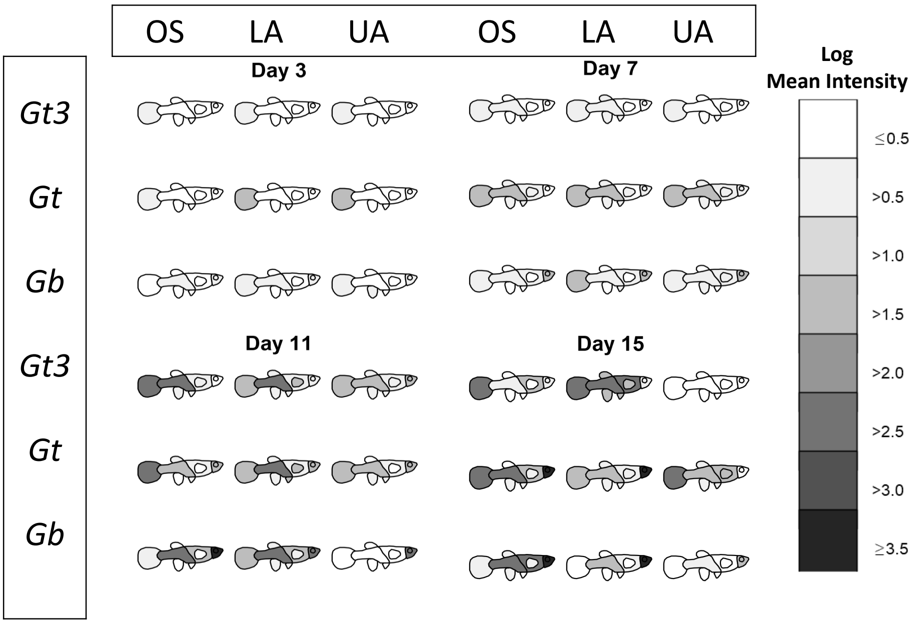

Figure 2.1: The movement of three different gyrodactylid parasites species/strain (*Gt3*, *Gt* and *Gb*) across eight host body parts (tail, lower body, anal fin, pelvic fins, dorsal fin, upper body, pectoral fins, head) of different fish stocks (Ornamental, LA and UA stocks) at four time-points. The degree of blackness indicates higher mean intensity (on log scale) over surviving fish.

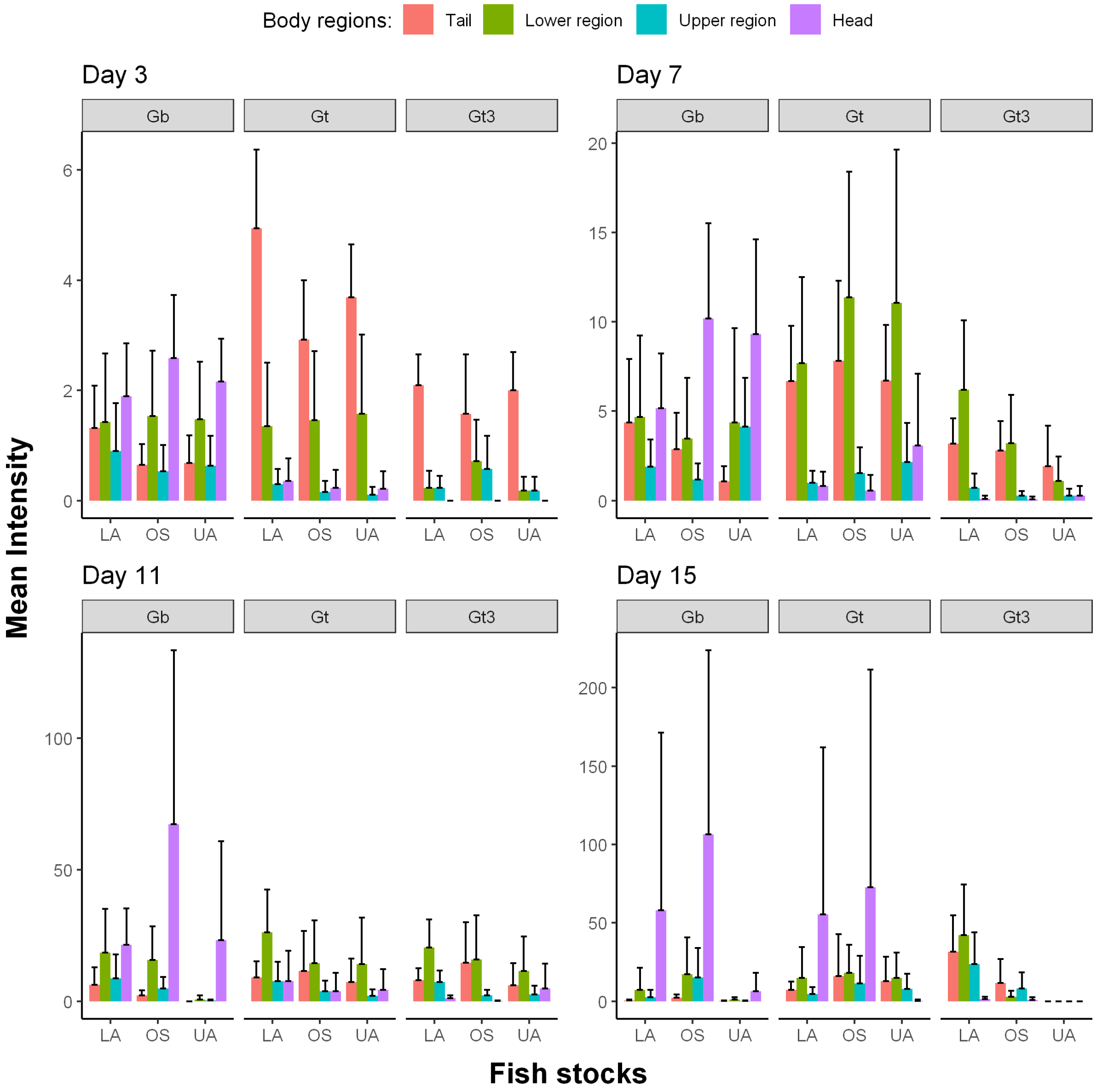

Figure 2.2: Mean intensities (with corresponding 95% confidence intervals) of three gyrodactylid strains (*Gt3*, *Gt* and *Gb*) at four main body regions (tail , lower region, upper region and head) across three fish stocks (Ornamental, LA and UA stocks) over surviving fish and across time.

Table 2.2: Peak time of gyrodactylid infection (in days) across three different parasite strains and three fish stocks for four body regions.

Parasite strains	Fish	Tail	Lower region	Upper region	Head
Gt3	OS	11	11	15	15
	LA	15	15	15	9
	UA	11	11	13	11
Gt	OS	17	13	15	15
	LA	11	11	11	15
	UA	9	9	17	9
Gb	OS	7	13	17	15
	LA	11	11	11	15
	UA	5	7	9	13

2.3.3 Multi-state Markov model for gyrodactylid infection progression

Individual fish after being infected can transition among three discrete host states: fish remains infected (state 1), fish alive with loss of infection (state 2) and fish dead (state 3), over the observation period. Let $\{X_i(t); t \geq 0\}$ be the state of fish i over time. We suppose that $\{X_i(t)\}$ is a time-inhomogeneous Markov chain with transition rate matrix $Q(t) = \{q_{rs}(t)\}$ for $r,s = 1,2,3$. For each $i = 1,2,\cdots,157$, we have observations $X_i = (X_{i0}, X_{i1}, \cdots, X_{i9})$ at times $t_0 = 0$, $t_1 = 1$, $t_2 = 3$, $\cdots$, $t_9 = 17$. The likelihood for the model parameters $\theta = \{q_{rs}(t)\}$ is given as

$$L(\theta) = \prod_{i=1}^{157} L_i(\theta|x_i), \tag{2.3}$$

where $L_i(\theta|x_i)$ is the likelihood contribution for each fish i obtained as product of state transition probabilities such that

$$L_i(\theta|x_i) = \prod_{j=1}^{9} p_{x_{ij-1},x_{ij}}(t_{j-1}, t_j) \tag{2.4}$$

with $p_{x_{ij-1},x_{ij}}(t_{j-1}, t_j) = P\{X_i(t_j) = x_{ij}|X_i(t_{j-1}) = x_{ij-1}\}$. We assumed that once a fish had lost its infection (state 2) or died (state 3), it cannot be reinfected due to the experimental design (move back to state 1) and thus the corresponding rates are 0. Hence, the transition rate matrix $Q(t)$ for the multi-state model with the three discrete host states is given as

$$Q(t) = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} q_{11}(t) & q_{12}(t) & q_{13}(t) \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \end{array} \quad \text{with} \quad q_{11}(t) = -q_{12}(t) - q_{13}(t),$$

where $q_{12}(t) > 0$ and $q_{13}(t) > 0$ are the rates at which an infected fish loses its infection and dies at time t respectively. Here, we modelled the rate matrix $Q(t)$ as a piecewise constant function with change points $t_1, t_2, \cdots, t_8$. For $t \in [t_{j-1}, t_j)$, we write $Q(t) = Q_j$. The transition probability matrix is

$$P(s,t) = (P_{ij}(s,t))_{ij} = (P(X(t) = j | X(s) = i))_{ij} = e^{\int_s^t Q(u)du}. \qquad (2.5)$$

The likelihood function for the model parameters is estimated using a maximum likelihood method, fitted using the msm package in R [159].

2.3.3.1 Estimating the probability of transition and parasite virulence given covariates

We examine how variables such as fish sex, fish size, fish stock and parasite strain may affect the transition rates $Q(t)$. Let $z_i = \{z_{i1}, z_{i2}, z_{i3}, z_{i4}\}$ be the realized values of the covariates (fish sex, fish size, fish stock and parasite strain) for fish i. Then, the transition rate matrix entries $q_{rs}(t)$ for $r, s = 1, 2, 3$ and $t \in [t_{j-1}, t_j)$ were taken as

$$q_{rs}(t, z_i) = q_{rsj}^{(0)} \exp(\beta_{rs1} z_{i1} + \beta_{rs2} z_{i2} + \beta_{rs3} z_{i3} + \beta_{rs4} z_{i4}) = q_{rsj}^{(0)} \exp(\beta_{rs}^T z_i), \qquad (2.6)$$

where $q_{rsj}^{(0)}$ is baseline intensity, and β_{rs} is a parameter vector. The likelihood is then maximized over $q_{rsj}^{(0)}$ and the regression coefficients β_{rs}; for $r = 1$ and $s = 2, 3$. The hazard ratios (HR) corresponding to each covariate are $\exp(\beta_{rs})$, for $r = 1$ and $s = 2, 3$. The transition probabilities were estimated from $q_{rs}(t, z_i)$ using equation 2.5. Given the four predictors (fish sex, fish size, fish stock and parasite strain) and two possible transitions from state 1 to either state 2 (q_{12}) or state 3 (q_{13}) in the proposed multi-state Markov model (defined by equation 2.6), there are 16^2 (or 256) possible variable permutations or models (which includes transitions independent of the underlying covariates).

A systematic variable and model selection was carried out using both Akaike information criterion (AIC) and Bayesian information criterion (BIC) statistics (due to the relative advantages of the two model selection criteria), where all possible variable permutations or models were considered. The AIC statistic assesses the model's goodness of fit while reducing the complexity of the underlying parameters; whereas the BIC statistics penalise adding more parameters or strongly penalise free parameters compared to the AIC statistic. According to Kuha [183], effective model selection can be achieved by using both AIC and BIC statistics, predominantly to identify models favoured by both criteria; although the study's methodological design, the main research questions, and the belief of a true model and its applicability to the study are crucial factors in determining whether to utilise the AIC or BIC [310]. The best model (among identified parsimonious or highly predictive models) was finally chosen based on a likelihood ratio test (LRT) at a 5% significance level. Detailed results on the variable selection for the multi-state model and its R codes (for reproducibility of results) can be found via the GitHub URL link: github.com/twumasiclement/In-Silico-Modelling-of-Parasite-Dynamics.

Let T_1 be the time spent in state 1, given that the fish or the process is in state 1 at time 0. Then, the mean sojourn time in state 1 is given as

$$E(T_1) = \sum_{j=1}^{\infty} E(T_1|\text{leave in period j}) \times \text{P}(\text{leave in period j}), \tag{2.7}$$

where

$$E(T_1 \mid \text{leave in period j}) = t_{j-1} + E(S_j|S_j \leq t_j - t_{j-1}) \tag{2.8}$$

with

$$S_j \sim \exp(q_{12}(j,z_i) + q_{13}(j,z_i)),$$

and $E(S_j|S_j \leq t_j - t_{j-1})$ is given by equation 2.11 according to Theorem 1. In equation 2.7, the probability that the process leaves in period j, denoted by P(leave in period j), is computed such that

$$\text{P}(\text{leave in period j}) = \begin{cases} P(S_j \leq t_j - t_{j-1}), & j = 1 \\ \left[1 - \sum_{j'=1}^{j-1} \text{P}(\text{leave in period } j')\right] \times P(S_j \leq t_j - t_{j-1}), & 2 \leq j \leq 7 \\ 1 - \left[\sum_{j'=1}^{7} \text{P}(\text{leave in period } j')\right], & j \geq 8 \end{cases}$$

with

$$P(S_j \leq t_j - t_{j-1}) = 1 - \mathrm{e}^{-(q_{12}(j,z_i)+q_{13}(j,z_i))(t_j - t_{j-1})} \qquad \text{for} \quad j \geq 1$$

in accordance to equation 2.10 under Theorem 1.

Theorem 1. *Let S_j be the time spent by infected fish during period j. Suppose that $S_j \sim \exp\left(q_{12}(j,z_i) + q_{13}(j,z_i)\right)$ with probability density*

$$f(S_j) = [q_{12}(j,z_i) + q_{13}(j,z_i)]\,\mathrm{e}^{-(q_{12}(j,z_i)+q_{13}(j,z_i))S_j}, \quad S_j > 0$$

where $q_{12}(j,z_i)$ and $q_{13}(j,z_i)$ are the transition rates from state 1 to state 2 and 3, respectively, given the covariates z_i for fish i; such that

$$E(S_j) = \frac{1}{q_{12}(j,z_i)+q_{13}(j,z_i)}.$$

Then,

$$E\left[S_j \mathbb{1}_{\{S_j \le t_j - t_{j-1}\}}\right] = E(S_j) - [t_j - t_{j-1} + E(S_j)]\,\mathrm{e}^{-(q_{12}(j,z_i)+q_{13}(j,z_i))(t_j-t_{j-1})} \tag{2.9}$$

and

$$P(S_j \le t_j - t_{j-1}) = 1 - \mathrm{e}^{-(q_{12}(j,z_i)+q_{13}(j,z_i))(t_j-t_{j-1})}. \tag{2.10}$$

Proof of Theorem 1.

For simplicity, let $S_j \sim \exp(\lambda)$ with probability density $f(S_j) = \lambda \mathrm{e}^{-\lambda S_j}, \quad S_j > 0$ and $E(S_j) = \frac{1}{\lambda}$, where $\lambda = q_{12}(j,z_i)+q_{13}(j,z_i)$. Suppose $\alpha = t_j - t_{j-1}$, then

$$\begin{aligned} E\left[S_j \mathbb{1}_{\{S_j \le \alpha\}}\right] &= \int_0^\alpha S_j f(S_j) dS_j = \lambda \int_0^\alpha S_j \mathrm{e}^{-\lambda S_j} dS_j \\ &= \lambda I, \quad \text{where} \quad I = \int_0^\alpha S_j \mathrm{e}^{-\lambda S_j} dS_j. \end{aligned}$$

Considering the integral I and using integration by parts,

$$I = \int_0^\alpha S_j \mathrm{e}^{-\lambda S_j} dS_j = uv\Big|_0^\alpha - \int_0^\alpha u'v dS_j,$$

where $u = S_j$, $u' = 1$, $v' = \mathrm{e}^{-\lambda S_j}$ and $v = -\frac{1}{\lambda}\mathrm{e}^{-\lambda S_j}$.

$$\begin{aligned} \implies I &= -\frac{S_j}{\lambda}\mathrm{e}^{-\lambda S_j}\Big|_0^\alpha + \int_0^\alpha \frac{1}{\lambda}\mathrm{e}^{-\lambda S_j} dS_j = -\frac{\alpha}{\lambda}\mathrm{e}^{-\lambda\alpha} - \frac{\alpha}{\lambda^2}\mathrm{e}^{-\lambda\alpha} + \frac{1}{\lambda^2} \\ &= -\frac{1}{\lambda}\mathrm{e}^{-\lambda\alpha}\left[a + \frac{1}{\lambda}\right] + \frac{1}{\lambda^2}. \end{aligned}$$

Hence,

$$\begin{aligned} E\left[S_j \mathbb{1}_{\{S_j \le \alpha\}}\right] &= \lambda I = \lambda\left(-\frac{1}{\lambda}\mathrm{e}^{-\lambda\alpha}\left[\alpha + \frac{1}{\lambda}\right] + \frac{1}{\lambda^2}\right) \\ &= \frac{1}{\lambda} - \left[\alpha + \frac{1}{\lambda}\right]\mathrm{e}^{-\lambda\alpha} = E(S_j) - [\alpha + E(S_j)]\,\mathrm{e}^{-\lambda\alpha}. \end{aligned}$$

Substituting for λ and α gives the required equation 2.9 such that

$$E\left[S_j \mathbb{1}_{\{S_j \le t_j - t_{j-1}\}}\right] = E(S_j) - [t_j - t_{j-1} + E(S_j)]\,\mathrm{e}^{-(q_{12}(j,z_i)+q_{13}(j,z_i))(t_j-t_{j-1})}.$$

Also, the required equation 2.10 can be obtained such that

$$
\begin{aligned}
P(S_j \leq t_j - t_{j-1}) = P(S_j \leq \alpha) &= \int_0^{\alpha} \lambda \mathrm{e}^{-\lambda S_j} dS_j \\
&= \lambda \int_0^{\alpha} \mathrm{e}^{-\lambda S_j} dS_j = \lambda \left[-\frac{1}{\lambda} \mathrm{e}^{-\lambda S_j} \right]_0^{\alpha} \\
&= 1 - \mathrm{e}^{-\lambda \alpha} = 1 - \mathrm{e}^{-(q_{12}(j,z_i)+q_{13}(j,z_i))(t_j - t_{j-1})}. \qquad \text{Q. E. D.}
\end{aligned}
$$

From Theorem 1, it can be deduced that

$$
\begin{aligned}
E(S_j | S_j \leq t_j - t_{j-1}) &= \frac{E\left[S_j \mathbb{1}_{\{S_j \leq t_j - t_{j-1}\}}\right]}{P(S_j \leq t_j - t_{j-1})} \\
&= \frac{E(S_j) - [t_j - t_{j-1} + E(S_j)]\,\mathrm{e}^{-(q_{12}(j,z_i)+q_{13}(j,z_i))(t_j - t_{j-1})}}{1 - \mathrm{e}^{-(q_{12}(j,z_i)+q_{13}(j,z_i))(t_j - t_{j-1})}},
\end{aligned} \tag{2.11}
$$

where

$$
E(S_j) = \frac{1}{q_{12}(j,z_i) + q_{13}(j,z_i)}.
$$

Also, given the fish or process is in state 1, then the probability of moving to state 2 or 3 next is given as

$$
\begin{aligned}
&P(\text{transition from state 1 to s}|\text{leave state 1}) = \\
&\quad \sum_{j=1}^{\infty} P(\text{transition from state 1 to s}|\text{leave in period j}) \times \mathrm{P}(\text{leave in period j}),
\end{aligned} \tag{2.12}
$$

where

$$
P(\text{transition from state 1 to s}|\text{leave in period j}) = \frac{q_{1s}(j,z_i)}{q_{12}(j,z_i) + q_{13}(j,z_i)}
$$

for $s = 2, 3$. We assume that $q_{12}(t, z_i) = q_{12}(15, z_i)$ and $q_{13}(t, z_i) = q_{13}(15, z_i)$ for $t \geq 15$.

2.3.3.2 Results on the multi-state Markov modelling

We used the time-inhomogeneous multi-state Markov model to examine the significant determinants of fish survival (fish sex, fish size, fish stock and parasite strain). The estimated hazard ratios (HR) corresponding to each significant predictor of the fitted model is summarized by Table 2.3. Figure 2.3 shows how the baseline transition rates from the infected state (state 1) to uninfected (state 2) and dead (state 3) states changed over the observed time intervals. Figure 2.4 shows that the fitted multi-state model gives a very good fit to the proportion of fish that will remain in each host infection status from the onset of infection to the end of the study period.

The likelihood of infected fish fighting off their infection was significantly influenced by fish size (HR=0.87, 95% C.I=0.76-0.99, p=0.037); such that larger fish are less likely to clear off their infection. However, fish sex, fish stock and parasite strain did influence the likelihood of infected fish dying, but not parasite extinction. Infected male fish were 52% more likely to die compared to female fish (HR=1.52, 95% C.I=1.04-2.22, p=0.031). The risk of death from the gyrodactylid infection among the OS fish (HR=0.24, 95% C.I=0.14-0.39, p<0.001) was 76% less likely compared to UA fish stock. LA fish (HR=0.39, 95% C.I=0.25-0.61, p<0.001) were 61% less likely to die from gyrodactylid infections relative to UA fish. Based on estimated hazard ratios, the rate of fish survival from the gyrodactylid infections was higher among OS stock; followed by LA stock and then UA stock. Fish infected by laboratory strain of *G. turnbulli* (HR=1.65, 95% C.I=1.03-2.65, p=0.037) were 65% more likely to die compared to the wild strain. The wild *G. bullatarudis* strain (HR=1.64, 95% C.I=1.02-2.62, p=0.039) was also 64% more likely to kill fish compared to the wild *G. turnbulli* strain. The estimates of the hazard ratios corresponding to *Gt3* and *Gb* relative to wild *G. turnbulli* strain suggest that there is no significant difference in the likelihood of fish mortality between *Gt3* and *Gb* strains. We quantified parasite virulence by estimating both the rates of host mortality (Figure 2.5) and host recovery (Figure 2.6) over time using the fitted multi-state Markov model.

We estimated the mean sojourn time in state 1 (the average amount of time fish can remain infected) and the probability of next transition from the infected state (state 1) to either recovery (state 2) or dead state (state 3) across all significant predictors (fish sex, fish size, fish stock and parasite strain) of the fitted multi-state Markov model. For any strain of gyrodactylid, large Ornamental female fish remained infected longer than fish with any other attributes (Table 2.4). Fish infected with the wild *G. turnbulli* strain on average remained infected longer than fish infected with *Gt3* or wild *G. bullatarudis* strains before either recovering or dying, irrespective of the fish size, stock and sex. The mean time for fish to remain infected with any parasite strain before fighting off their infection or dying was between 6 and 14 days. It was found that an infected fish had a higher probability of dying than recovering from the infection irrespective of the type of gyrodactylid infection, fish stock, sex, and size (Table 2.5). Large male fish were more likely to die than small or medium-sized male or female fish of any size; whereas the chance of host recovery was higher among OS fish stock compared to the Trinidadian fish stocks. The fish infected with wild *G. turnbulli* strain had a greater probability of fighting off their infections than fish infected with either *Gt3* or *Gb* strain.

Table 2.3: Estimated hazard ratios (HR) from the multi-state Markov model across significant predictors (fish sex, fish size, fish stock and parasite strain) with their respective 95% confidence intervals (C.I).

Covariates	Transitions	HR	Lower C.I	Upper C.I	p-value
Fish size	$1 \to 2$	0.87	0.76	0.99	0.037
Fish sex					
Male (Ref: Female)	$1 \to 3$	1.52	1.04	2.22	0.031
Fish stock					
OS (Ref: UA)	$1 \to 3$	0.24	0.14	0.39	<0.001
LA (Ref: UA)	$1 \to 3$	0.39	0.25	0.61	<0.001
Parasite strain					
Gt3 (Ref: *Gt*)	$1 \to 3$	1.65	1.03	2.65	0.037
Gb (Ref: *Gt*)	$1 \to 3$	1.64	1.02	2.62	0.039

Table 2.4: Mean sojourn time (in days) for fish to remain infected across significant predictors (fish sex, fish size, fish stock and parasite strain) based on the fitted multi-state Markov model.

Parasite	**Fish**	**Male fish**			**Female fish**		
strain	**stock**	Small (11 mm)	Medium (17 mm)	Large (26 mm)	Small (11 mm)	Medium (17 mm)	Large (26 mm)
Gt3	OS	10.69	11.33	11.79	11.40	12.13	12.64
	LA	9.52	10.03	10.40	10.49	11.12	11.56
	UA	6.78	7.04	7.22	8.06	8.43	8.69
Gt	OS	11.52	12.26	12.79	12.01	12.81	13.37
	LA	10.67	11.31	11.76	11.38	12.11	12.62
	UA	8.32	8.71	8.99	9.49	10.00	10.36
Gb	OS	10.71	11.35	11.81	11.42	12.14	12.66
	LA	9.54	10.06	10.43	10.51	11.14	11.58
	UA	6.81	7.07	7.25	8.09	8.46	8.72

Table 2.5: Probability of next transition from the infected state 1 to either the recovery state 2 or the dead state 3 across significant predictors (fish sex, fish size, fish stock and parasite strain) based on the fitted multi-state Markov model.

		Male fish						**Female fish**					
Parasite strain	**Fish stock**	Small (11 mm)		Medium (17 mm)		Large (26 mm)		Small (11 mm)		Medium (17 mm)		Large (26 mm)	
		p_{12}	p_{13}	p_{12}	p_{13}	p_{12}	p_{13}	p_{12}	p_{13}	p_{12}	p_{13}	p_{12}	p_{13}
Gt3	OS	0.460	0.540	0.357	0.643	0.266	0.734	0.565	0.435	0.460	0.540	0.358	0.642
	LA	0.338	0.662	0.249	0.751	0.177	0.823	0.436	0.564	0.335	0.665	0.247	0.753
	UA	0.177	0.823	0.122	0.878	0.082	0.918	0.238	0.762	0.168	0.832	0.116	0.884
Gt	OS	0.586	0.414	0.481	0.529	0.378	0.622	0.683	0.317	0.587	0.413	0.483	0.517
	LA	0.457	0.543	0.355	0.645	0.264	0.736	0.562	0.438	0.456	0.543	0.355	0.645
	UA	0.253	0.747	0.179	0.821	0.124	0.875	0.335	0.665	0.246	0.754	0.176	0.824
Gb	OS	0.463	0.537	0.359	0.641	0.268	0.732	0.567	0.433	0.462	0.538	0.360	0.640
	LA	0.340	0.660	0.250	0.750	0.179	0.821	0.438	0.562	0.337	0.663	0.249	0.751
	UA	0.178	0.822	0.123	0.877	0.083	0.917	0.240	0.760	0.169	0.831	0.116	0.884

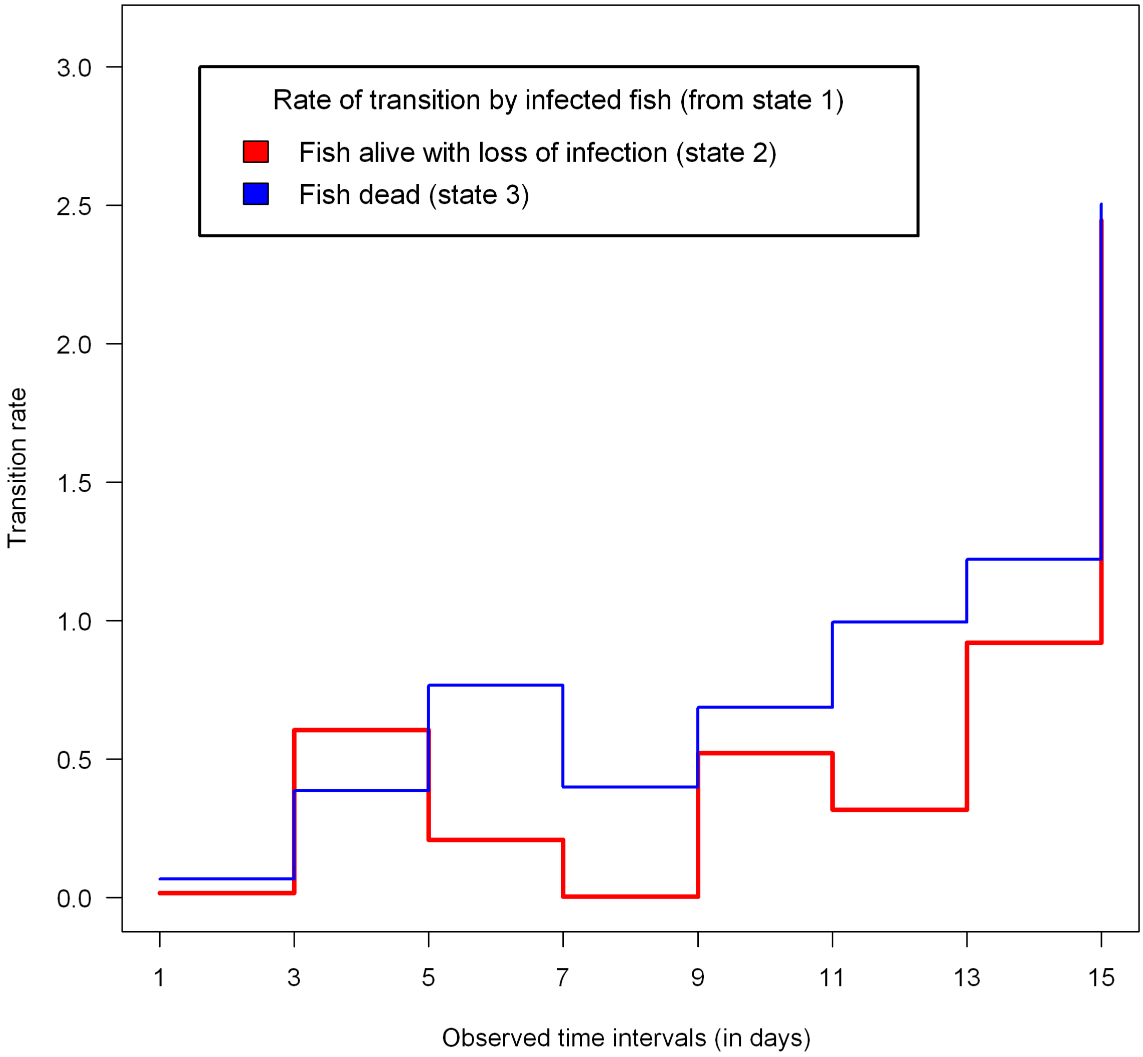

Figure 2.3: Piecewise-constant plot of estimated baseline transition rates from infected host state to uninfected and dead states at different observed time intervals in the time-inhomogeneous multi-state Markov model.

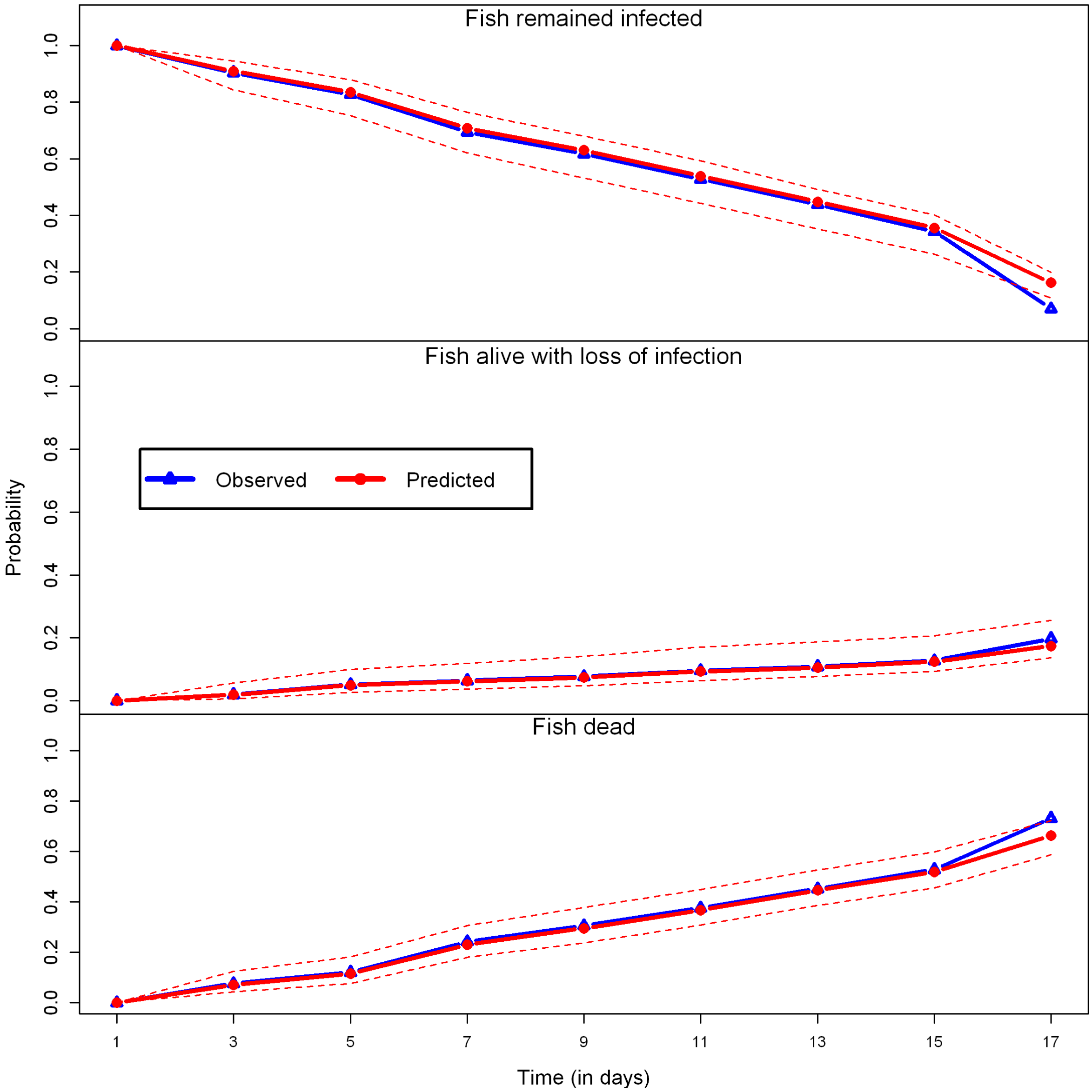

Figure 2.4: Comparison between observed and expected proportion of fish that will remain in each host infection state from days 1 to 17 after the onset of gyrodactylid infection based on the fitted multi-state Markov model (mean absolute percentage error=7.85%).

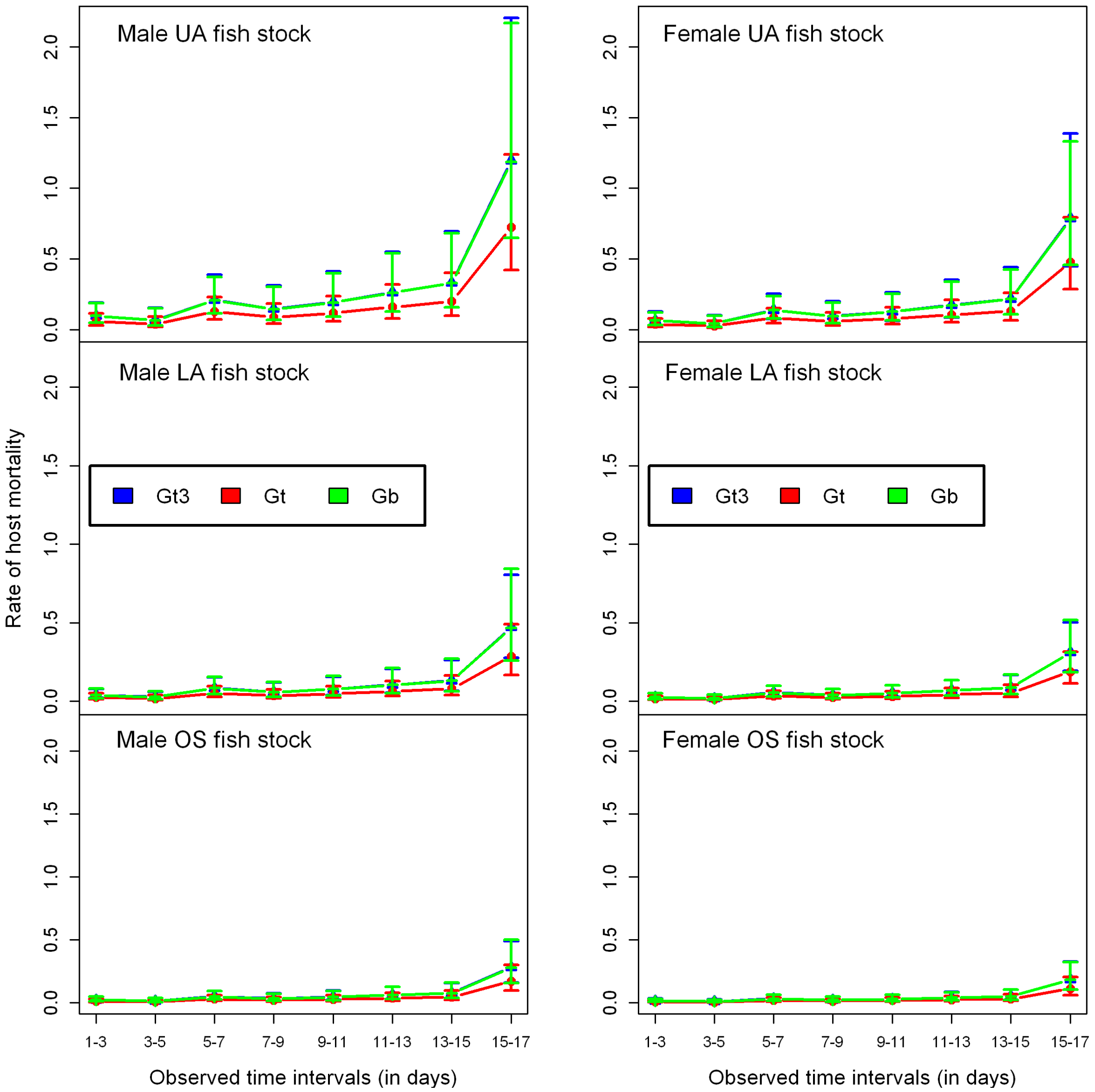

Figure 2.5: Predicted host mortality rates of parasite strains (*Gt3*, *Gt* and *Gb*) on the fish stocks (Ornamental, LA and UA stocks) over time for both male and female fish respectively.

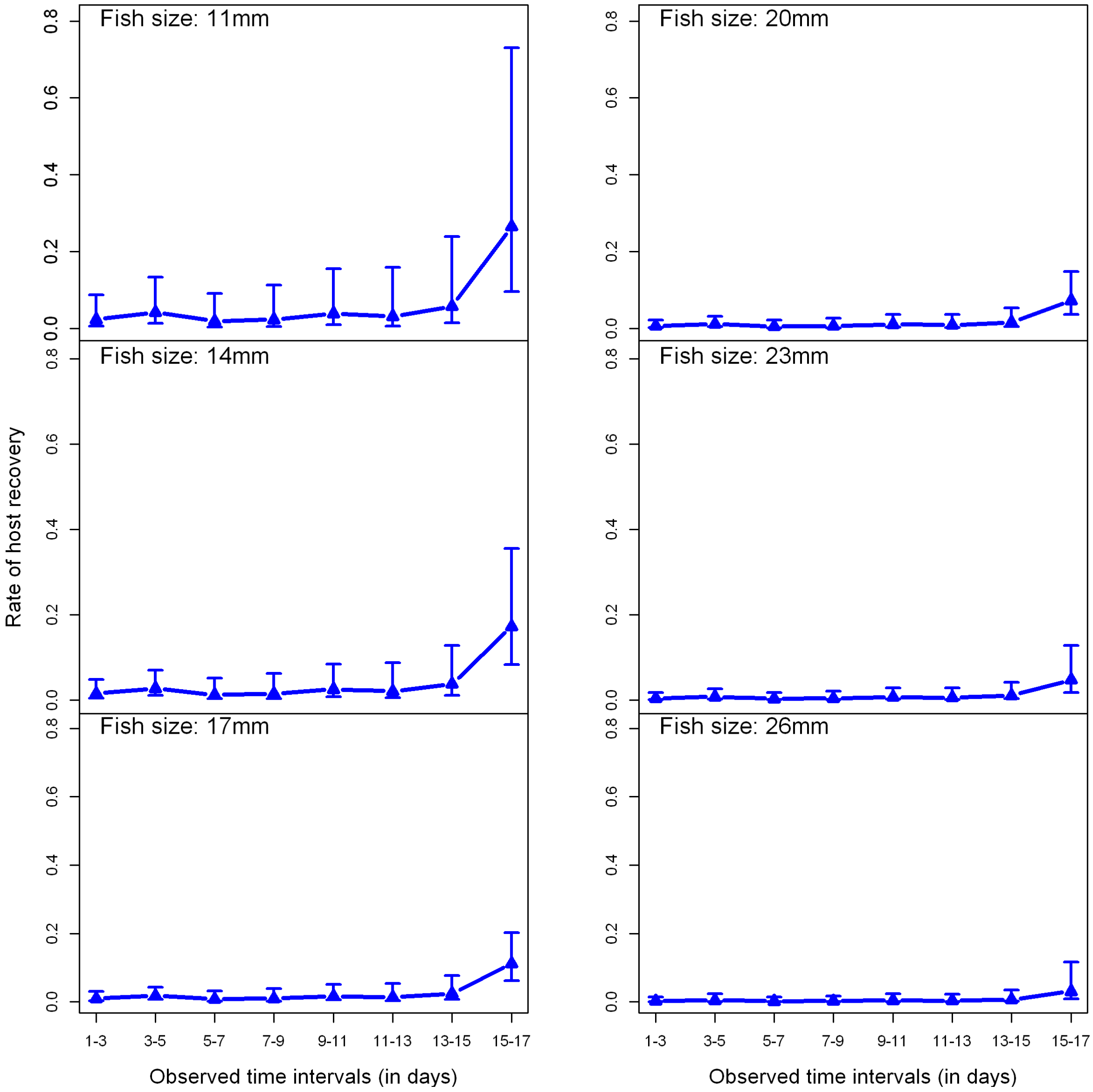

Figure 2.6: Predicted host recovery rates over time at different fish sizes (11, 14, 17, 20, 23 and 26 mm).

2.4 Discussion

2.4.1 Insights into the gyrodactylid-fish system

In this study, we built on previous studies of the infrapopulation dynamics of three different gyrodactylid strains (two strains of *G. turnbulli* and one strain of *G. bullatarudis*) among three different fish stocks (OS, LA and UA stocks) in relation to parasite habitat preference, host survival and parasites' virulence [see 129, 132, 56, 287]. Here, we re-analysed empirical data to investigate further our understanding of these parasites' spatial and temporal variation. Concerning parasite habitat preference, it was previously hypothesised that there is a niche partition with *G. turnbulli* occurring caudally [129] and *G. bullatarudis* rostrally [132]. We have confirmed here that the microhabitat preferences of the *G. turnbulli* (laboratory and wild type) and *G. bullatarudis* strains depend on the type of host, and can change over time for the wild *G. turnbulli* strain. A quantitative measure of the significance of differences in spatial parasite distribution over time and across different fish stocks has been established using a Multivariate Kruskal-Wallis test and associated post-hoc statistical tests. In the previous analysis of the existing empirical data, a binomial logistic regression was employed to estimate survival rates such that fish classified as either dead or alive at the end of the observation period [56]. Information about whether fish alive either remained infected or recovered from the infection was not included in the earlier model, which could have impacted the significance of some covariates and other estimates. In addition, the virulence of the gyrodactylid strains was only measured by the proportion of host casualties [56]. Here, using a more sophisticated mathematical model, we have been able to include host recovery, and fish sex was identified as a significant factor of host survival compared to the previous study [56]. We have also estimated for the first time, the average duration that fish can remain infectious and the probability that infected fish will either recover or die from each of the three parasite strains across the three guppy populations, sexes, and different fish sizes (small, medium and large sizes). We have quantified both host recovery and mortality over time as relevant metrics of parasite virulence for the gyrodactylid-fish system.

As noted in this study, the captive-inbred *G. turnbulli* strain preferred the tail of three different fish stocks (Ornamental, Lower Aripo River and Upper Aripo River stocks); whereas the wild *G. turnbulli* initially preferred the tail but then switched to the head. The wild *Gyrodactylus bullatarudis* consistently showed a rostral preference on all fish. The mean intensity of parasites was higher on OS and LA fish than UA stocks across all body regions over time, probably related to the higher mortality of the UA fish. Lower numbers of parasites on the pectoral, pelvic, dorsal and anal fins compared to the tail, lower body, upper body, and head regions might be affected by fish being maintained in isolation or due to difference in the surface area of these body regions. Individual host isolation meant there was no opportunity for host-to-host transmission to occur via the fins [as suggested by 129]. Thus, the parasites might be making a behavioural decision to enhance their fitness in response to the absence of alternative hosts and or reduce competition at small-sized body regions over time. The peak time to infection varied spatially across parasite strains and fish stocks. Such variation likely represents a trade-off between successful parasite exploitation and the host's localised immune response [reviewed by 15]. Parasite distribution on infected hosts could also be driven by multiple abiotic and biotic factors [137, 141, 200, 257].

The fitted multi-state model revealed that fish sex, fish stock and parasite strain influenced fish mortality. LA and OS fish stocks survived for longer than UA fish. The Ornamental guppy population was infectious longer than the Trinidadian fish stocks (LA and UA fish) based on the estimated average duration of infection, however, the OS guppies had a higher chance of host recovery compared to the LA and UA fish stocks. The OS guppies were twice as likely to fight off the infection even with a higher infection mean time than the Trinidadian fish stocks due to superior innate immune defences or immunocompetence towards single-species infections [56, 203, 303, 304]. Larger fish were infectious over a longer period than small or medium-sized fish, whereas female fish from all three guppy populations experienced a longer duration of infection than male fish. It was found that fish infected by the wild strain of *G. turnbulli* on average remained infected longer than the laboratory *G. turnbulli* strain and the wild *G. bullatarudis* strain. Additionally, probability of host recovery from the wild *G. turnbulli* strain was consistently higher among all fish stocks and sexes. Previous experiments showed that LA guppies have a superior immune response to the UA fish [303]; whilst the Ornamental guppy stock performed better in parasite resistance than the UA stock [56]. This current study revealed that a longer period of host infection leads to a higher chance of host recovery and a smaller chance of host mortality. Thus, the low mean parasite intensity and low infection duration among UA guppies compared to OS and LA stocks, explains why UA fish were more likely to die over time relatively. As in the previous study, the laboratory strain of *G. turnbulli* and wild strain of *G. bullatarudis* were more likely to cause fish mortality than the wild strain of *G. turnbulli*, but we found that infected male fish were twice as likely to die relative to female fish. The main reason for this new finding of fish sex as a significant determinant of host mortality is the use of a multi-state model that is able to incorporate host mortality and recovery simultaneously. Other parasite-fish studies have identified fish sex as a significant factor of host mortality [318]. Only fish size significantly influenced the rate of infection loss; namely larger fish acquired more parasites as infections progressed resulting in low parasite extinction compared to smaller fish [306]. Host recovery is thus dependent on host size, with probability of infection loss low but increasing gradually over time. Nevertheless, it was found that the chance of host mortality was more likely to occur than host recovery irrespective of host size.

Parasite virulence was described in terms of host mortality and recovery. We found that host mortality and host recovery are significantly time-dependent and generally higher towards the end of the infection period. Previously, *Gt3* was identified as causing most host deaths, followed by *G. bullatarudis* and then the wild *G. turnbulli*; but their respective host mortality rates were not quantified, nor did we previously consider how this changed over time, nor the effect of the different fish stocks [56]. Here, we found no significant difference in host mortality rates between *Gt3* and *Gb* parasite strains over time. Male fish from the three different guppy populations (OS, LA and UA stocks) consistently had a higher rate of host mortality than female fish stocks over time. This could be explained by the fact that the female fish are infectious longer than the male fish as revealed by the estimated mean sojourn time of infection; and thus, the female host populations are able to develop innate or adaptive host immunity faster than the male fish stocks over time.

In summary, both host-parasite and strain-specific microhabitat preferences were identified. The multi-state Markov model was also usefully employed to provide temporal and additional insights into host survival and parasite virulence (described by host mortality and recovery). The multi-state Markov model assumed that fish could not be reinfected after infection loss due to the study's experimental design, but this could be adapted in future studies to include transmission. The model's results may change due to reinfection in a social setting where population mixing between hosts is possible. The developed time-inhomogeneous multi-state Markov model could be extended and applied to a range of different host-parasite systems.

2.4.2 Mathematical implications of the study findings

The current study could inform the modelling of other biological systems and survival analyses where the entire infection history of an individual (or host) is of interest. Multi-state models provide a robust approach to modelling almost any kind of longitudinal time-to-event data. In the current multi-state Markov model, we could not include spatial information and other relevant information about parasite fecundity, age group (young or old parasite), parasite mortality, parasite mobility, and host immune response while exploring host survival and parasite infrapopulation dynamics. Hence, a more sophisticated (individual-based) stochastic simulation model including these data will be needed to understand the gyrodactylid-fish system better. The findings from the spatial-temporal parasite dynamics will impact some aspects of the sophisticated stochastic simulation model (for the gyrodactylid-fish system) in several ways, including the specification of the model's discrete states and model assumptions as well as other specific hypotheses to investigate (as presented under Chapter 6). Such a stochastic model (conditioned on relevant information such as fish sex, fish size, fish type, and parasite strain) should be able to simulate the spread of different age parasites over the external surfaces of fish over at least a 17-day infection period with population carrying capacity (dependant on host size and area of host body regions).

Unlike the multi-state Markov model, the novel individual-based stochastic model can investigate specific hypotheses of interest or open biological questions concerning gyrodactylids' birth, death, mobility rates of the different parasite strains, and immune responses of different host populations (amongst others). Furthermore, the stochastic simulation model can predict the infrapopulation dynamics of the gyrodactylid-fish system beyond the standard 17-day experimental period and be modified to investigate mixed-gyrodactylid infection dynamics on a single host (which is unknown to biologists).

Based on findings from the spatial and temporal parasite dynamics of the *Gyrodactylus* species (as a result of lower mean parasite intensities at some host body areas as observed in this study), the stochastic individual-based model must be set up such that for each fish host, the eight body regions (tail, lower body, upper body, anal fin, dorsal fin, pelvic fins, pectoral fins, and head) must be re-categorised into four major body locations: tail, lower region (comprising of the lower body, anal fin, pelvic fins and dorsal fin), upper region (made up of the upper body and pectoral fins) and the head (as observed in Figure 2.2).

The multidimensional stochastic model should be parameterised by young and older parasites' birth, death, and movement rates conditioned on the host's immune response and parasite strain. Host mortality should be assumed to occur at a rate proportional to the total number of parasites on fish (and conditioned on host sex and size). However, the microhabitat-specific immune responses occur as a function of the parasite abundance in each of the four major body regions of the host (and are dependent on host sex and fish type). The stochastic model should also include parasite body preference for moving back and forth on the host (conditioned on parasite strain). Consequently, the sophisticated stochastic simulation model will help answer other open biological questions as well as serve as a more realistic simulator to aid in the data-generation process of this system and conduct (numerical) biological experiments that are difficult to explore or control within standard experimental settings.

In general, the statistical tests and the multi-state model used in this study could be helpful for other host-parasite systems. Specifically, the rank-based multivariate Kruskal-Wallis test (with its post-hoc tests) and the time-inhomogeneous multi-state Markov model could be adopted for other biological systems to investigate the spatial-temporal parasite distribution as well as host survival. We did not estimate other relevant epidemiological parameters (such as the probability that a recovered host will be reinfected or die and the mean sojourn time for a recovered individual to remain uninfected or reinfected) from the multi-state model, as for this study, we focused just on parasite infrapopulations. Future studies will examine host-to-host transmission to holistically understand the spread of gyrodactylid parasites and the host-parasite interactions among different populations of fish.

Chapter 3

Review of mathematical models for host-parasite systems

3.1 Introduction

Mathematical modelling of biological or host-parasite systems has seen tremendous developments and broad applications in the field of theoretical and applied ecology [30]. Mathematical models give a logical framework for developing, testing, and evaluating ecological hypotheses and biological systems. Most of these classical models are applied either at the individual or population level. Hence, these models can be categorised as either individual-based models (IBMs) or population-based models (PBMs). IBMs strictly model each individual by keeping track of the state of each member of the population; whereas, PBMs keep track of the total number of individuals in each state. However, common modelling setbacks are primarily attributed to the underlying assumptions made about the models, which are either too simple or too complex. Consequently, it makes some mathematical models unrealistic with respect to the biological or host-parasite system under study [77]. All the mathematical models reviewed in this section and under subsequent sections of Chapter 3 are models developed from previous studies.

Even though individual hosts or parasites may differ genetically, physiologically, or behaviourally, one of the most common modelling assumptions in PBMs is that the distribution of hosts' parasites can be aggregated into a single state variable that signifies population size [113, 322]. Many traditional ecological or epidemiological models, such as logistic, Lotka-Volterra or predator-prey, and compartmental epidemic models, assume that all individuals (within a subgroup) are identical and can be lumped together to represent the population size (as a single state variable). However, the single state variable modelling principle is only valid in a practical sense from two different perspectives. That is i) when there is no information loss by considering population averages or aggregations or ii) all parasites or individuals represented by a population number are identical. Individual variations between organisms are recognised in this viewpoint, but they are irrelevant in the context of the modelling problem under investigation; therefore, models that average the entire distribution of the individuals are statistically sufficient.

Nonetheless, the setbacks regarding population aggregation are that biological organisms are naturally distinct in physiological and behavioural traits (determined by their specific genetic, age structure and other environmental factors); and for spatially dependent systems, organisms or individuals mainly affect other organisms within their spatial-temporal neighbourhood [60, 219]. Also, a small number of hosts are infected with many parasites, while most hosts are either parasite-free or have a lower parasite load over time. The origin of this nearly universal pattern is critical to our knowledge of host-parasite interactions, and it significantly impacts various aspects of their ecology and evolution. Nevertheless, the data-generating processes that characterise parasite aggregation in the standard statistical framework or parasitological data are mostly not explicitly described or further explored [113]. Consequently, many studies have been done to bridge all these modelling gaps associated with PBMs by adopting IBMs through computer simulations or *in silico* models [153]. According to Metz and Diekmann [219], individual-based modelling approaches can be broadly categorised into either i) individual state (or i-state) distribution and ii) i-state configuration models. The mathematical modelling of the i-state distribution methods are dominated by classical transition matrix (e.g., Markov models [298, 328]), and partial differential equation (e.g., reaction-diffusion equations for Lotka-Volterra competition system [222]) approaches. In contrast, Monte Carlo computer modelling is the central methodology used in the i-state configuration approach (e.g., stochastic simulation algorithms [12, 36]). The classification of individual-based ecological models into i-state distribution and i-state configuration models reflect only a technical component of the model structure [219]. Thus, it implies that there are two numerical approaches to define individuals in the model. Nonetheless, this classification provides little useful information from a biological standpoint, as well as in terms of the model's ability to describe and predict system dynamics [301]. Incorporating evidence from the individual level to investigate processes at the population, community, and ecosystem levels can also improve PBMs.

In summary, IBMs have advantages over PBMs making them relevant to adopt in both epidemiological and parasitological studies [30]. Additionally, IBMs can integrate the intrinsic stochastic nature of infections, possible interactions among parasites and their hosts, and other events such as host immune response. Although IBMs appear more useful and accurate than PBMs due to their flexibility and modelling accuracy, they require enormous computational efforts since birth, death, and possible movements are tracked for each parasite and plausibly their host, especially for spatially explicit IBMs. Despite the advantages of IBMs in modelling host-parasite systems, only a few have been developed or applied to parasitological studies, for instance the gyrodactylid-fish system, since they require in-depth information about the biology of both parasites and their host [107, 198, 306]. Also, the tractability of some parasites and their host populations makes them highly preferable for IBM [306].

In general, modellers require expert knowledge from biologists to summarise a biological system of interest, according to the study's goal, to model the biological system [96]. Thus, the biological aspect of model building entails providing observed or experimental data, expert opinion, or knowledge of similar systems. In contrast, the modeller side focuses on selecting and adapting existing methods or developing new ones appropriate for the system and biological questions under investigation. The intersection of biology and mathematical modelling is vital for defining the model's framework (within or between hosts, populations, type of parasitic infection, etc.) and identifying general knowledge alongside implausible or contentious aspects [96]. The biological hypotheses of interest can then be formulated. Thus, biological knowledge of the host-parasite system under investigation provides the needed elements to formulate and evaluate mathematical models; whereas, modelling leads to the formation of novel hypotheses and the identification of relevant information and research gaps that should be examined through experimental or observational research. Figure 3.1 is a conceptual framework summarising the relationship between the knowledge of the biological system and the mathematical model together with model fitting and hypothesis testing.

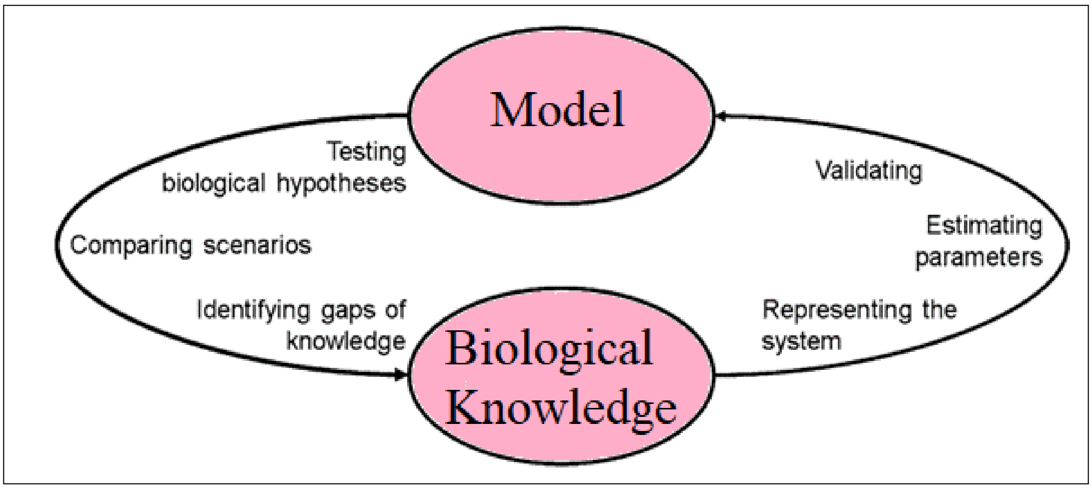

Figure 3.1: Mutual input of biology and mathematical modelling [adapted from 96].

The subsequent sections of Chapter 3 are organised as follows: Section 3.2 gives a brief history about mathematical modelling of infection dynamics for host-parasite systems; whereas, section 3.3 discusses the modelling of host-parasite interactions for both microparasitic and macroparasitic infections, and reviews existing PBMs and IBMs for modelling of host-parasite systems. Finally, section 3.4 presents previous works on the modelling of *Gyrodactylus* infections.

3.2 Brief history of host-parasite mathematical models

Over the years, mathematical models of parasitic infections have aided in predicting host-pathogen interaction outcomes, testing pertinent hypotheses, and assessing general or unknown knowledge regarding disease transmission and evolution in various scenarios [8, 172]. However, they have not reached their full potential due to their underlying limitations, including their robustness and model assumptions. Consequently, it has resulted in the development of new and improved methodologies in recent times. Furthermore, with the rise in the frequency of emerging and re-emerging infectious disease outbreaks in recent decades, the use of mathematical models to generate short-term and long-term predictions have grown [65]. Hence, it is necessary to evaluate the assumptions that underpin disease transmission and control models as well as how they influence estimates of essential epidemiological parameters and epidemic projections. Their applications range from parasite population biology and between or within-host dynamics to implementing intervention strategies against pathogens of public and animal health interests. Nonetheless, it was not until Louis Pasteur (the founder of modern immunology) developed the "germ theory of diseases", which asserts that micro-organisms known as pathogens or germs (e.g., bacteria, viruses, fungi, and viroids, amongst others) can cause disease transmission, that the scientific study of infectious disease epidemiology began in earnest [98]. Then, in 1890, Robert Koch (also known as the father of microbiology with Louis Pasteur and as the father of medical bacteriology) discovered a vaccine therapy (known as tuberculin medication for tuberculosis) based on the germ theory [114, 185].

The study of parasitic infections was pioneered by John Graunt, who first applied basic numerical techniques (without the use of a mathematical model) for competing risk analysis of several diseases and other causes of death over 20 years in a study on the *Bills of Mortality* for London parishes (which included diseases such as smallpox, swinepox, convulsion, plague, and measles, amongst others) as early as 1662 [109]. In the twentieth century, Chiang and other modellers formally developed mathematical models for competing risk analysis for human and animal mortality [63, 95, 277]. Daniel Bernoulli is credited with being the first to use a deterministic model based on a differential equation for the number of survivors in a cohort of individuals subjected to smallpox epidemic disease in 1760, where he also investigated the efficacy of variolation treatments against smallpox [8]. In 1906, Hamer's study of measles led to the popular term known as the "law of mass-action", the fundamental principle for the compartmental model of disease spread in mathematical epidemiology, under the homogeneous-mixing assumption

[8, 123].

This law in infectious disease dynamics among human and animal populations assumes that the infection transmission rate in a given well-mixing population is proportional jointly to the number of infected (I) and susceptible (S) individuals; such that new infections occur at a rate equal to either βSI (for density-dependent transmissions) or $\beta S\frac{I}{N}$ (for frequency-dependent transmissions) where β is the transmission or effective contact rate (which depends on the per capita contact rate and transmission probability) and N is the total population size [320]. The effective contact rate (β) can also be age-dependent, and methods for estimating age-stratified contact rates exist in the literature [see 241]. It can be challenging to fit sophisticated or multi-parameter compartmental models to various data sources with the needed uncertainty estimates for major model parameters of interest (including β). Given a (possibly multi-parameter) compartmental model (whose likelihood function is mostly mathematically intractable or unknown), likelihood-free methods of inference such as bootstrap filter [112], approximate Bayesian computation [209, 288], and synthetic likelihood methods [325] (amongst others) have been proposed in the literature to estimate underlying model parameters (including the effective contact rate).

Density-dependent transmission models assume that transmission scales linearly with population density (where the population density is the population size per area) and that the mean number of contacts with infected individuals is dependent on the density; in contrast, frequency-dependent transmission models assume that infection transmission or the rate of contact is independent of both population size and the density [29]. In both classes of models, homogeneous mixing is assumed with no specified geographical or social structure [101]. The mode of transmission is frequently used to guide the choice between these two models, though the transmission mechanism is not explicitly considered in deriving their respective infection rates. Heesterbeek [144], on the other hand, revealed that the mass-action principle in epidemiology was formally established by Ronald Ross and Anderson McKendrick (in 1902), who arrived at the concept simultaneously but via distinct paths (contrary to what has been believed over the years). More specifically, Anderson McKendrick identified the universal application of the mass-action principle, whereas Ronald Ross developed it from an innovative chemical standpoint. Ross further constructed a continuous-time model based on Hamer's discrete model for studying malaria disease infection [8]. In a nutshell, the law of mass-action made epidemiology a science [144]. Then, in 1927, Kermack and McKendrick proposed the epidemic threshold theorem and developed basic compartmental or deterministic models for examining the transmission dynamics of viral and bacterial infectious agents within a population of hosts, amongst other [86, 171]. Although the classical compartmental models such as the Susceptible-Infectious-Removed (SIR) models are typically referred to as the Kermack and McKendrick models, it has been pointed out that the 1927 paper discusses a more generic framework with fewer assumptions. For a general overview of SIR models, see a few works by Bohner et al. [42], Satsuma et al. [266], and Weiss [314].

Reed and Frost, and Major Greenwood developed discrete-time stochastic models in 1928 and 1931, which proceeded via generations of infectives [74]. Bartlett [25] investigated a continuous-time stochastic SIR model, which sparked a vast literature. Finally, in 1931, Greenwood [116] proposed the idea that randomness could play a role in infection transmission and that transmission could happen or not happen with a certain probability during a given contact. Furthermore, in a study on the mathematical theory of infectious diseases and its applications in 1957, Bailey [14] presented both deterministic and stochastic epidemic models, as well as the estimation of their parameters. The basic reproduction concept (which estimates the average number of infected contacts per infected individual) had its origins in the work of Ronald Ross, Alfred Lotka, and others; however, it was first applied in epidemiology in 1952 by George Macdonald (who constructed population models of the spread of malaria disease) [285]. For a more detailed historical perspective about the basic reproductive number (R_0), which is a pivotal epidemic quantity or an epidemic invasion criterion (for $R_0 > 1$), see works by Heesterbeek

[145]. The three core principles underpinning modern theoretical epidemiology are the mass-action principle coupled with the epidemic threshold theory and randomness in disease transmission [96]. The use of mathematical modelling has increased substantially in the last two decades across all areas of infectious disease research, from ecology, biology, microbiology and pathogen evolution to large-scale epidemiology and public health [9, 18, 85, 87, 168, 208, 328].

3.3 Literature on modelling infection dynamics

3.3.1 Introduction

Each member or individual of a population is classified according to their state (i.e., the characteristics that influence their ability to acquire and spread infection). At a minimum, an individual's state indicates their disease status, such as whether they are susceptible to infection, infectious, or have recovered from infection. Other characteristics, such as age or spatial location, may be necessary to characterise the states of individuals, depending on the level of information used by our host-parasite models. Generally, depending on the location of parasites' microhabitat, they can be categorised as either endoparasitic (i.e., parasites that live inside the host such as intercellular parasites like parasitic worms, protozoans of vertebrates, and other helminths, etc.) or ectoparasitic (i.e., parasites that reside outside the host such as monogenean parasites, arthropods, and other protozoans, amongst others). These groups of parasites can also be either microparasitic or macroparasitic in relation to their size, generational time length, and mode of reproduction.

The classification of ecological models as either IBM or PBM can be much more meaningful when based on some biologically essential attributes of the models. Four appropriate classification criteria are [301]: i) the extent to which the complexity of the life cycle of an individual (e.g., host or parasite) is reflected in the model, ii) whether or not resource dynamics (such as food, space, or habitat quality) are explicitly taken into account, iii) whether or not real or natural numbers are used to represent population size, and iv) the degree to which variations among individuals of the same age is taken into account. These criteria can be used to classify most theoretical ecology models that describe the dynamics of ecological systems. Hence, PBMs and IBMs are two common disease modelling frameworks for host-parasite systems. Despite the disparities between PBMs and simulation-based IBMs, researchers have discovered that the two frameworks share similarities and other hybrid models which combines IBMs and PBMs exist by leveraging their respective advantages [311]. A thorough discussion about the similarities between IBMs and PBMs, as well as the application of some existing hybrid models, can be found in a study by Gallagher [108]. For disease modelling, IBMs are effective when limited knowledge about a developing outbreak, and thus, IBMs that sufficiently reflect reality and provide useful insight for decision-making are desired by mathematical modellers [44]. The modelling of host-parasite systems (for both microparasitic and macroparasitic infections) together with an overview of existing population-based and individual-based models of host-parasite systems are presented in sections 3.3.2–3.3.4.

3.3.2 Modelling host-parasite systems

Mathematical models help better understand how infections spread within or between host populations and provide a simple summary of epidemiological data [28]. Biological or general ecological systems may include the interaction between competitors for limited resources (e.g., food), infrapopulation or interpopulation variations, the effects of mutualists or the trophic effects of predator-prey interactions. The link between parasites and host populations can be viewed as an extension of the predator-prey relationship in general [231]. Infectious agents are categorised into microparasites and macroparasites, partly dependent on the type of model required to characterise infection transmission. Thus, to model disease infection among animal or human populations, it is necessary to distinguish between the type of parasitic infection. Ecologists and epidemiologists benefit from the distinction between macroparasites and microparasites since these groups of parasites vary considerably in terms of within-host replication, ability to induce a lasting host immune response, and how they are quantified in natural populations. The choice of the modelling approaches for host-parasite interactions is thus partly dependent on the classification of parasitic infection. In addition, in-depth knowledge about the system is crucial in modelling host-parasite interactions and formulating relevant hypotheses (refer to Figure 3.1). Hence, modelling approaches are utilised in various ways depending on what is known about the biological system under study and the research questions to be addressed [146].

The first goal to model host-parasite interactions of any biological system could be to summarise existing knowledge and build a formal representation of the system to make it easier to grasp the underlying complex processes and establish general qualitative assumptions with the help of existing empirical data [96]. For instance, prior to developing a novel stochastic simulation model to understand the gyrodactylid-fish system better, the current study had to understand the biology of the system (see section 1.4) and descriptively explore the spatial-temporal parasite dynamics of *Gyrodactylus* on their fish host as well as test other hypotheses using multi-state Markov model and advanced statistical tests (see Chapter 2). Possible model representations of general biological systems include but are not limited to analytical formulations (e.g., deterministic dynamical systems, stochastic processes, etc.), computer-based or simulation-based models (e.g., Monte Carlo simulation algorithms), and graphical models (e.g., social network models, decision trees and directed or undirected graphical models, amongst others).

After model identification and understanding the biological system being modelled, the second aim is to determine the relative importance of each of the numerous mechanisms involved in system dynamics (important, secondary, or irrelevant) [96]. A detailed description of the system is then required with explicitly stated assumptions and biologically relevant model parameters. The model can then test biological hypotheses (explaining the model's structure, parameter values, or underlying transition function) by comparing distinct sub-models dynamical behaviour that includes or excludes the hypotheses. The multi-state Markov model (described in section 2.3.3) and other investigations about the infrapopulation dynamics of the three different *Gyrodactylus* strains (presented in section 2.3.2), for example, gave temporal and extra insights into parasite virulence, microhabitat preferences, host survival, and the significant factors influencing these disease outcomes across three different guppy populations. Consequently, the findings helped define novel biological questions, other model assumptions, and the model parameters to consider in the complex stochastic model for the gyrodactylid-fish system (developed in Chapter 6).

Third, suppose the mathematical model has been fitted or calibrated and validated based on the observed empirical data of the underlying biological system [96]. In that case, it can be used to further forecast future system states based on observed previous states and assumptions about future mechanisms. A typical example is that biologists usually study the infrapopulation dynamics of *Gyrodactylus* parasites on their host within a 17-day infection period; whereas, very little is known about the infection dynamics beyond these observation periods. Hence, with the help of a robust simulation model, predictions beyond the standard 17 days can be made using the fitted model. Of course, quantitative forecasts are still subject to some uncertainty following model validation; however, qualitative forecasts can be offered for several situations (only if past data are unavailable). Moreover, specific complex explorations that may be difficult to investigate under experimental settings (such as studying the infection dynamics of mixed-gyrodactylid strains on a single host) can be experimented with using the fitted model or modified model version (especially for agent-based models). Sections 3.3.2.1 and 3.3.2.2 present mathematical models for modelling microparasitic and macroparasitic infections, respectively. Figure 3.2 is a graphical summary of the modelling process of biological systems from model identification to testing hypotheses and making predictions based on the fitted mathematical model.

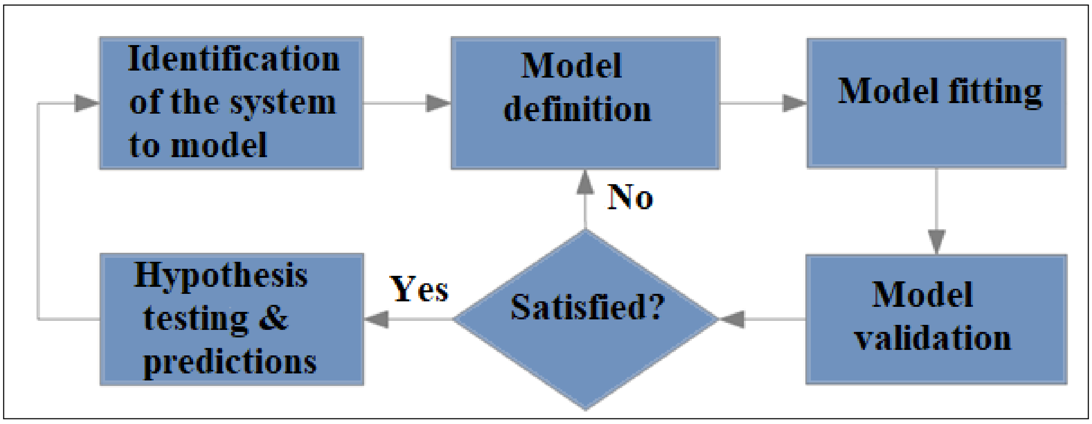
Identification of the system to model
Model definition
Model fitting
Model validation
Satisfied?
No
Yes
Hypothesis testing & predictions

Figure 3.2: Graphical summary of the modelling process of biological systems.

3.3.2.1 Modelling microparasitic infection

Microparasites are small-sized infectious agents that reproduce directly within the host and have a short generation time (e.g., viruses, pathogenic bacteria, protozoa, etc.) [289]. Depending on hosts' immunocompetence, they often gain immunity to infection over time, and infection duration is usually short compared to the host lifespan. Microparasitic infections can also be transmitted directly or indirectly through an intermediary host. Mathematical models of microparasitic infections are mostly population-based models which classify individuals in a closed population into three main categories or sub-populations: susceptible (where members of the population are at risk of infection), infected (where members of the population who are infectious after exposure), and recovered or immune (where members of the population are either dead or cannot be reinfected), and monitor temporal changes in the number of hosts within each group under the assumption that the entire population are homogeneously mixing. If the host survives the infection, it can either transition into the recovered class and become immune for a short time (or indefinitely in certain situations) or relapse into the susceptible class.

These types of models for host-microparasite dynamics, initially established by Kermack and McKendrick [171], are called compartmental models [8]. The basic fundamental microparasite model (given by equation 3.1 and outlined by Figure 3.3) under the density-dependent transmission assumption (where transmission or contact rate depends on the population density such as directly transmitted diseases), was developed by Anderson and May [7] with differential equations defined as:

$$
\begin{aligned}
\frac{dX}{dt} &= aN - bX - \beta XY + \gamma Y \\
\frac{dY}{dt} &= \beta XY - (\alpha + b + \gamma)Y,
\end{aligned}
\tag{3.1}
$$

where $\frac{dX}{dt}$ is the rate of change of susceptible host population $X(t)$, $\frac{dY}{dt}$ is the rate of change of infected host population $Y(t)$, and $N(t) = X(t) + Y(t)$ is the total host population at time t. Here, the per-capita birth rate of the host does not depend on the infection, and the net birth rate is given as aN (where a is the per-capita birth rate of the host). In addition, susceptible individuals are infected at rate β (where β represent the transmission coefficient) and die at rate b (where b is the per-capita death rate of the host); whereas, the infected host population die at a rate of $b + \alpha$, with α denoting the parasite-induced host mortality rate. Finally, if infected individuals survive the infection, they relapse into the susceptible state at rate γ (where γ is the recovery rate). Like other compartmental models, if the recovery rate (γ) is constant, then the distribution of infectious periods is exponential with mean $\frac{1}{\gamma}$. This assumption corresponds to the probability of recovery being independent of the time since infection in biological terms. In most applications, this is far from reality, but it simplifies the model formulation substantially. Otherwise, keeping track of when each infective acquired the infection is required.

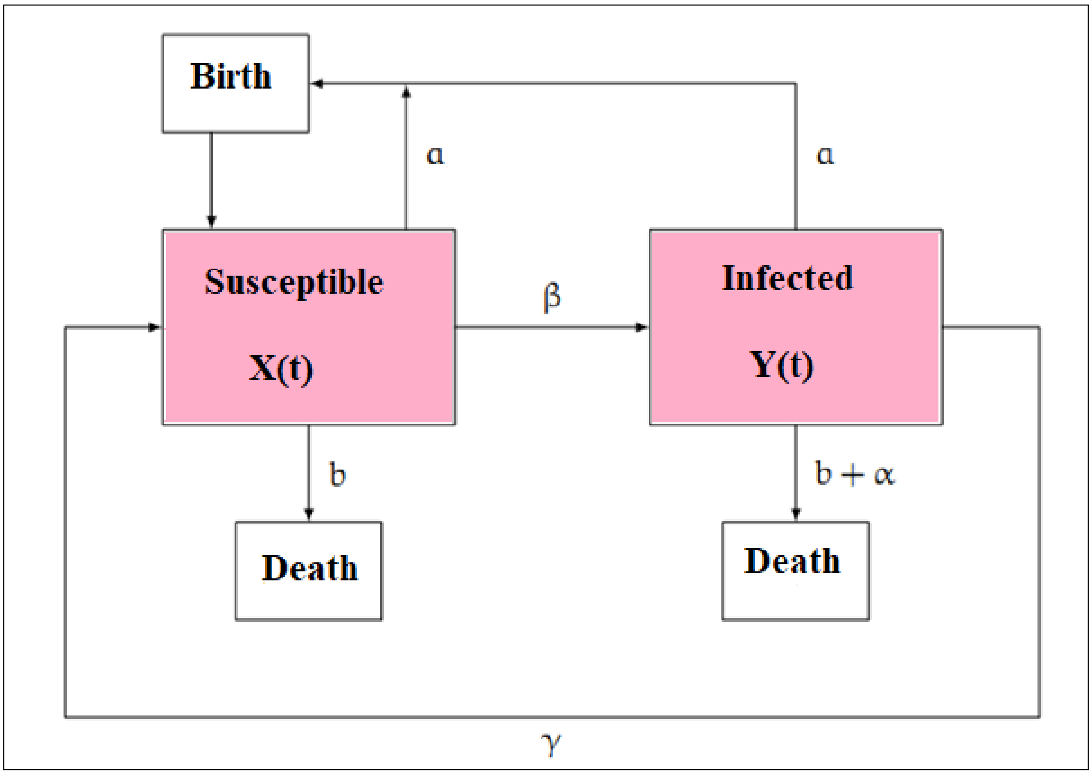

Figure 3.3: Conceptual framework of a simple host-microparasite interaction [adapted from 8].

This model was later modified and expanded to incorporate: i) parasite-induced host reproduction reduction, ii) infection latent periods, iii) vertical transmission, iv) disease and stress, v) density-dependent constraints, and vi) free-living infective stages [81]. Their simple model captured the essence of the dynamical interaction between invertebrate hosts and the microparasites they directly transmit by integrating parts of traditional epidemiology (where the host population is constant) and prey-predator dynamics (which conventionally emphasise how prey and predator populations may be regulated by their interaction) [7]. Transmission of some microparasites can also be relatively constant throughout a wide range of host populations, a mechanism known as a frequency-dependent transmission where transmission or contact rate is independent of the population density (e.g., vector-borne and sexually transmitted diseases) [294]. Due to frequency-dependent transmission in such instances, there is no threshold density for pathogen invasion; in theory, such pathogens can survive at arbitrarily low host densities. Figure 3.4 briefly outlines microparasitic infection dynamics within host populations from susceptible state to recovered or removed state. Nonetheless, the number of distinct groups can be increased or decreased depending on the complexity of the modelling problem and the disease infection under study. Thus, several variants of these compartmental models (with or without demography) and their diverse applications exist in the literature [see 48, 218, 262, 273]. Within the susceptible-infected (SI) framework, compartmental models can be as simple as the classical SI model or as quite sophisticated. For instance, Pandey et al. [233] described an epidemic model with 26 compartments. $MSEIR$, $MSEIRS$, $SEIR$, $SEIRS$, SIR, $SIRS$, SEI, $SEIS$, SI, and SIS are other typical compartmental models, where M refers to passive infant immunity, and E represent the exposed state but not yet infectious [146]. Compartmental models can be constructed deterministically using ordinary differential equations (ODEs) and difference equations or stochastically using continuous-time Markov chains and stochastic differential equations (SDEs). For example, the ODE epidemic model serves as a framework for developing equivalent stochastic models and a point of comparison with stochastic epidemic models [4]; thus, these two classes of models are different perspectives on the same infectious dynamics.

Disease infections are naturally stochastic, especially at the individual level, and stochastic models thus aid in understanding that random fluctuations can explain variations in disease transmission [28]. However, stochastic epidemic models are best adapted to studying infection dynamics in small populations, unlike their deterministic versions, and explaining infection dynamics at early stages. The prevailing view is that deterministic models should be thought of as approximations to stochastic models, and thus, the deterministic models are rightly infinite population limit of their stochastic models with homogeneous or non-homogeneous mixing populations [184, 261]. According to Kurtz [184], the approximation holds when all population sizes (i.e., the sizes of all subgroups defined in the model) are large, and the number of contacts made by infected individuals throughout their infectious phase is high; nonetheless, these assumptions are frequently inappropriate in practice. Appropriate adjustments to the basic SIR model extend the model's applicability by improving the consistency of the model assumptions with reality. Moreover, each extension necessitates additional information, such as community sizes or group contact patterns. As a result, it is vital to establish a careful balance between sophisticated and overly simplistic models. Random fluctuations can address minor features of stochastic models; thus, they can be developed with fewer details.

The desirability of deterministic models stems from the fact that they are mathematically more tractable than their stochastic counterparts [28]. However, many assumptions are made in compartmental models, including that hosts are uninfected at birth, that the disease does not affect host fecundity, that hosts can be immune to the disease over time, and that host populations are large such that stochastic processes can be ignored. Furthermore, because the growth of microparasites occurs quickly after exposure within the host, the intra-host infection dynamics can be disregarded when modelling with compartmental models or other population-based models. Infected hosts may die before microparasites can produce many new infections if they are excessively virulent, whereas non-virulent pathogens can become highly abundant but have low population-level effects. Therefore, simple microparasite models can provide essential insights for considering pathogen risks to wild or captive populations. For instance, models indicate that the effects of infectious microparasitic disease on the host populations are influenced by several factors, including pathogen effects on individual host fitness. In addition, models for microparasitic infections can produce important infection control predictions, such as the effects of vaccination or culling on the likelihood of pathogen eradication [8].

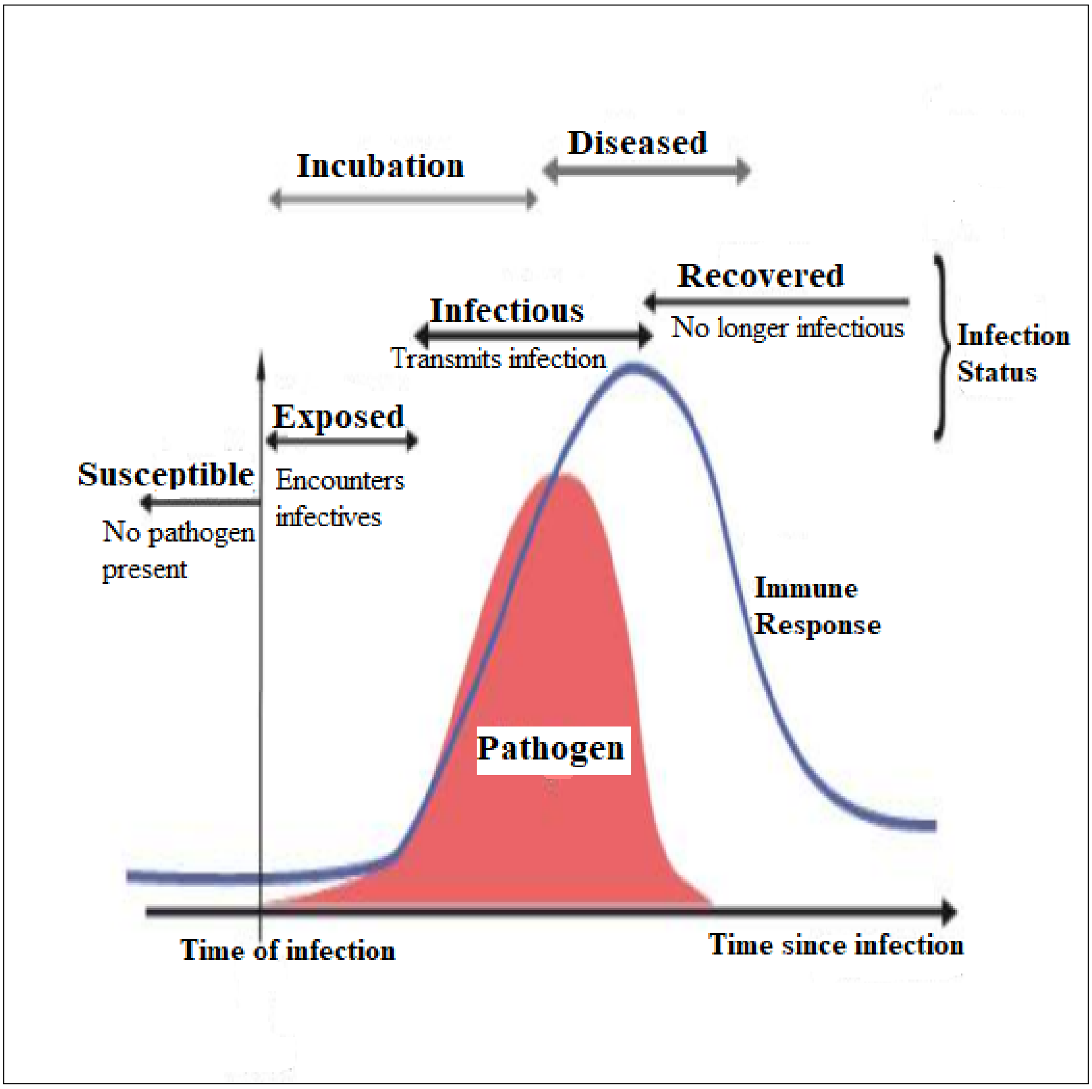

Figure 3.4: Infection dynamics of micro-parasitic infection in human and animal populations [adapted from 170].

3.3.2.2 Modelling macroparasitic infection

In contrast to microparasites, macroparasites have no direct reproduction within the host (e.g., parasitic arthropods, helminths, etc.) and are usually larger with more extended generations times [8]. For macroparasites, transmission stages (eggs and larvae) are produced and released into the environment or their host population. Extensive empirical studies have revealed that macroparasites are virtually always aggregated across host populations, with most individuals harbouring modest numbers of parasites but a few individuals hosting many [275]. According to experimental research, the magnitude of spatial aggregation in the infective stage distribution is reflected in the level of parasite aggregation across hosts [174]. Individual differences in hosts' exposure to parasite infective stages and disparities in their susceptibility once an infectious agent has been encountered produce variations like these. Furthermore, in the absence of any exposure heterogeneity, even minor differences in susceptibility between hosts can quickly establish non-random, aggregated parasite distributions [6]. The relative importance of these different mechanisms and the role of interactions between mechanisms in intensifying individual host differences in parasite burdens are still unknown [321].

Moreover, the number of macroparasites harboured by individual hosts often determines the severity of macroparasitic infections and the reproductive capacity of adult macroparasites (i.e., macroparasite virulence), and only a fewer number of the host population may survive high parasite abundance [275]. Modellers keep track of the number of adult macroparasites per host since macroparasitic infection outcomes (such as the survival and fecundity of macroparasites and their hosts) depend significantly on infection intensity. Mathematical models that investigate these problems mostly become intractable [117], while experimental studies and computer simulations can become increasingly complicated [321]. However, models for studying macroparasite infections must account for the variations in parasite population dynamics and aggregation in parasite abundance. In contrast to the compartmental models for studying microparasitic infection within host populations, distributional models are instead employed to model host-macroparasite interactions [8]. These distributional models are more sophisticated than compartmental models because they must account for parasite distribution among hosts [8]. Anderson and May [6, 212] proposed the core macroparasite model on which subsequent models are based; whereas, Dobson and Hudson and others [88] further modified the Anderson-May macroparasite model (outlined at the top part of Figure 3.5) to take into account the presence of free-living infective stages or larvae, arrested parasite development, and parasites with complex life cycles (which can incorporate several intermediate hosts as well as a definitive host). These models, which are a version of the predator-prey model, commonly characterise the density of the entire host population, the adult parasite abundance within hosts, and the total number of free-living parasite stages or larvae in the external environment. Interestingly, highly aggregated parasite distributions tend to maintain host–macroparasite interactions, whereas random or regular parasite distributions tend to destabilise them, resulting in host and parasite abundance population cycles (as shown at the bottom part of Figure 3.5).

Dobson and Hudson's modified macroparasite model is defined by the system of differential equations (given by equation 3.2) such that [88]:

$$\begin{aligned}\frac{dH}{dt} &= (a-b)H-(\alpha+\delta)P \\ \frac{dP}{dt} &= \beta WH-(\mu+b+\alpha)P-\alpha\frac{P^2}{H}\left(\frac{k+1}{k}\right) \\ \frac{dW}{dt} &= \lambda P-\gamma W-\beta WH,\end{aligned} \tag{3.2}$$

where $\frac{dH}{dt}$ is the rate of change of the total host population $H(t)$ (with grouse considered as hosts in their model), $\frac{dP}{dt}$ denotes the rate of change of the number of adult parasite population, and $\frac{dW}{dt}$ represents the rate of change of the total population of free-living larvae $W(t)$ at time t. As in the basic microparasite model (given by equation 3.1 and Figure 3.3), per-capita host birth and death rates in the modified macroparasite model are denoted by a and b, respectively. Here, the per-capita rates at which adult parasites induce host infecundity and mortality are respectively denoted by δ and α. The model assumes that reproduction takes place outside of the host via transmission stages (e.g., eggs or larvae), and the overall host death rate increases linearly with parasite burden. In addition, older parasites reproduce free-living infective stages (or larvae) at rate λ and die due to three different processes: parasite mortality (μ), host mortality (b), and parasite-induced host mortality (α). As a result, the model assumes that once hosts die, their parasites extinct. Free-living egg and larval stages die in the external environment at a rate γ, while hosts (e.g., grouse) consume them at a rate of β, resulting in new adult infections. Furthermore, the model described by Figure 3.5 further assumes that parasites aggregate within hosts according to a negative binomial distribution, with the degree of aggregation inversely proportional to k. Nonetheless, for more complex host-macroparasite systems, the exact probability distribution of the number of aggregated parasites per host may be unknown or mathematically intractable. The mortality of adult parasites is influenced by within-host aggregation (as indicated in equation 3.2), with parasite mortality increasing when k is small (and parasites are highly clustered). Dobson and Hudson [88] defined the basic reproductive number of macroparasites (R_0) as the product of the average number of new infections caused by a single adult parasite and the average life expectancy of adult and larval stages; where

$$R_0 = \frac{\beta\lambda H}{(\mu+b+\alpha)(\gamma+\beta H)}. \tag{3.3}$$

As with microparasitic infections, macroparasite invasion and persistence within host populations occur when $R_0 > 1$. When parasites decrease host fecundity (i.e., $\delta > 0$), the host–macroparasite interaction becomes even more destabilised, increasing the likelihood of parasite-induced host population cycles over time (for additional quantities and applications from this macroparasite model, see [88]). Later versions of Anderson and May's basic model included: i) non-random parasite distributions, ii) non-linear parasite-induced host deaths, iii) density dependence in parasite population growth, iv) parasite-induced reduction in host reproduction, v) parasites that reproduce within their hosts, and vi) the effect of time delays [6, 212]. The existing mathematical models for host-macroparasite systems can be extended, modified and adapted for other host populations (including hosts that do not feed directly on their parasites but can induce host immune response over time).

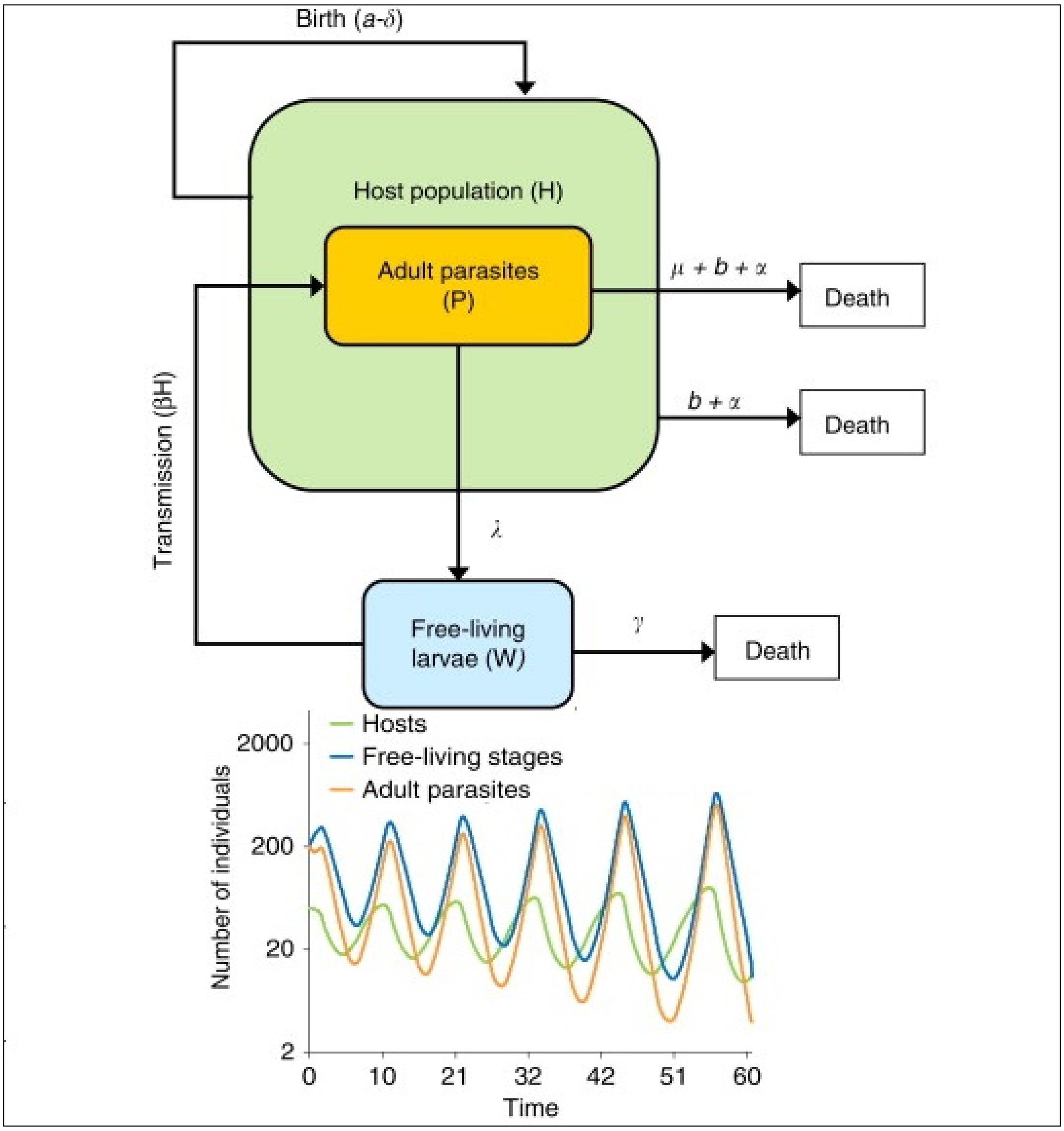

Figure 3.5: Conceptual framework of a simple host-macroparasite interaction model (top), and the infection dynamics of host population, adult parasites and free-living transmission stages over time (bottom) [adapted from 88].

3.3.3 Overview of population-based models

Population-based models (PBMs) dominate epidemiological and ecological studies compared to individual-based models (IBMs), and they frequently produce basic models, such as systems of ordinary differential equations or difference equations, that can be analysed mathematically and numerically [197]. In population ecology, the spatial-temporal variations in abundance and distribution of species (including plants, animals or humans) are often explored; whereas, changes in the number of individuals of a given species or interacting species over time are usually described using dynamical system modelling (based on differential or difference equations). PBMs thus have been proposed to predict and understand these population dynamics. To get the basic knowledge of species' spatial-temporal abundance and distribution using PBMs, four main parameters which are not limited to the number of births (B), deaths (D), immigration (I) and emigration (E), are often tracked; such that the changes in the population size (N) through space and time is given as [46]

$$N(t) = B(t) - D(t) + I(t) - E(t). \tag{3.4}$$

Equation 3.4 suggests that the population increases in number through births or immigration and decrease through death or emigration at any time t; nonetheless, there can be other sources of population increments or decrements depending on the dynamical ecological system under study. Broadly, continuous-time models, discrete-time models, and stochastic models are the three basic types of PBMs applied to problems in population ecological modelling [46]. Population-based mathematical models can be approached from two perspectives. Firstly, as deterministic population models (i.e., using difference or differential equations), such as the continuous-time Lotka-Volterra model of interspecific competition or the discrete-time Nicholson-Bailey predator-prey model [178, 202], or secondly, as stochastic population models, in which the occurrence of events is considered probabilistic even though the underlying rates remain constant [214, 230]. Even though these PBMs often ignore some underlying biological or ecological realism of the system due to mathematical generalities, they provide a framework through which the occurrence of fundamental or complex relationships and processes of the system can be formulated and explored [138].

In mathematical biology, integrodifference equations are widely adopted to model the dispersal and growth of populations [201]. An alternative approach to the previously discussed deterministic and stochastic epidemic population-based models (including deterministic and stochastic integrodifference equations and compartmental models) in theoretical ecology and epidemiology is branching process models (BPMs), which are classes of either discrete-time or continuous-time stochastic individual-based processes often employed to approximate the beginning of an epidemic[3, 160, 221]. The Galton-Watson discrete-time process (described in [173]) is the most common formulation of the branching process; nonetheless, its continuous-time processes can be derived via embeddability of the discrete-time branching processes. Specifically, BPMs are suitable for modelling problems across many species of animals, plants and other organisms where the growth and development of the population over time are of interest [13]. In the simplest case, the original intent of branching processes was to formulate a mathematical model in which each individual of a given population in the nth generation produces a random number of offspring in generation $n+1$, under fixed probability distribution (obtained from a probability generating function) [221]. Multi-type branching processes, for example, are more complex extensions of that simple branching process [121, 162]; whereas, other general forms of the branching process formulated as an embedded random walk exist in the literature [35, 246]

To formulate the simple discrete-time case, suppose Z_n is the state in period n (i.e., the population size of generation n), and assume $X_{n,k}$ is an independent and identically distributed random variable representing the number of offspring (k) in the nth generation, where $n \in \{0,1,2,\cdots\}$ and $k \in \{0,1,2,\cdots,Z_n\}$. Also, suppose $p_k = P(Z_1 = k)$ is the probability mass function of $Z_1 = X$ (where p_k denote the probability that an individual produces k offspring in one generation), with corresponding probability generating function given as $f(s) = \sum\limits_{k=0}^{\infty} p_k s^k$ (for s on the unit interval). Then the recurrence equation for the state at generation $n+1$ is given as [312]

$$Z_{n+1} = \sum_{k=0}^{Z_n} X_{n,k}, \quad Z_0 = 1. \tag{3.5}$$

In epidemiological studies and the modelling of other systems with similar dynamics, BPMs are often used to determine: i) the long-term survival of such a process or its ultimate extinction probability such that $\lim\limits_{n\to\infty} P(Z_n = 0) = f(0)$, ii) the epidemic threshold known as the basic reproductive number (R_0), iii) the intrinsic exponential growth rate of the process (e.g., the Malthusian parameter for epidemic branching process), iv) expected population size of a particular generation, and v) the effects of control mechanisms. For example, in branching process theory, the mean reproductive rate of the branching process μ (which also estimates R_0) is

$$\mu = E[X] = \sum_{k=0}^{\infty} kp_k = f'(1) \tag{3.6}$$

and by employing the standard Wald's equation of expectation to equation 3.5, the expected population size of the nth generation is

$$E[Z_n] = \xi\mu^n, \quad \text{for} \quad Z_0 = \xi. \tag{3.7}$$

By imposing the Markovian property of the branching process (where the state Z_n depends on Z_{n-1}) and equation 3.6, leads to the efficient Harris estimator or Method of Moments estimator of μ based on the first N generations is defined as [103, 326]:

$$\hat{\mu} = \lim_{N\to\infty} \frac{\sum\limits_{n=1}^{N} Z_n}{\sum\limits_{n=0}^{N-1} Z_n} \approx R_0. \tag{3.8}$$

From equations 3.6–3.8, the expected number of individuals in the population extinct if $\mu < 1$ with probability of ultimate extinction occurring at $\mu = 1$; whereas at $\mu > 1$, the probability of ultimate extinction < 1 (but not necessarily zero). Unlike in similar deterministic models (e.g., epidemic compartmental models where $R_0 > 1$), there remains a chance of extinction even when the mean reproductive rate $\mu > 1$ [254]. A more comprehensive theoretical framework and justifications of mathematical models for branching processes, including multidimensional analogue of the Galton-Watson model and continuous-time branching processes, have been presented in Harris's study [135]. Finally, these BPMs closely approximate the IBMs and provide insight into the demographic and environmental stochasticity parameters that influence the critical domain size for a stochastic population (necessary for the population to persist; but the critical domain size is dependent on the model structure and assumptions) [254]. For vast applications of branching processes in epidemiological and ecological modelling, see works by Vajargah and Moradi [302], Jacob [160], Inés et al. [155] and Hautphenne et al. [142]. For a variety of problems, branching processes can be simulated, particularly in the field of evolutionary biology, and such simulations aid in the development and validation of these models' estimating methodologies, as well as hypothesis testing [122].

Moreover, continuous-time population models, characterised by differential equations that dominate the literature, are: i) Pure birth processes (suited to modelling systems where over a short period, the population proliferates without being constrained by crowding, competition, or contests and do not die but give birth at a constant rate), ii) Pure death processes (suited to ecological processes concerned with how an organism's lifespan and survival affect population fluctuations over time, where individuals do not give birth and do not suffer the constraints of crowding, competition or contests but die at a constant rate), iii) Birth-death processes (suited to systems where populations change as a result of birth and death processes), iv) Logistic growth models (describe population growth in the presence of a limiting resource or carrying capacity where co-existing species or individuals are embedded in webs of competitive and trophic interactions), and v) Predator-prey processes (which are a class of population-based models used to explore the effects of interspecific and intraspecific competition among species or populations) [46]. Different species or populations can compete for limited resources like food and territory. These interspecific competitive (between species) interactions have a well-established mathematical framework [150], and theoretical results for a broad class of models involving predation, competition, and cooperation exist [255]. Although correlation does not necessarily indicate causation (e.g., a negative correlation between populations of two organisms does not always imply interspecific competition), basic mathematical models of competition assume that a species' growth rate is inhibited by either intraspecific (within species) or interspecific processes.

The single-species logistic growth model (such as the standard Pearl-Verhulst logistic growth model) can be extended to a two-species form assuming $N_1(t)$ and $N_2(t)$ denote the numbers of individuals of species 1 and 2 at time t, respectively. Suppose r_i are the growth rate of species i, s_{ii} is the intraspecific effect of species i on itself, and s_{ij} is the interspecific effect of species j on species i for $i,j = 1,2$. Biologically, for species $i,j = 1,2$, it is assumed that s_{ii} or $s_{ij} < 0$ implies an inhibitory effect, s_{ii} or $s_{ij} = 0$ suggests no effect and s_{ii} or $s_{ij} > 0$ indicates an enhanced effect. Then, the general model form (under the deterministic scheme) for two-interacting populations with direct competition within and between species is given as [255]

$$\begin{aligned} \frac{dN_1}{dt} &= N_1(r_1 + s_{11}N_1 + s_{12}N_2) \\ \frac{dN_2}{dt} &= N_2(r_2 + s_{21}N_1 + s_{22}N_2), \end{aligned} \tag{3.9}$$

where $r_i > 0$ with $s_{ii} < 0$ and $s_{ij} < 0$. However, it can be inferred from equation 3.9 that for non-interacting populations, $r_i > 0$ with $s_{ii} < 0$ and $s_{ij} = 0$ for $i,j = 1,2$ (i.e., independent logistic case). Also, the second species lives on the waste products of the first (i.e., scavenging), but otherwise does it neither harm nor good if $r_i < 0$ with $s_{ii} < 0$, $s_{12} = 0$ and $s_{21} > 0$. Additionally, there is a symbiotic relationship between the two species or populations if $r_i < 0$ with $s_{ii} = 0$ and $s_{ij} > 0$ for $i,j = 1,2$. Since the overall dynamics of the general two-species model given by equation 3.9 is dependent of six real parameters (which may be < 0, $= 0$ or > 0), there are $3^6 = 729$ possible model formulations; nevertheless, only a few which involves either predation or competition may have biological importance. The traditional Lotka-Volterra predator-prey model (under the deterministic scheme) can be derived from equation 3.9 by setting $r_1 > 0$, $r_2 < 0$, $s_{11} = s_{22} = 0$, $s_{12} < 0$ and $s_{21} > 0$; where, $N_1(t)$ is the number of prey (or hosts), $N_2(t)$ is the number of predators (or parasites) at time t. Here, it is assumed that in the absence of predators (species 2), prey (species 1) increases at a rate of r_1; whereas, predators die at a rate of r_2. Also, within-species competition is ignored at $s_{11} = s_{22} = 0$; whereas, s_{12} measures the death rate of prey due to being eaten or attacked by predators, and s_{21} determines the ability of the predator in catching prey. The more predators there are, the faster the prey population will be reduced, and the availability of predator food resources increases as the number of prey increases.

Biological experiments or mathematical modelling are the conventional techniques to study parasites' influence on communities; the former is logistically problematic in the field. The latter relies on various simplifying assumptions. Both can only examine interactions among a few species at a time. Network analysis, adapted from other mathematical domains, is gaining attention in community ecology and is now applied to study complex host-parasite interactions [242]. It enables an entire complicated system to be analysed rather than just one or a few components at a time. The application of network theory to understand and predict the spread of parasitic infections through host populations via social or sexual interactions has been successful [92, 238]. Network analysis can also be used to uncover recurrent coevolutionary units within a more extensive host-parasite system and determine how a community will react to perturbations like introducing new species through migration, invasion or the removal of species following local extinction. These models also provide a fruitful alternative framework in which to investigate the transmission of infection in human and animal populations [169], and network epidemic models under random and non-random mixing assumptions have been proposed in the literature [327]. A major reason for studying epidemic models on social networks is to understand better which network features have the most significant impact on spreading and, in particular, how public health measures such as vaccination, (quicker) diagnosis and treatment, isolation, travel restrictions, and so on can be used to reduce spreading [49].

However, modelling transmission over networks is mathematically and computationally challenging due to the intrinsic high-dimensionality of networks. As a result, even the most basic network epidemic models leave many problems unexplained. Efforts to increase the practical utility of network models by integrating realistic elements of contact networks and host-parasite biology (for instance, waning immunity) have made some headway, but robust analytical results are still limited. A more general theory is required to comprehend the impact of network structure on infection dynamics and control. The effect of network models for heterogeneously mixing populations on parameter estimation and the epidemic outcome is under-studied [237]. Although these models allow for examining the effect of clustering and, in some cases, degree correlation on epidemic features, it must be acknowledged that the networks they generate are relatively unique and difficult to generalise. Furthermore, epidemics based on different network models with the same degree distribution, clustering coefficient, and degree correlation may have different properties [19]. Understanding how network properties affect epidemiological quantities of interest is a commonly stated difficulty for sophisticated network models. Several studies have improved upon methodologies for social network analysis. For an extensive discussion of the computational statistical methods and models (including exponential-family random graph models, dynamic Markovian and non-Markovian models of networks, joint model of networks, measurement error models and partially sampled networks, etc.) and parameter estimation methods in social network analysis, see works by Carrington et al. [59] and Hunter et al. [152].

3.3.4 Overview of agent-based models

Agent-based models (ABMs), also known as (spatially explicit) IBMs in ecology, are a population and community modelling approach that allows for a high level of individual and interaction complexity [119]. This class of models can be seen as a natural extension of the Lenz-Ising model [225] and Cellular Automata-like models [324]. IBM of population dynamics is a popular approach in current theoretical ecology [30], although it has only been used in a few parasitological research so far [see 107, 199]. IBMs are typically more complicated than PBMs and are better suited to simulation experiments or *in silico* modelling than statistical analysis [197]. Specifically, ABMs can simulate either single-species or multiple-species populations of biological systems with both motile and sessile individuals or organisms [30]. Each autonomous individual or agent (e.g., parasite, host or otherwise) may have a unique set of state variables or traits (e.g., spatial location and physiological features) and peculiar behavioural attributes in terms of growth, reproduction, habitat selection, foraging, and dispersal [78]. These characteristics may differ between individuals and might change over time. Nonetheless, agent-based models can also describe changes in the number of individuals rather than population density and explicitly account for resource dynamics. Moreover, the advancement of computers shifted the focus of ecological modelling to individuals rather than population averages used in population-based models (e.g., compartmental modelling approach) [301]. Additionally, individual-based models are bottom-up models in which population-level behaviours emerge from interactions among autonomous individuals and their abiotic environment, as opposed to traditional differential equation population models, which are described in terms of imposed top-down population parameters (such as birth and death rates) [78].

Models of spatially explicit population dynamics (such as IBMs, interacting particle systems, neighbourhood models and spatial point processes) can be classified based on whether population sizes, space, and time are characterised as discrete or continuous entities [43, 76, 90]. Individuals' fates in the model are characterised by sets of rules or assumptions (dependent on age, size, sex, genotype, etc.) that determine their performance. Individual-based models have the benefit over traditional models in that they can include any number of individual-level mechanisms. Therefore, agent-based models are utilised any time one or more of the following elements are regarded vital for addressing a research question or solving an applied problem, but are difficult or impossible to represent in population-level differential equations [30]: i) individual and temporal variations, ii) local interactions between individuals, and iii) adaptive behaviour (which includes physiological features and energy budgets). However, because each individual is treated as a separate entity in IBMs, simulation of the underlying system can be challenging for complex systems: they can lead to high computational costs (both in terms of time and memory). Nonetheless, as computers get more potent through parallel computing and high memory sizes, this becomes less of an issue when tracking each individual's states across the entire population.

The usage of spatially explicit IBM simulation for biological systems necessitates a description of the environment (e.g., host) and each individual (parasite) living in it as well as interactions between individuals and the environment. At (almost) every step of this description of agent-based models, at least a few alternatives are a priori plausible, including the model framework, environment specification for individuals to coexist, number of model parameters of interest, model complexity, and simulation method or approach [30]. Knowing whether and how a choice from among the alternatives changes spatio-temporal patterns should help distinguish the effects of model formulation and biological processes; nonetheless, several modellers have argued that IBMs are quite robust to at least some of these aforementioned alternatives [91, 323]. Thus, to develop agent-based models, good knowledge of the biological system simulated is required. For example, to specify a schematic representation of the environment (i.e., host) of an IBM, a question usually asked involves whether a (discrete) lattice should be modelled as regular or irregular, and if so, whether it should be made up of squares, triangles, or hexagons in the former case. In addition, the physical qualities of the environment can be assumed to be homogeneous or heterogeneous; movement can be directed rather than diffusive, the initial population distribution might be random or have a particular spatial pattern, and so on [30].

Apart from making a substantial contribution to spatial pattern formation, IBMs also give a means of determining the population-level outcomes of specific individual-level behaviour [30]. Computer simulations must be run repeatedly to offer information about average or typical population responses for IBMs to be reliable. These simulation experiments are better suited to answer specific questions about the model or the biological system rather than uncovering whole model behaviour due to the high dimensionality of model parameter space in some instances. Mean-field models are analytical models (represented by differential or difference equations) that do not consider spatial or probable variation (but of great importance) and thus, are in sharp contrast to the individually oriented modelling spectrum. The mean-field modelling approach has at least one advantage: the agreement between mean-field models and IBMs simulated under homogeneous mixing settings provides a valuable starting point for exploring the complexities caused by assumptions that destroy these conditions. Thus, mean-field model techniques described extensively by Berec [30] can be adapted in the construction of spatio-temporal model frameworks like IBMs.

Individual-based models have been criticised for lacking the formal framework and analytical procedures accessible in mathematical models made up of differential equations and Markov chain models [119]. This is partly due to individual heterogeneity or complex interaction structures that can cause impacts on system dynamics that are difficult to account for using population-based framework [20]. To address this problem, understanding the transition from the most informative individual level to the levels at which system behaviour is typically observed is vital. Thus, a Markov chain approach can help derive and evaluate models on specific levels on the one hand and understand the temporal and spatial patterns that may emerge in that transition on the other [20]. A rigorous investigation of a family of agent-based models that specify the dynamics of a complex system at the individual level using the Markov chain approach has been proposed by Banisch [20]. It uses lumpability and information theory to link the individual and population levels of observation, providing a basic framework for aggregation in agent-based and related computational models.

The starting point is a microscopic Markov chain description of the dynamical process that is completely consistent with the dynamical behaviour of the ABM, which is derived by treating the state space of a large Markov chain as the set of all possible agent configurations [20]. This is known as a micro chain, and using the random mapping representation of a Markov process (defined in [164]), an explicit formal representation incorporating microscopic transition rates may be obtained for a class of models. The circumstances where the macro model is still Markov may be identified using well-known lumpability constraints, and in this case, a complete picture of the dynamics is obtained, including the transient stage, which is the most informative phase in applications. The sort of probability distribution used to construct the stochastic element of the model, which determines the updating process and drives the dynamics, plays a critical role in this regard. The problem of aggregation in ABMs, particularly the lumpability constraints, can be incorporated into a broader framework that employs information theory [defined in 274] to identify different levels and relevant scales in complex dynamical systems.

Izquierdo et al. [157] introduced the possibilities of using time-homogeneous Markov chains in the investigation of ABMs based on computer models (with a well-defined mathematical function written in a programming language, where pseudo-random number generators are used to simulate random variables in the computer models). They argued that when a computer model is analysed as a time-homogeneous Markov chain, many model features not explicit prior to the analysis become apparent. The key concept is to incorporate all possible agent system configurations as the state space of a huge Markov chain. While Izquierdo et al. (2009) relied on numerical computations to estimate the stochastic transition matrices of the models, Banisch [20] showed how to derive the transition probabilities $\hat{P}$ explicitly in terms of the update function u and a probability distribution ω accounting for the stochastic parts of the model. Thus, realisations of ABMs with a sequential update strategy can be thought of as random walks on regular graphs. Consider a simple agent-based system defined by a set of N of agents (e.g., hosts or otherwise), where each one is characterised by individual attributes (e.g., physiological features, spatial location and behaviour) from a finite list of possibilities (denoted by S). Suppose $\Sigma = S^N$ is the configuration space representing the set all possible combinations of attributes of agents, and let $\mathbf{x} = (x_1, x_2, \cdots, x_N)$ with $x_i \in S$ for $i = 1, 2, \cdots, N$ denote an agent configuration (where $\mathbf{x}$ is a vector of discrete numbers). At each time step of the time-homogeneous Markov simulation model, the agents' attributes' updating procedure typically consists of two phases. First, a subset of agents is chosen at random using some probability distribution ω, and then, the attributes of the agents are updated according to a rule defined by u, which depends on the subset of agents selected at this time. ABMs can be represented using this specification by the so-called random map representation of Markov chains, under the existence and regularity conditions [188]. A more detailed mathematical theory and applications of Markov chain aggregation for agent-based models are presented by Banisch [20], and Izquierdo et al. [157] has demonstrated an in-depth approach of computer or agent-based modelling using time-homogeneous Markov chains. A wide range of individual-based models and techniques, as well as their applications in ecological modelling, have been explored by DeAngelis [77].

3.4 Modelling of *Gyrodactylus* infection dynamics

Gyrodactylus parasites are unique in that they are ectoparasites with microparasitic characteristics (as discussed in section 1.4). Several modelling studies have been conducted over the years to understand the infection dynamics of these parasites on their fish host (especially for *G. salaris* infection). Scott and Anderson [271] conducted a study in 1984 to understand *Gyrodactylus turnbulli* infection dynamics, which examined using SIR models with parameter values based on experimental data of guppy population. Their goal was to establish which factors have a direct impact on parasite transmission dynamics. des Clers [82] also designed age-structured population models to evaluate the effects of *G. salaris* on various stages of the salmon life cycle in 1993. *G. salaris* on salmon has also been studied using other methodologies such as Monte Carlo models [148, 232]. Paisley et al. [232] utilised this modelling technique to assess the risk of *G. salaris* being introduced to the Tana river in Norway, while Høgåsen and Brun [148] adopted the same technique to estimate inter-river transmission risk of *G. salaris* by migrating Atlantic salmon smolts. In addition, other studies have used qualitative risk assessment and analysis techniques to detect routes of transmission and the risk of *G. salaris* being introduced into the UK, as well as the risk of *G. salaris* spreading to uninfected areas of Europe [234, 235, 236].

Jansen et al. [161] employed a dispersal model to investigate the possibility of secondary infections by testing the hypothesis of parasite inter-river dispersal; whereas, van Oosterhout et al. [306] developed an individual-based computer model to simulate the dynamics of gyrodactylid parasite infection and naive hosts' immunological defence (i.e., among fish that have never been exposed to these parasites). Their computer model can predict the progression of gyrodactylid infections in a single host and provide predictions regarding parasites' optimal life history; however, it suffers from a few biological realism of the gyrodactylid-fish system (especially in terms of species-specific microhabitat preference, parasite fecundity per age and host mortality, amongst others). Another agent-based simulation model has been developed to quantify estimation error of *G. salaris* population growth rate on a single salmon host conditioned on stochastic variability in survivorship and reproduction [248]. Their agent-based simulation model assumed two distinct death functions: i) constant parasite death throughout the simulation and ii) parasite death is positively associated with parasite age (chance of death increases with parasite age) as in a study by Cable et al. [54]. Their findings revealed that estimations of error structures of population growth rate are normally distributed, especially in populations of more than 20 parasites, and that this rate can be an important measure for comparing gyrodactylid populations of more than 20 to 30 parasites. Nevertheless, in populations with fewer than 20 parasites, the error is disproportionately high, making comparisons of gyrodactylid population growth on different hosts using the population growth parameter less relevant [248]. Furthermore, Ramrez et al. [248] discovered that decreasing parasite population growth rates could not be explained based on stochastic error, implying that the cause is biological. Finally, they concluded that the bulk of gyrodactylid-host studies identical to their study are a few to identify significant changes in local adaptation of gyrodactylid monogeneans amongst fish populations.

As can be seen, most of the previous modelling works on the gyrodactylid-fish system mainly focused on *G. salaris* infection dynamics among Atlantic salmon populations, with the majority of the research focusing on the use of statistical and computer models to assess and estimate the risk of the parasite's spread to new rivers [148, 161, 232, 234, 235, 236]; whereas a few modelling studies on host-parasite dynamics have been carried out on other parasite species (e.g., *G. turnbulli* and *G. bullatarudis* strains) and other host populations [257, 271, 306]. Moreover, the Anderson-May deterministic models (previously discussed in section 3.3.2) were adapted and extended to analyse the *G. salaris*-Atlantic salmon infection system further in another study [81]. It allowed predictions on the long-term consequences or impact of infections in UK regions free of *G. salaris*. As a result, models of host-parasite interactions in the other gyrodactylid-fish systems need to be investigated since much is already known about the *G. salaris*-salmon systems. In addition to comprehending the short-term infection dynamics of gyrodactylids as often done in the previous modelling studies, it is critical to understand the long-term effects of gyrodactylid parasite infections.

Chapter 4

Birth-death process with catastrophic extinction

4.1 Introduction

This study investigates a continuous-time Markov process dubbed as the linear birth-death process with catastrophic extinction (B-D-C process). The primary motivation is to consider the B-D-C process as an auxiliary model for a more complex stochastic model that simulates the spread of different strains of *Gyrodactylus* parasites over the external surfaces of the host (see Chapters 5 and 6). Here, the B-D-C process is used to refine the summary statistics of a modified approximate Bayesian computation (ABC) in calibrating the multidimensional stochastic model based on estimates of the B-D-C model parameters (see Chapters 5 and 6). The simulation of the B-D-C process using a tau-leaping algorithm also provides additional insights on how to accelerate the simulation of the sophisticated stochastic model by proposing a good error threshold based on the trade-off between simulation accuracy and computational speed.

The constant-rate linear B-D-C process is a discrete state-space stochastic process where each host gives birth to new hosts at a constant rate $\lambda > 0$, dies at constant rate $\mu > 0$ and the entire population becomes extinct due to a catastrophic event at a constant $\rho > 0$ (for a formal definition, see section 4.1.1). Thus, the linear B-D-C process is an extension of the classical linear birth-death process where the process is subjected to catastrophes that result in parasite population extinction [83]. Due to the application of the B-D-C process in the current study (for other modelling purposes), the catastrophe rate is assumed to depend on the parasite population size since host mortality (defined as the catastrophe event) for ectoparasitic infection occurs at a rate proportional to parasite abundance. Although these class of stochastic models are simple in terms of their model framework, the exact transition function and parameter estimation can be challenging to obtain in the setting of discretely observed processes [75, 167].

In this current thesis chapter, we derive the analytical transition function of the B-D-C process (section 4.1.2). The derived transition function is further validated analytically using mathematical induction and is numerically validated based on Monte Carlo estimation (sections 4.1.2–4.1.4). Additionally, we estimate the B-D-C model parameters by comparing different estimation methods (Maximum likelihood estimation, generalised method of moments and embedded Galton-Watson approach) based on three different *in silico* simulation experiments where parasite population size is large, moderate or low (section 4.2). The bias, variance, and mean square error of the parameter estimates and the estimation methods' computational times are compared. Finally, we develop and compare two different hybrid τ-leaping algorithm based on leap-size selection methods proposed by Gillespie [110] and Gillespie and Petzold [111], respectively, to accelerate the simulation of the B-D-C process (section 4.3). We propose a good error threshold by exploring the trade-off between simulation accuracy and computational speed of the three different *in silico* simulation experiments where parasite numbers are high (Case 1), moderate (Case 2) or low (Case 3). The differences between the two τ-leaping methods are the leap-size selection procedure (which is proportional to the simulation error bound) and their respective leap conditions. All the mathematical theorems under Chapter 4 are proposed and proved for the first time in the current study.

4.1.1 Definition of the Linear B-D-C process

Let $\{X_t, t \geq 0\}$, the number of parasites on host at any time t, be a linear birth and death process with catastrophic extinction (B-D-C process) defined on the state space, $S = \{0, 1, 2, \cdots\}$, determined in accordance with the following scheme:

Event	**Transition**	**Rate**
Birth	$X_t \to X_t + 1$	λX_t
Death	$X_t \to X_t - 1$	μX_t
Catastrophe	$X_t \to 0$	ρX_t

where $\lambda > 0$, $\mu > 0$ and $\rho > 0$ are the birth, death and catastrophe rates respectively. The catastrophe event is defined as the state where the parasite population suddenly hit 0 due to host mortality. The exact transition probabilities of the defined B-D-C process,

$$P_{m,n}(t) = P\{X(t) = n|X(0) = m\}$$

can be obtained from the probability generating function $G_m(z,t)$ as presented by Lemma 1. The probability generating function (PGF) given by Lemma 1 is taken from Karlin and Tavaré [167].

Lemma 1. Given the rates λ, μ and ρ, suppose v_0 and v_1 are the roots of the equation

$$\lambda v + (\mu/v) = \lambda + \mu + \rho; \quad 0 < v_0 < 1 < v_1.$$

Then, the probability generating function of the B-D-C process, given $X(0) = m \geq 1$, is defined as:

$$\begin{aligned} G_m(z,t) = \sum_{n=0}^{\infty} P_{m,n}(t) z^n &= \left[\frac{\nu_0\nu_1(1-\sigma) + z(\nu_1\sigma - \nu_0)}{\nu_1 - \sigma\nu_0 - z(1-\sigma)}\right]^m + C(t) \\ &= \left(\frac{k_1 + k_2 z}{1 - k_3 z}\right)^m + C(t) \quad \text{for} \quad |z| < 1; \end{aligned} \tag{4.1}$$

where $k_1 = \frac{\nu_0\nu_1(1-\sigma)}{\nu_1-\sigma\nu_0}$, $k_2 = \frac{\nu_1\sigma-\nu_0}{\nu_1-\sigma\nu_0}$, $k_3 = \frac{1-\sigma}{\nu_1-\sigma\nu_0}$, $\sigma = \mathrm{e}^{-\lambda(\nu_1-\nu_0)t}$, $\nu_0 = \frac{(\lambda+\mu+\rho)-\sqrt{(\lambda+\mu+\rho)^2-4\mu\lambda}}{2\lambda}$, $\nu_1 = \frac{(\lambda+\mu+\rho)+\sqrt{(\lambda+\mu+\rho)^2-4\mu\lambda}}{2\lambda}$, and the probability of catastrophic extinction, $C(t)$, is given as

$$C(t) = 1 - \left(\frac{k_1 + k_2}{1 - k_3}\right)^m.$$

Remark. If the catastrophe rate $\rho = 0$, then the linear B-D-C process $\{X_t, t \geq 0\}$ (with birth rate $\lambda > 0$, death rate $\mu > 0$ and catastrophe rate $\rho = 0$) is the standard linear birth-death process with its probability generating function given as

$$\tilde{G}_m(z,t) = \left[\frac{(1-\sigma) + (\sigma - \gamma)z}{1 - \sigma\gamma - z\gamma(1-\sigma)}\right]^m, \quad m \geq 1, \quad |z| < 1$$

where $\sigma = \mathrm{e}^{-(\mu-\lambda)t}$ and $\gamma = \lambda/\mu$. Also, for catastrophe rate $\rho > 0$, the linear B-D-C process conditioned on non-extinction is a birth-death process (where $C(t) = 0$).

4.1.2 Derivation of the transition function and theoretical moments of B-D-C process from its PGF

We propose and prove the analytical form of the *nth* derivative of the PGF given by equation 4.1 (w.r.t. z) of the B-D-C process (in accordance to Theorem 2, proposed for the first time in the current study) using mathematical induction. The exact transition probability function and other theoretical moments (mean, variance and the 3rd uncentred moment) are explicitly derived for further modelling purposes.

Theorem 2. *Given the probability generating function $G_m(z,t)$ defined in Lemma 1, the nth derivative w.r.t. z is given as*

$$G_m^{(n)}(z,t) = \sum_{j=1}^{\min(m,n)} \gamma_j^{(n)} \frac{m!}{(m-j)!} k_3^{n-j} (k_2 + k_1 k_3)^j \left(\frac{k_1 + k_2 z}{1 - k_3 z} \right)^{m-j} (1 - k_3 z)^{-(n+j)}, \quad m,n \geq 1 \tag{4.2}$$

where $\gamma_j^{(n)}$ is defined recursively by

$$\gamma_j^{(n)} = \gamma_{j-1}^{(n-1)} + (n+j-1)\gamma_j^{(n-1)} \quad \text{for} \quad j = 2,3,\cdots,\min(m,n-1)$$

and

$$\gamma_1^{(n)} = \min(m,n)\gamma_1^{(n-1)},$$

with

$$\gamma_n^{(n)} = \gamma_{n+1}^{(n+1)} = 1, \quad \forall n \in \mathbb{N}.$$

Proof by mathematical induction.

For $n=1$, $m\geq 1$

$$\begin{aligned}\frac{\partial}{\partial z}G_m(z,t) &= \frac{\partial}{\partial z}\left[\left(\frac{k_1+k_2z}{1-k_3z}\right)^m + C(t)\right]\\ &= m(k_2+k_1k_3)\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-1}(1-k_3z)^{-2}, \quad \gamma_1^{(1)}=1\\ &= \gamma_1^{(1)}\frac{m!}{(m-1)!}(k_2+k_1k_3)\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-1}(1-k_3z)^{-2}\\ &= \sum_{j=1}^{1}\gamma_j^{(1)}\frac{m!}{(m-j)!}k_3^{1-j}(k_2+k_1k_3)^j\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-j}(1-k_3z)^{-(1+j)}\\ &= G_m^{(1)}(z,t) \quad \text{as required.}\end{aligned}$$

Now, suppose equation 4.2 holds for $n=k$, we show it holds for $n=k+1$.

Case 1: $k+1\leq m$

$G_m^{(k+1)}(z,t) = \frac{\partial}{\partial z}G_m^{(k)}(z,t)$

$$= \frac{\partial}{\partial z}\left[\sum_{j=1}^{k}\gamma_j^{(k)}\frac{m!}{(m-j)!}k_3^{k-j}(k_2+k_1k_3)^j\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-j}(1-k_3z)^{-(k+j)}\right]$$

$$= \frac{\partial}{\partial z}\left[\gamma_1^{(k)}\frac{m!}{(m-1)!}k_3^{k-1}(k_2+k_1k_3)\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-1}(1-k_3z)^{-(k+1)}\right]$$

$$+\frac{\partial}{\partial z}\left[\gamma_2^{(k)}\frac{m!}{(m-2)!}k_3^{k-2}(k_2+k_1k_3)^2\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-2}(1-k_3z)^{-(k+2)}\right]+\cdots+$$

$$\frac{\partial}{\partial z}\left[\gamma_{k-1}^{(k)}\frac{m!}{(m-(k-1))!}k_3^{k-(k-1)}(k_2+k_1k_3)^{k-1}\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-(k-1)}(1-k_3z)^{-(k+k-1)}\right]+$$

$$\frac{\partial}{\partial z}\left[\gamma_k^{(k)}\frac{m!}{(m-k)!}k_3^{k-k}(k_2+k_1k_3)^k\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-k}(1-k_3z)^{-(k+k)}\right]$$

$$= \gamma_1^{(k)} \frac{m!}{(m-1)!} k_3^{k-1}(k_2 + k_1 k_3) \left[\left(\frac{k_1 + k_2 z}{1 - k_3 z} \right)^{m-1} \frac{\partial}{\partial z} (1 - k_3 z)^{-(k+1)} + \right.$$

$$\left. (1 - k_3 z)^{-(k+1)} \frac{\partial}{\partial z} \left(\frac{k_1 + k_2 z}{1 - k_3 z} \right)^{m-1} \right] +$$

$$\gamma_2^{(k)} \frac{m!}{(m-2)!} k_3^{k-2}(k_2 + k_1 k_3)^2 \left[\left(\frac{k_1 + k_2 z}{1 - k_3 z} \right)^{m-2} \frac{\partial}{\partial z} (1 - k_3 z)^{-(k+2)} + \right.$$

$$\left. (1 - k_3 z)^{-(k+2)} \frac{\partial}{\partial z} \left(\frac{k_1 + k_2 z}{1 - k_3 z} \right)^{m-2} \right] + \cdots +$$

$$\gamma_{k-1}^{(k)} \frac{m!}{(m-(k-1))!} k_3^{k-(k-1)}(k_2 + k_1 k_3)^{k-1} \left[\left(\frac{k_1 + k_2 z}{1 - k_3 z} \right)^{m-(k-1)} \frac{\partial}{\partial z} (1 - k_3 z)^{-(k+k-1)} + \right.$$

$$\left. (1 - k_3 z)^{-(k+k-1)} \frac{\partial}{\partial z} \left(\frac{k_1 + k_2 z}{1 - k_3 z} \right)^{m-(k-1)} \right] +$$

$$\gamma_k^{(k)} \frac{m!}{(m-k)!} k_3^{k-k}(k_2 + k_1 k_3)^k \left[\left(\frac{k_1 + k_2 z}{1 - k_3 z} \right)^{m-k} \frac{\partial}{\partial z} (1 - k_3 z)^{-(k+k)} + \right.$$

$$\left. (1 - k_3 z)^{-(k+k)} \frac{\partial}{\partial z} \left(\frac{k_1 + k_2 z}{1 - k_3 z} \right)^{m-k} \right]$$

$$= \gamma_1^{(k)} \frac{m!}{(m-1)!} k_3^{k-1}(k_2 + k_1 k_3) \left[\left(\frac{k_1 + k_2 z}{1 - k_3 z} \right)^{m-1} k_3(k+1)(1 - k_3 z)^{-(k+2)} + \right.$$

$$\left. (1 - k_3 z)^{-(k+1)}(m-1) \left(\frac{k_1 + k_2 z}{1 - k_3 z} \right)^{m-2} \frac{k_2 + k_1 k_3}{(1 - k_3 z)^2} \right] +$$

$$\gamma_2^{(k)} \frac{m!}{(m-2)!} k_3^{k-2}(k_2 + k_1 k_3)^2 \left[\left(\frac{k_1 + k_2 z}{1 - k_3 z} \right)^{m-2} k_3(k+2)(1 - k_3 z)^{-(k+3)} + \right.$$

$$\left. (1 - k_3 z)^{-(k+2)}(m-2) \left(\frac{k_1 + k_2 z}{1 - k_3 z} \right)^{m-3} \frac{k_2 + k_1 k_3}{(1 - k_3 z)^2} \right] + \cdots +$$

$$\gamma_{k-1}^{(k)} \frac{m!}{(m-(k-1))!} k_3^{k-(k-1)}(k_2 + k_1 k_3)^{k-1} \left[\left(\frac{k_1 + k_2 z}{1 - k_3 z} \right)^{m-(k-1)} k_3(k+k-1) \cdot \right.$$

$$(1 - k_3 z)^{-(k+(k-1)-1)} +$$

$$\left. (1 - k_3 z)^{-(k+(k-1))}(m-(k-1)) \left(\frac{k_1 + k_2 z}{1 - k_3 z} \right)^{m-k} \frac{k_2 + k_1 k_3}{(1 - k_3 z)^2} \right] +$$

$$\gamma_k^{(k)} \frac{m!}{(m-k)!} k_3^{k-k}(k_2 + k_1 k_3)^k \left[\left(\frac{k_1 + k_2 z}{1 - k_3 z} \right)^{m-k} k_3(k+k) \right.$$

$$\left. (1 - k_3 z)^{-(k+k+1)} + (1 - k_3 z)^{-(k+2)}(m-k) \left(\frac{k_1 + k_2 z}{1 - k_3 z} \right)^{m-(k+1)} \frac{k_2 + k_1 k_3}{(1 - k_3 z)^2} \right]$$

Expanding the terms gives,

$$G_m^{(k+1)}(z,t) = \tfrac{\partial}{\partial z} G_m^{(k)}(z,t)$$
$$= \gamma_1^{(k)}(k+1)\frac{m!}{(m-1)!}k_3^k(k_2+k_1k_3)\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-1}(1-k_3z)^{-(k+2)}$$
$$+\gamma_1^{(k)}\frac{m!}{(m-2)!}k_3^{k-1}(k_2+k_1k_3)^2\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-2}(1-k_3z)^{-(k+3)}$$
$$+\gamma_2^{(k)}(k+2)\frac{m!}{(m-2)!}k_3^{k-1}(k_2+k_1k_3)^2\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-2}(1-k_3z)^{-(k+3)}$$
$$+\gamma_2^{(k)}\frac{m!}{(m-3)!}k_3^{k-2}(k_2+k_1k_3)^3\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-3}(1-k_3z)^{-(k+4)}+\cdots$$
$$+\gamma_{k-1}^{(k)}(k+k-1)\frac{m!}{(m-(k-1))!}k_3^2(k_2+k_1k_3)^{k-1}\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-(k-1)}(1-k_3z)^{-(k+k)}$$
$$+\gamma_{k-1}^{(k)}\frac{m!}{(m-k)!}k_3(k_2+k_1k_3)^k\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-k}(1-k_3z)^{-(k+k+1)}$$
$$+\gamma_k^{(k)}(k+k)\frac{m!}{(m-k)!}k_3(k_2+k_1k_3)^k\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-k}(1-k_3z)^{-(k+k+1)}$$
$$+\gamma_k^{(k)}\frac{m!}{(m-(k+1))!}(k_2+k_1k_3)^{k+1}\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-(k+1)}(1-k_3z)^{-(k+k+2)}$$

Grouping like terms and reorganizing algebraically gives,

$$G_m^{(k+1)}(z,t) = \tfrac{\partial}{\partial z} G_m^{(k)}(z,t)$$
$$= \gamma_1^{(k)}(k+1)\frac{m!}{(m-1)!}k_3^{(k+1)-1}(k_2+k_1k_3)\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-1}(1-k_3z)^{-(k+1+1)}$$
$$+[\gamma_1^{(k)}+\gamma_2^{(k)}(k+1+1)]\frac{m!}{(m-2)!}k_3^{(k+1)-2}(k_2+k_1k_3)^2\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-2}(1-k_3z)^{-(k+1+2)}$$
$$+[\gamma_2^{(k)}+\gamma_3^{(k)}(k+1+2)]\frac{m!}{(m-3)!}k_3^{(k+1)-3}(k_2+k_1k_3)^3\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-3}(1-k_3z)^{-(k+1+3)}$$
$$+\cdots+$$
$$[\gamma_{k-1}^{(k)}+\gamma_k^{(k)}(k+1+k-1)]\frac{m!}{(m-k)!}k_3^{(k+1)-k}(k_2$$
$$+k_1k_3)^k\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-k}(1-k_3z)^{-(k+1+k)}$$
$$+\gamma_k^{(k)}\frac{m!}{(m-(k+1))!}(k_2+k_1k_3)^{k+1}\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-(k+1)}(1-k_3z)^{-(k+1+k+1)}$$

$$=\sum_{j=1}^{k+1}\gamma_j^{(k+1)}\frac{m!}{(m-j)!}k_3^{(k+1)-j}(k_2+k_1k_3)^j\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-j}(1-k_3z)^{-(k+1+j)}$$

where $\gamma_j^{(k+1)} = \gamma_{j-1}^{(k)} + (k+1+j-1)\gamma_j^{(k)}$ for $j = 2,3,\cdots,k$ and $\gamma_1^{(k+1)} = (k+1)\gamma_1^{(k)}$, and $\gamma_{k+1}^{(k+1)} = \gamma_k^{(k)} = 1$ as required.

Case 2: $k+1 \geq m$

$$G_m^{(k+1)}(z,t) = \tfrac{\partial}{\partial z} G_m^{(k)}(z,t)$$

$$= \frac{\partial}{\partial z}\left[\sum_{j=1}^{m} \gamma_j^{(k)} \frac{m!}{(m-j)!} k_3^{k-j}(k_2+k_1k_3)^j \left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-j} (1-k_3z)^{-(k+j)}\right]$$

$$= \frac{\partial}{\partial z}\left[\gamma_1^{(k)} \frac{m!}{(m-1)!} k_3^{k-1}(k_2+k_1k_3) \left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-1} (1-k_3z)^{-(k+1)}\right]$$

$$+\frac{\partial}{\partial z}\left[\gamma_2^{(k)} \frac{m!}{(m-2)!} k_3^{k-2}(k_2+k_1k_3)^2 \left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-2} (1-k_3z)^{-(k+2)}\right] + \cdots +$$

$$\frac{\partial}{\partial z}\left[\gamma_{m-1}^{(k)} \frac{m!}{(m-(m-1))!} k_3^{k-(m-1)}(k_2+k_1k_3)^{m-1} \left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-(m-1)} (1-k_3z)^{-(k+m-1)}\right] +$$

$$\frac{\partial}{\partial z}\left[\gamma_m^{(k)} \frac{m!}{(m-m)!} k_3^{k-m}(k_2+k_1k_3)^m \left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-m} (1-k_3z)^{-(k+m)}\right]$$

$$= \gamma_1^{(k)} \frac{m!}{(m-1)!} k_3^{k-1}(k_2+k_1k_3) \left[\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-1} \frac{\partial}{\partial z}(1-k_3z)^{-(k+1)} +\right.$$

$$\left.(1-k_3z)^{-(k+1)} \frac{\partial}{\partial z}\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-1}\right] +$$

$$\gamma_2^{(k)} \frac{m!}{(m-2)!} k_3^{k-2}(k_2+k_1k_3)^2 \left[\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-2} \frac{\partial}{\partial z}(1-k_3z)^{-(k+2)} +\right.$$

$$\left.(1-k_3z)^{-(k+2)} \frac{\partial}{\partial z}\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-2}\right] + \cdots +$$

$$\gamma_{m-1}^{(k)} \frac{m!}{(m-(m-1))!} k_3^{k-(m-1)}(k_2+k_1k_3)^{m-1} \left[\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-(m-1)} \frac{\partial}{\partial z}(1-k_3z)^{-(k+m-1)}\right.$$

$$+$$

$$\left.(1-k_3z)^{-(k+m-1)} \frac{\partial}{\partial z}\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-(m-1)}\right] +$$

$$\gamma_m^{(k)} \frac{m!}{(m-m)!} k_3^{k-m}(k_2+k_1k_3)^m \left[\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-m} \frac{\partial}{\partial z}(1-k_3z)^{-(k+m)} +\right.$$

$$\left.(1-k_3z)^{-(k+m)} \frac{\partial}{\partial z}\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-m}\right]$$

$$= \gamma_1^{(k)} \frac{m!}{(m-1)!} k_3^{k-1}(k_2 + k_1k_3) \left[\left(\frac{k_1 + k_2z}{1 - k_3z} \right)^{m-1} k_3(k+1)(1-k_3z)^{-(k+2)} + \right.$$

$$\left. (1-k_3z)^{-(k+1)}(m-1) \left(\frac{k_1 + k_2z}{1 - k_3z} \right)^{m-2} \frac{k_2 + k_1k_3}{(1-k_3z)^2} \right] +$$

$$\gamma_2^{(k)} \frac{m!}{(m-2)!} k_3^{k-2}(k_2 + k_1k_3)^2 \left[\left(\frac{k_1 + k_2z}{1 - k_3z} \right)^{m-2} k_3(k+2)(1-k_3z)^{-(k+3)} + \right.$$

$$\left. (1-k_3z)^{-(k+2)}(m-2) \left(\frac{k_1 + k_2z}{1 - k_3z} \right)^{m-3} \frac{k_2 + k_1k_3}{(1-k_3z)^2} \right] + \cdots +$$

$$\gamma_{m-1}^{(k)} \frac{m!}{(m-(m-1))!} k_3^{k-(m-1)}(k_2 + k_1k_3)^{m-1} \left[\left(\frac{k_1 + k_2z}{1 - k_3z} \right) k_3(k+m-1) \right.$$

$$\left. (1-k_3z)^{-(k+m)} + (1-k_3z)^{-(k+m-1)} \frac{k_2 + k_1k_3}{(1-k_3z)^2} \right] +$$

$$\gamma_m^{(k)} \frac{m!}{(m-m)!} k_3^{k-m}(k_2 + k_1k_3)^m \left[k_3(k+m)(1-k_3z)^{-(k+1+m)} \right]$$

Expanding the terms gives,

$$G_m^{(k+1)}(z,t) = \tfrac{\partial}{\partial z} G_m^{(k)}(z,t)$$

$$= \gamma_1^{(k)}(k+1) \frac{m!}{(m-1)!} k_3^{k}(k_2 + k_1k_3) \left(\frac{k_1 + k_2z}{1 - k_3z} \right)^{m-1} (1-k_3z)^{-(k+2)}$$

$$+ \gamma_1^{(k)} \frac{m!}{(m-2)!} k_3^{k-1}(k_2 + k_1k_3)^2 \left(\frac{k_1 + k_2z}{1 - k_3z} \right)^{m-2} (1-k_3z)^{-(k+3)}$$

$$+ \gamma_2^{(k)}(k+2) \frac{m!}{(m-2)!} k_3^{k-1}(k_2 + k_1k_3)^2 \left(\frac{k_1 + k_2z}{1 - k_3z} \right)^{m-2} (1-k_3z)^{-(k+3)}$$

$$+ \gamma_2^{(k)} \frac{m!}{(m-3)!} k_3^{k-2}(k_2 + k_1k_3)^3 \left(\frac{k_1 + k_2z}{1 - k_3z} \right)^{m-3} (1-k_3z)^{-(k+4)}$$

$$+ \gamma_3^{(k)}(k+3) \frac{m!}{(m-3)!} k_3^{k-2}(k_2 + k_1k_3)^3 \left(\frac{k_1 + k_2z}{1 - k_3z} \right)^{m-3} (1-k_3z)^{-(k+4)} + \cdots$$

$$+ \gamma_{m-2}^{(k)} \frac{m!}{(m-(m-1))!} k_3^{k-(m-2)}(k_2 + k_1k_3)^{m-1} \left(\frac{k_1 + k_2z}{1 - k_3z} \right)^{m-(m-1)} (1-k_3z)^{-(k+m)}$$

$$+ \gamma_{m-1}^{(k)}(k+m-1) \frac{m!}{(m-(m-1))!} k_3^{k-(m-2)}(k_2 + k_1k_3)^{m-1} \left(\frac{k_1 + k_2z}{1 - k_3z} \right)^{m-(m-1)} \cdot$$

$$(1-k_3z)^{-(k+m)}$$

$$+ \gamma_{m-1}^{(k)} \frac{m!}{(m-m)!} k_3^{k-(m-1)}(k_2 + k_1k_3)^m (1-k_3z)^{-(k+1+m)}$$

$$+ \gamma_m^{(k)}(k+m) \frac{m!}{(m-m)!} k_3^{k-(m-1)}(k_2 + k_1k_3)^m (1-k_3z)^{-(k+1+m)}$$

Grouping like terms and reorganizing algebraically gives,

$$G_m^{(k+1)}(z,t) = \frac{\partial}{\partial z} G_m^{(k)}(z,t)$$

$$= \gamma_1^{(k)}(k+1)\frac{m!}{(m-1)!}k_3^{(k+1)-1}(k_2+k_1k_3)\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-1}(1-k_3z)^{-(k+1+1)}$$

$$+[\gamma_1^{(k)} + \gamma_{m,2}^{(k)}(k+1+1)]\frac{m!}{(m-2)!}k_3^{(k+1)-2}(k_2+k_1k_3)^2\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-2}(1-k_3z)^{-(k+1+2)}$$

$$+[\gamma_2^{(k)} + \gamma_3^{(k)}(k+1+2)]\frac{m!}{(m-3)!}k_3^{(k+1)-3}(k_2+k_1k_3)^3\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-3}(1-k_3z)^{-(k+1+3)}$$

$$+\cdots$$

$$+[\gamma_{m-2}^{(k)} + \gamma_{m-1}^{(k)}(k+1+m-2)]\frac{m!}{(m-(m-1))!}k_3^{(k+1)-(m-1)}(k_2+k_1k_3)^{(m-1)}\cdot$$

$$\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-(m-1)}(1-k_3z)^{-(k+1+m-1)}$$

$$+[\gamma_{m-1}^{(k)} + \gamma_m^{(k)}(k+1+m-1)]\frac{m!}{(m-m)!}k_3^{(k+1)-m}(k_2+k_1k_3)^m\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-m}\cdot$$

$$(1-k_3z)^{-(k+1+m)}$$

$$= \sum_{j=1}^{m}\gamma_j^{(k+1)}\frac{m!}{(m-j)!}k_3^{(k+1)-j}(k_2+k_1k_3)^j\left(\frac{k_1+k_2z}{1-k_3z}\right)^{m-j}(1-k_3z)^{-(k+1+j)};$$

where $\gamma_j^{(k+1)} = \gamma_{j-1}^{(k)} + (k+1+j-1)\gamma_j^{(k)}$ for all $2 \le j \le m$, $\gamma_1^{(k+1)} = (k+1)\gamma_1^{(k)}$, Q.E.D. and $\gamma_1^{(1)} = 1$ as required.

Remark. Given Theorem 2, we can directly derive the exact transition function of the B-D-C process $X(t)$ from the *nth* derivative of its PGF given by equation 4.1 (w.r.t. z) as indicated in Corollary 2.1.

Corollary 2.1. The analytical form of the transition function of the process $X(t)$ for $m,n\geq 0$ is given as:

$$P_{m,n}(t) = \begin{cases} \sum_{j=1}^{\min(m,n)} \gamma_j^{(n)} \frac{m!}{n!(m-j)!} k_3^{n-j}(k_2+k_1k_3)^j k_1^{m-j} & m,n\geq 1 \\ k_1^m + C(t) & n=0, \quad m\geq 1 \\ 1 & m,n=0 \\ 0 & \text{otherwise.} \end{cases} \tag{4.3}$$

Proof. Suppose $m,n \geq 0$

Case 1: $m,n \geq 1$

$$\begin{aligned}
P_{m,n}(t) = P\{X(t) = n | X(0) = m\} &= \frac{G_m^{(n)}(0,t)}{n!} \\
&= \sum_{j=1}^{\min(m,n)} \gamma_j^{(n)} \frac{m!}{n!\,(m-j)!} k_3^{n-j} (k_2 + k_1 k_3)^j \left(\frac{k_1 + k_2 \cdot 0}{1 - k_3 \cdot 0} \right)^{m-j} (1 - k_3 \cdot 0)^{-(n+j)} \\
&= \sum_{j=1}^{\min(m,n)} \gamma_j^{(n)} \frac{m!}{n!\,(m-j)!} k_3^{n-j} (k_2 + k_1 k_3)^j k_1^{m-j}.
\end{aligned}$$

Case 2: Let $n = 0$ and $m \geq 1$

Given the generating function $G_m(z,t)$,

$$\begin{aligned}
P_{m,0}(t) &= G_m(0,t) \\
&= \left(\frac{k_1 + k_2 \cdot 0}{1 - k_3 \cdot 0} \right)^m + C(t) = k_1^m + C(t).
\end{aligned}$$

Case 3: Suppose $m = n = 0$

Qualitatively, given that there are $m = 0$ parasites at time $t = 0$, there is a certain probability of having $n = 0$ parasites at any time $t > 0$ since no birth can occur. Now, since

$$P_{m,0}(t) = k_1^m + C(t)$$

; where

$$C(t) = 1 - \left(\frac{k_1 + k_2}{1 - k_3} \right)^m \quad \text{for} \quad k_1, k_2, k_3 > 0,$$

$$\implies P_{0,0}(t) = P\{X(t) = 0 | X(0) = 0\} = P_{m,0}(t)|_{m=0} = k_1^0 + 1 - \left(\frac{k_1 + k_2}{1 - k_3} \right)^0 = 1 \qquad \text{Q. E. D.}$$

Hence, the transition function $P_{m,n}(t)$ given by Corollary 2.1 is true for $m,n \geq 0$ as required.

Remark. Given the *nth* derivative of the B-D-C process $X(t)$ as defined under Theorem 2, we can further derive at least the first to third theoretical moments of $X(t)$ as indicated in Corollary 2.2.

Corollary 2.2. The expected value, variance and the 3rd uncentered moment of the B-D-C process $X(t)$ can be derived from equation 4.1 such that

$$\begin{aligned} E\{X(t)|X(0)=m\} &= \frac{\partial}{\partial z}G_m(z,t)\bigg|_{z=1} \\ &= \frac{m(k_2+k_1k_3)}{(1-k_3)^2}\left(\frac{k_1+k_2}{1-k_3}\right)^{m-1}. \end{aligned} \tag{4.4}$$

$$Var\{X(t)|X(0)=m\} = \frac{\partial^2}{\partial z^2}G_m(z,t)\bigg|_{z=1} + E\{X(t)|X(0)=m\} - [E\{X(t)|X(0)=m\}]^2, \tag{4.5}$$

where

$$\frac{\partial^2}{\partial z^2}G_m(z,t)\bigg|_{z=1} = \frac{2mk_3(k_2+k_1k_3)}{(1-k_3)^3}\left(\frac{k_1+k_2}{1-k_3}\right)^{m-1} + \frac{m(m-1)(k_2+k_1k_3)^2}{(1-k_3)^4}\left(\frac{k_1+k_2}{1-k_3}\right)^{m-2}.$$

Additionally,

$$E\{X^3(t)|X(0)=m\} = \frac{\partial^3}{\partial z^3}G_m(z,t)\bigg|_{z=1} + 3E\{X^2(t)|X(0)=m\} - 2E\{X(t)|X(0)=m\} \tag{4.6}$$

with

$$\begin{aligned} \frac{\partial^3}{\partial z^3}G_m(z,t)\bigg|_{z=1} = \frac{6m(k_2+k_1k_3)k_3^2}{(1-k_3)^4}\left(\frac{k_1+k_2}{1-k_3}\right)^{m-1} + \frac{6m(m-1)(k_2+k_1k_3)^2k_3}{(1-k_3)^5}\left(\frac{k_1+k_2}{1-k_3}\right)^{m-2} \\ + \frac{m(m-1)(m-2)(k_2+k_1k_3)^3}{(1-k_3)^6}\left(\frac{k_1+k_2}{1-k_3}\right)^{m-3}, \end{aligned}$$

and

$$E\{X^2(t)|X(0)=m\} = Var\{X(t)|X(0)=m\} + [E\{X(t)|X(0)=m\}]^2. \tag{4.7}$$

The 1st, 2nd and 3rd uncentered theoretical moments are useful for the generalised method of moments estimation of the B-D-C process described in Section 4.2.

4.1.3 Numerical validation of the derived transition function of the B-D-C process

The derived transition function $P_{m,n}(t)$ and its theoretical moments (mean and variance functions) were further validated numerically based on 1 million exact stochastic simulations of the B-D-C process (for pseudo-code and R codes of the exact SSA of the B-D-C process, see Algorithm 1 and Appendix C) at given parameter values ($\lambda = 0.512$, $\mu = 0.35$ and $\rho = 0.003$) and $X_0 = 2$ as a case study. Algorithm 1 is a novel B-D-C simulation algorithm developed (in the current study) by adapting the standard Monte Carlo stochastic simulation technique to also include host survival status. The analytical transition probabilities (computed from corollary 2.1) and their corresponding Monte Carlo estimates ($\hat{p}_{2,n}(t) = k_n(t)/N$, where $k_n(t)$ is the frequency of occurrence for each value of $n \geq 0$ at time t and N is the total number of simulations) were compared over time (Figure 4.1). Figure 4.1 is a goodness-of-fit plot of the analytical and Monte Carlo estimates of the transition probabilities for different values $n \geq 0$ at six time points ($t = 1$ to $t = 6$). The estimated 95% confidence intervals for the Monte Carlo estimates of the transition probabilities were very small in size due to the smaller values of their respective standard errors as a result of the large number of simulations; and thus, not presented. However, the analytical transition probabilities were within the estimated confidence intervals at least 90% of the time (i.e., coverage probability=0.90). This is because, in practice, a 95% confidence level (based on normal approximation) does not necessarily guarantee a 95% coverage probability. The sampling distribution of the true mean and variance functions (equations 4.4 and 4.5) of the B-D-C process were respectively compared with their corresponding Monte Carlo estimates (Figure 4.2).

Remark. We can deduce from Figures 4.1 and 4.2 that the Monte Carlo estimates of the transition probabilities, mean and variance of the B-D-C process $X(t)$ are consistent with the theoretical values. Hence, the derived transition function can be used to accurately estimate transition probabilities of the B-D-C process.

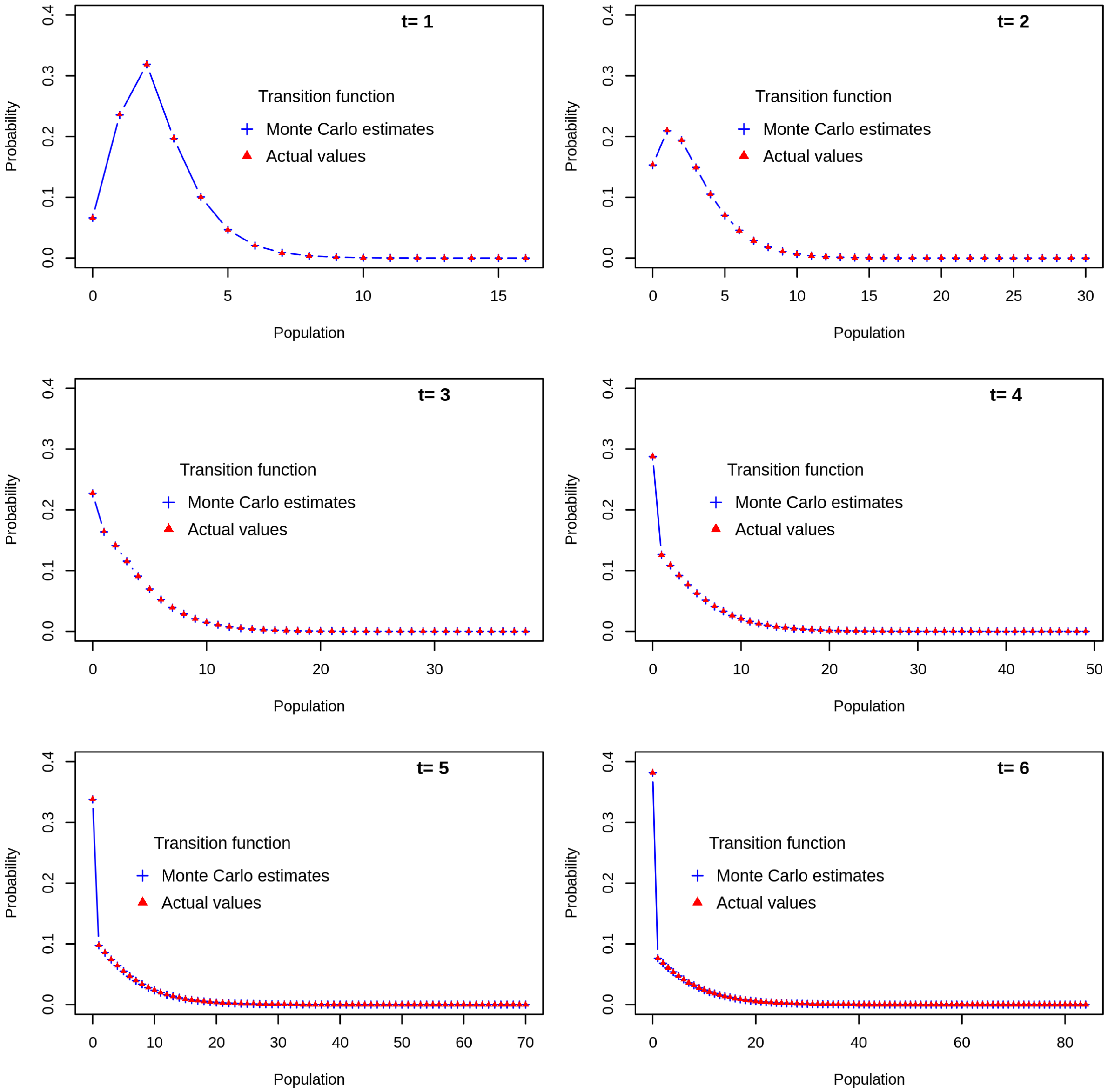

Figure 4.1: Comparison between analytical and numerical estimates of transition probabilities of the B-D-C process at given parameter values ($\lambda = 0.512$, $\mu = 0.35$ and $\rho = 0.003$) from $t = 1$ to $t = 6$ (in days).

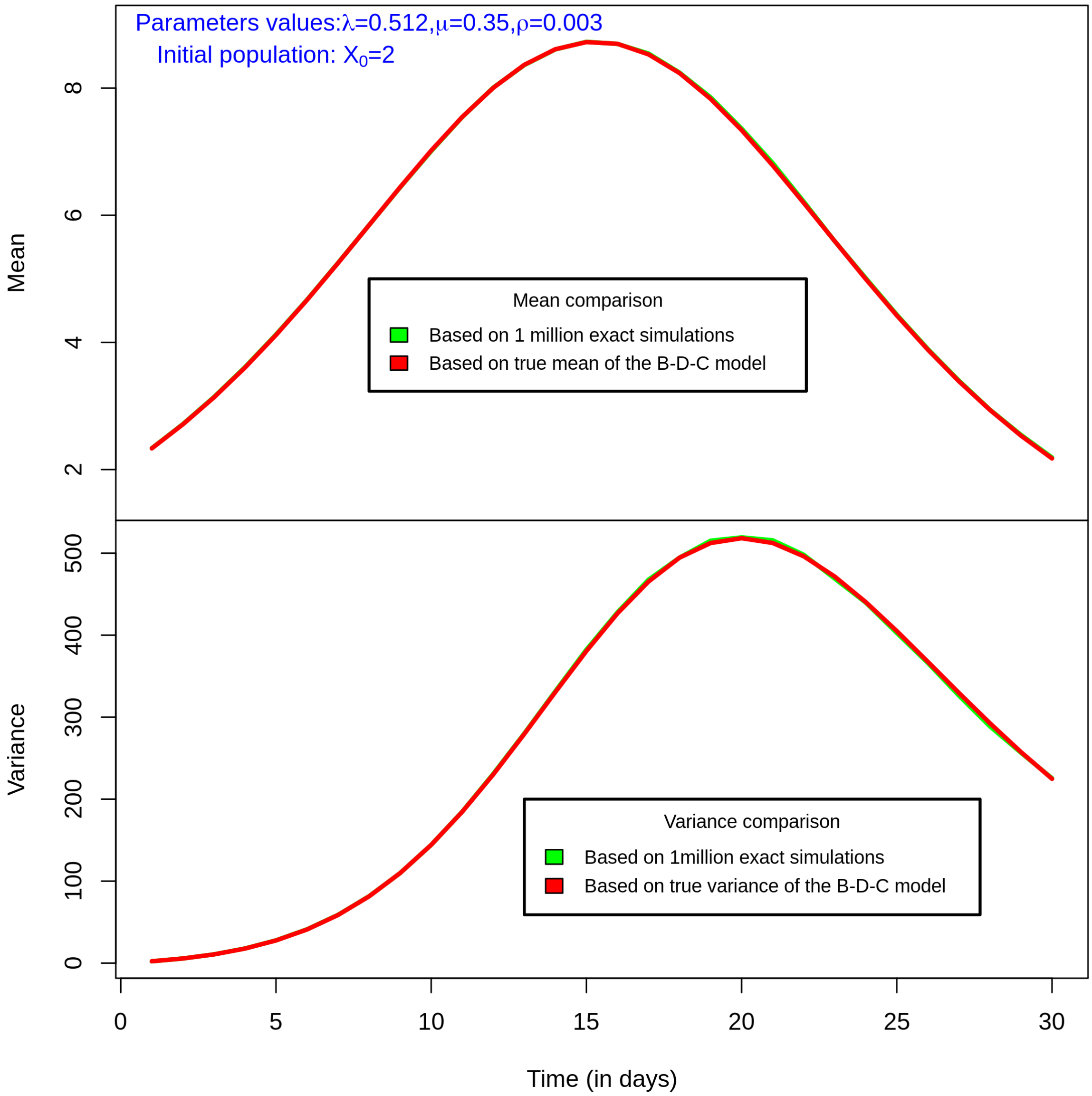

Figure 4.2: Comparison between analytical and Monte Carlo estimates of the mean and variance of the B-D-C process over time ($0 \leq t \leq 30$ days).

Algorithm 1: Exact SSA of the B-D-C process (pseudo-code)

```
   Input: $X$, $\lambda$, $\mu$, $\rho$, $t$, $t_{\text{final}}$ and host survival status ($s$).
   Output: Parasite numbers and survival status (alive: $s = 1$; dead: $s = 2$)
           recorded at discrete times ($t = 1, 2, \cdots, t_{\text{final}}$).
1  while $t < t_{final}$ and $s = 1$ do
2      Set initial time $t = t_0$, state $X = X_0$ and $s = 1$.
3      Calculate rates corresponding to birth ($a_1$), death ($a_2$) and catastrophe ($a_3$);
         such that $a_1 = \lambda X$, $a_2 = \mu X$ and $a_3 = \rho X$.
4      Compute the total rate, $a_0 = \sum_{j=1}^{3} a_j$,  for $j = 1, 2, 3$ (from step 3).
5      Determine the event to occur using a random number $u$ from $Uniform(0, a_0)$
6      if  $0 \leq u \leq a_1$ then
7          set $X = X + 1$
8      else
9          
10     end
11     else if $a_1 < u \leq a_1 + a_2$ then
12         set $X = X - 1$
13
14     else
15         set $X = 0$ and $s = 2$ (then stop the simulation).
16     end
17
18     Generate time increment $\tau$ from $Exponential(a_0)$, and update the time such
         that $t = t + \tau$.
19     Record $(X, s)$ at the desired discrete times.
20 end
```

4.1.4 Behaviour of the transition, mean and variance functions of the B-D-C process

We further explored the behaviour of the exact transition function by varying the values of m ($1 \leq m \leq 10$) and n ($1 \leq n \leq 10$) at pre-specified parameter values ($\lambda = 0.512$, $\mu = 0.35$ and $\rho = 0.003$). Figure 4.3 shows that for any fixed value of $m > 0$ and $n > 0$, transition probabilities decrease over time for $t > 0$ and the probability of transitioning to state n decreases with increasing values of m. The behaviour of the B-D-C process's mean and variance functions was also explored over time at $m = 2$ and different parameter values by varying the rates in a full factorial design (Figure 4.4). This combination of parameter values was pre-specified to determine how the choice of values can affect the mean (or parasite population) distribution over time and help determine instances where the parasite population from the B-D-C process is low, moderate, or high (for the sake of three different *in silico* simulation experiments to be carried out to assess the computational speed and accuracy of the B-D-C parameter estimation techniques proposed under the next section 4.2). Generally, the variance of the process $X(t)$ is far greater than its mean at any value of $t > 1$ and any given parameter values.

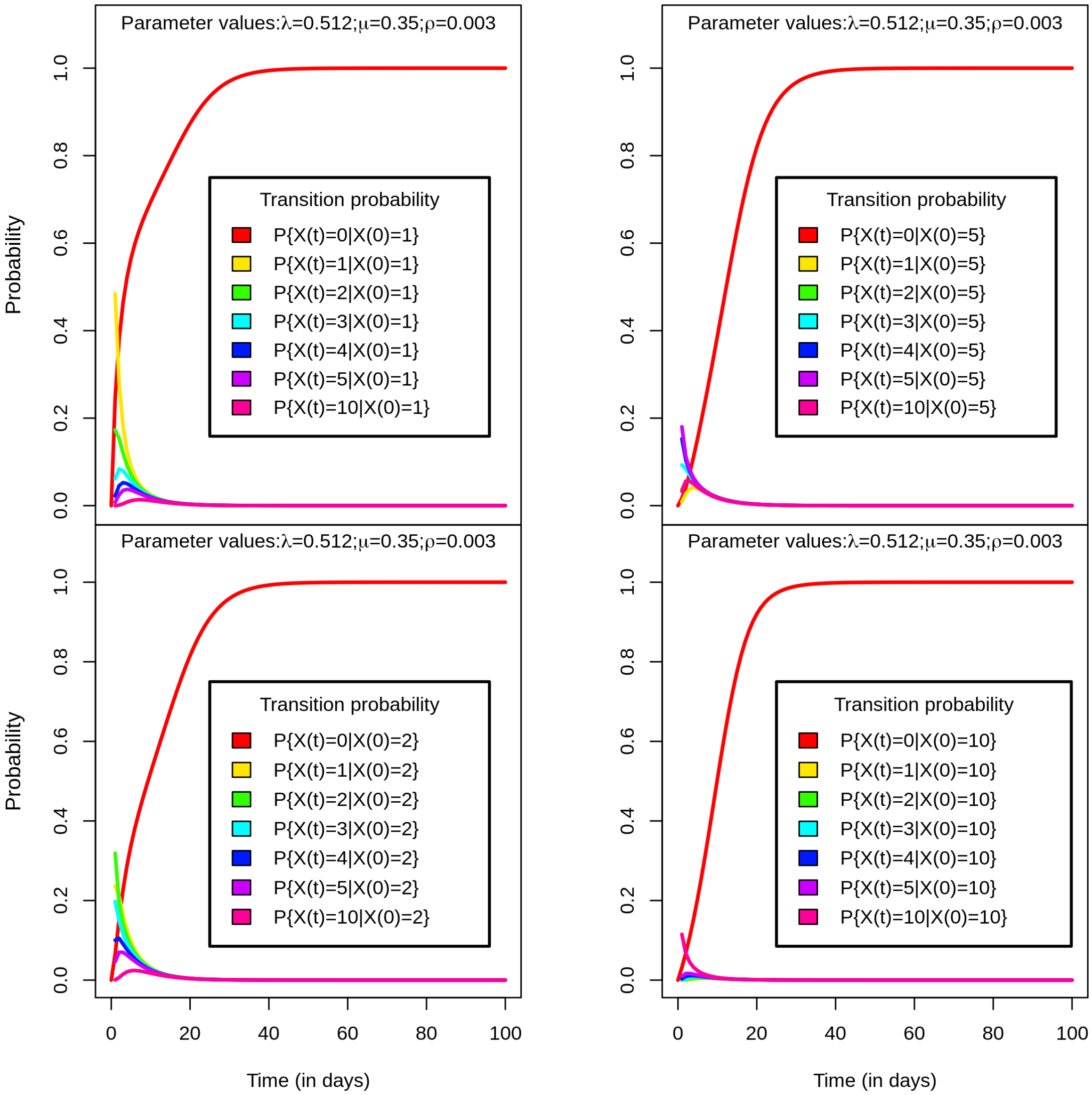

Figure 4.3: Exact transition probabilities of the B-D-C process at pre-specified parameter values ($\lambda = 0.512$, $\mu = 0.35$ and $\rho = 0.003$) for different values of m and n ($0 \leq m, n \leq 10$) over time ($0 \leq t \leq 100$ days).

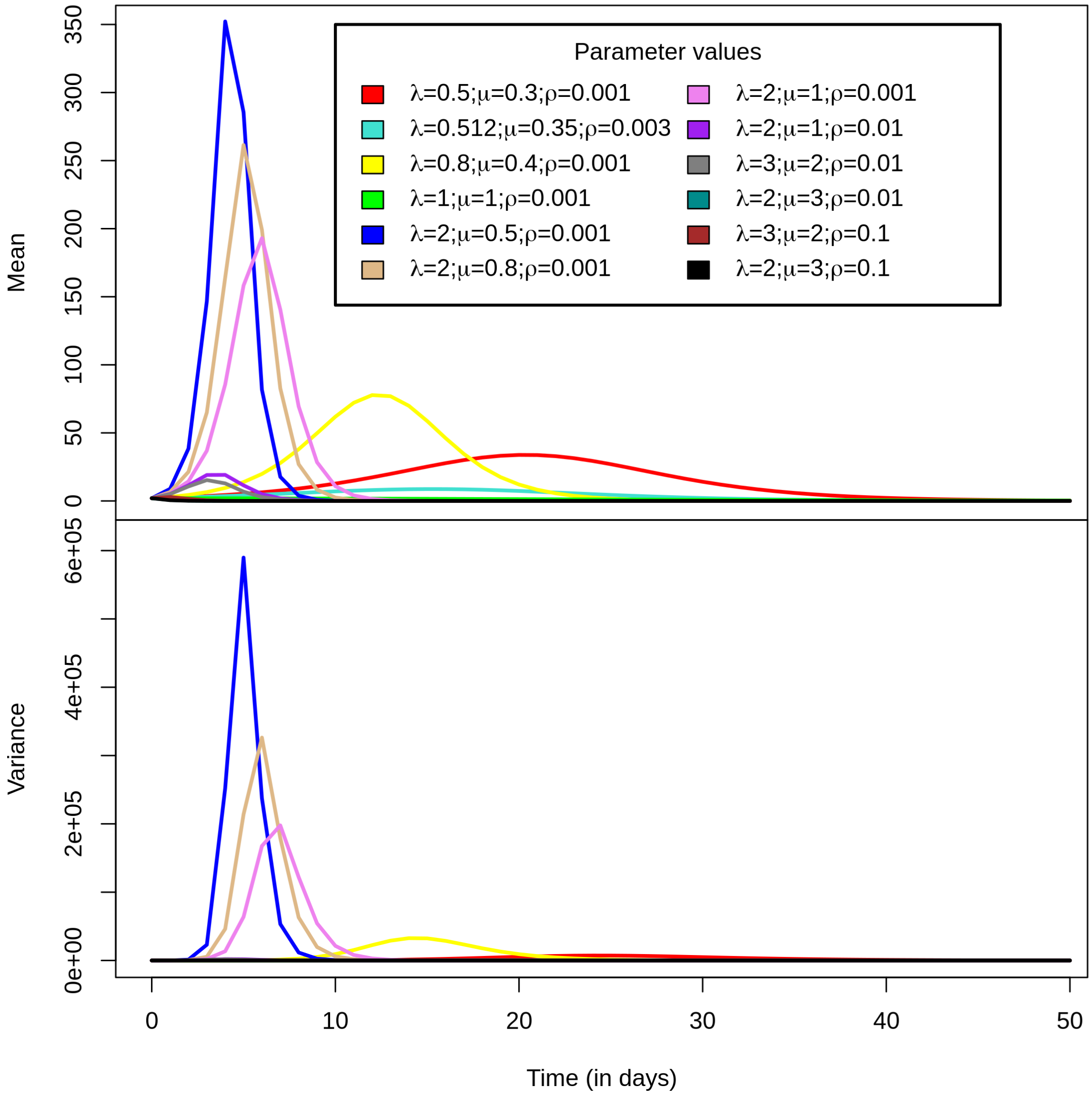

Figure 4.4: Exact mean and variance of the B-D-C process at different pre-specified parameter values ($0.5 \leq \lambda \leq 3$; $0.3 \leq \mu \leq 3$; $0.001 \leq \rho \leq 0.1$) over time ($0 \leq t \leq 50$ days).

4.2 Parameter estimation of the B-D-C process

We estimated the parameters of the B-D-C model by comparing between three estimation methods: Maximum likelihood estimation (MLE), Generalised method of moments (GMM) and Embedded Galton-Watson approach (GW), based on three different simulation experiments repeated 100 times (with $X_0 = 2$); such that the parasite population size were low ($\lambda = 3$, $\mu = 2$ and $\rho = 0.1$), moderate ($\lambda = 2$, $\mu = 1$ and $\rho = 0.01$) or high ($\lambda = 0.5$, $\mu = 0.3$ and $\rho = 0.001$). The main motivation is to identify which estimation method achieves good estimates (based on their bias, variance and mean square error of the estimates, respectively) but with faster computational time for the different simulation experiments. These parameter estimates will improve the summary statistics of a modified ABC based on a complex stochastic simulation model for the gyrodactylid-host system (see Chapters 5 and 6). The estimation methods were set-up to distinguish between the two zero states of the B-D-C process at any time t due to natural death of parasites or catastrophic extinction, such that

$$P\{X(t) = 0 \ \text{ and host alive}|X(0) = m\} = k_1^m$$

and

$$P\{X(t) = 0 \ \text{ and host dead}|X(0) = m\} = 1 - \left(\frac{k_1 + k_2}{1 - k_3}\right)^m,$$

with k_1, k_2 and k_3 defined as before.

For each *in silico* simulation experiment (Cases 1-3), we simulated data for 50 independent hosts. We recorded parasite numbers for each host over a 17-day period (odd-numbered days from day 1 to 17) to represent one simulation realisation. The simulation is set-up like the observed empirical data for the sophisticated stochastic model. We fit the MLE (described in Section 4.2.1) as well as the GMM and GW estimators (described in Sections 4.2.2 and 4.2.3, respectively) to the 50 simulated data for 100 different simulation realisations, respectively. The bias, variance and mean square error of the parameter estimates are computed for comparison. For any biased estimator, we explored its consistency as the sample size becomes large. The three different *in silico* simulation experiments were considered in determining whether the estimation methods will give reasonable estimates irrespective of the parasite population size over time as well as investigate the computational speed in such instances. A major assumption made for these parameter estimation methods is that data from different simulations and *in silico* simulation experiments are independent and identically distributed.

4.2.1 Maximum likelihood estimator

Suppose $X = \{X(t_1), X(t_2), \cdots X(t_9)\}$ is the number of parasites on a host at time t_i, $i = 1,2,3,\cdots,9$ and follows the B-D-C process with transition function defined in Corollary 2.1. Let n_r be the total number of hosts for each realisation for $r = 1,2,\cdots,100$ (n_r was set at 50 for each simulation realisation). Suppose n_{ki} is the number of parasites on the kth host for $k = 1,2,\cdots,n_r$ at time t_i, $i = 1,2,3,\cdots,9$. By employing the Markov property, the likelihood function corresponding to each simulation realisation for r is given as

$$\begin{aligned}
L(X,\lambda,\mu,\rho) &= \prod_{k=1}^{n_r} P\{X(t_9) = n_{k9}, X(t_8) = n_{k8}, \cdots, X(t_1) = n_{k1} | X(t_0) = n_{k0}\} \\
&= \prod_{k=1}^{n_r} P\{X(t_1) = n_{k1} | X(0) = n_{k0}\} \times P\{X(t_2) = n_{k2} | X(t_1) = n_{k1}\} \times \cdots \times \\
&\qquad P\{X(t_9) = n_{k9} | X(t_8) = n_{k8}\} \\
&= \prod_{k=1}^{n_r} \prod_{i=1}^{9} P\{X(t_i - t_{i-1}) = n_{ki} | X(0) = n_{k0}\}. \qquad (4.8)
\end{aligned}$$

Taking logarithm of $L(X,\lambda,\mu,\rho)$ (equation 4.8) gives the log-likelihood function $l(X,\lambda,\mu,\rho)$ (equation 4.9):

$$l(X,\lambda,\mu,\rho) = \sum_{k=1}^{n_r}\sum_{i=1}^{9} log\left(P\{X(t_i - t_{i-1}) = n_{ki}|X(t_0) = n_{k0}\}\right) \quad \text{for} \quad r = 1,2,\cdots,100. \tag{4.9}$$

Now, suppose $\theta = [\lambda,\mu,\rho]^T$ and $\Theta = \{\theta \in \mathbb{R}^3|\lambda > 0,\ \mu > 0,\ \rho > 0\}$ is the parameter space of θ, then the MLE ($\hat{\theta}_{\text{MLE}}$) is obtained such that

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta\in\Theta} l(X,\theta). \tag{4.10}$$

The maximum likelihood estimates were obtained numerically in R across the entire parameter space Θ by maximising the log-likelihood function for each simulation experiment (Cases 1-3). Due to round-off errors in storing the transition probabilities for large parasite numbers in R version 3.6.3 [247], arbitrary precision arithmetic in Julia version 1.5.3 [33] was employed to store these probabilities as double-precision binary floating-point numbers (i.e., 64-bit floating-point numbers with 15 decimal digits of precision); however, arbitrary-precision arithmetic is considerably slow. Additionally, the probabilities were efficiently computed recursively by pre-calculating various components of the transition function before computing the log-likelihood function (for the Julia codes of the recursive transition function and log-likelihood function, see Appendix D).

4.2.2 Generalised method of moments estimator

Let $\varphi_d(\theta,t_i) = E\{X_r^d(t_i)|X(0) = m\}$ represent the first d exact theoretical moments and $\bar{\varphi}_d(X,t_i)$ denoting their corresponding sample moments for $d = 1,2,3$ for simulation realisation r such that

$$\varphi_1(\theta,t_i) = \frac{m(k_2+k_1k_3)}{(1-k_3)^2}\left(\frac{k_1+k_2}{1-k_3}\right)^{m-1},$$

$$\varphi_2(\theta,t_i) = \frac{2mk_3(k_2+k_1k_3)}{(1-k_3)^3}\left(\frac{k_1+k_2}{1-k_3}\right)^{m-1} + \frac{m(m-1)(k_2+k_1k_3)^2}{(1-k_3)^4}\left(\frac{k_1+k_2}{1-k_3}\right)^{m-2} + \varphi_1(\theta,t_i),$$

$$\varphi_3(\theta,t_i) = \frac{6m(k_2+k_1k_3)k_3^2}{(1-k_3)^4}\left(\frac{k_1+k_2}{1-k_3}\right)^{m-1} + \frac{6m(m-1)(k_2+k_1k_3)^2k_3}{(1-k_3)^5}\left(\frac{k_1+k_2}{1-k_3}\right)^{m-2}$$
$$+\frac{m(m-1)(m-2)(k_2+k_1k_3)^3}{(1-k_3)^6}\left(\frac{k_1+k_2}{1-k_3}\right)^{m-3} + 3\varphi_2(\theta,t_i) - 2\varphi_1(\theta,t_i);$$

and their corresponding sample moments at time t_i, $i = 1,2\cdots,9$ is given by

$$\bar{\varphi}_d(X,t_i) = \frac{1}{n_r}\sum_{k=1}^{n_r} X_k^d(t_i) \text{ for } d = 1,2,3 \text{ and } r = 1,2,\cdots,100.$$

Suppose $\varphi_4(\theta,t_i)$ is the theoretical probability of catastrophe and let $\bar{\varphi}_4(X,t_i)$ be its sample estimate at time t_i such that

$$\varphi_4(\theta,t_i) = 1-\left(\frac{k_1+k_2}{1-k_3}\right)^m \qquad \text{and} \qquad \bar{\varphi}_4(X,t_i) = \frac{\omega_{ir}}{n_r};$$

where ω_{ir} is the number of hosts who died at time t_i out of the n_r total number of hosts who were alive at time 0 in the rth simulation realisation. Here, we assume an infected host dies and parasite population extinction occur at a rate proportional to the number of parasites $X(t_i)$ on host at time t_i.

Let $g(X,\theta)$ be a 9×4 matrix with entries $g_{ij}(X,\theta)$ (for $1 \le i \le 9$ and $1 \le j \le 4$) defined such that

$$g_{ij}(X,\theta) = \varphi_j(\theta,t_i) - \bar{\varphi}_j(X,t_i) \tag{4.11}$$

and satisfying the 9×1 unconditional moment conditions:

$$E\left[g(X,\theta)\right] = 0. \tag{4.12}$$

A two-step generalised method of moments technique originally proposed by Hansen [124] was employed to numerically obtain efficient GMM. Now, suppose that

$$\bar{g}_j(\theta) = \frac{1}{9}\sum_{i=1}^{9} g_{ij}(X,\theta) \text{ for } j = 1,2,3,4. \tag{4.13}$$

Then, the GMM estimator is given by

$$\hat{\theta}_{\text{GMM}} = \underset{\theta \in \Theta}{\arg\min}\, \bar{g}(\theta)^T \Omega(\theta^*)^{-1} \bar{g}(\theta), \tag{4.14}$$

where $\Omega(\theta^*)^{-1}$ is a positive-definite weighting matrix (a diagonal matrix with leading diagonal representing the multiplicative inverse of the variance of jth column entries of matrix g evaluated at θ^* as defined in equation 4.15), and θ^* is the initial parameter estimates at the first estimation step with the weighting matrix being a 4×4 identity matrix. For the weighting matrix at the second estimation step, we assume that the column entries or components (g_j) of matrix $g(X, \theta^*)$ are uncorrelated for $j = 1, 2, 3, 4$. The 2-step algorithm for obtaining $\hat{\theta}_{\text{GMM}}$ is therefore:

1. Estimate $\theta^* = \underset{\theta \in \Theta}{\arg\min}\, \bar{g}(\theta)^T \bar{g}(\theta)$.

2. Estimate the weighting matrix $\Omega(\theta^*)^{-1}$ given θ^*

$$\Omega(\theta^*)^{-1} = \begin{array}{c} \\ g_1 \\ g_2 \\ g_3 \\ g_4 \end{array} \begin{array}{c} \begin{array}{cccc} g_1 & g_2 & g_3 & g_4 \end{array} \\ \begin{pmatrix} \frac{1}{\sigma^2_{g_1(\theta^*)}} & 0 & 0 & 0 \\ 0 & \frac{1}{\sigma^2_{g_2(\theta^*)}} & 0 & 0 \\ 0 & 0 & \frac{1}{\sigma^2_{g_3(\theta^*)}} & 0 \\ 0 & 0 & 0 & \frac{1}{\sigma^2_{g_4(\theta^*)}} \end{pmatrix} \end{array}.$$

where $\sigma^2_{g_j(\theta^*)}$ is the variance of jth column entries of matrix $g(X, \theta^*)$ such that

$$\sigma^2_{g_j(\theta^*)} = \frac{\sum_{i=1}^{9} \left[g_{ij}(X, \theta^*) - \bar{g}_j(\theta^*)\right]^2}{8} \qquad \text{for} \quad j = 1, 2, 3, 4. \tag{4.15}$$

3. Compute the required estimates $\hat{\theta}_{\text{GMM}}$ given by equation 4.14.

By the weak law of large numbers,

$$\bar{\varphi}_j(X, t_i) \xrightarrow{p} \varphi_j(\theta, t_i) \text{ as } n_r \to \infty \text{ for } 1 \leq i \leq 9,\ 1 \leq j \leq 4 \text{ and } 1 \leq r \leq 100.$$

Hence, if the moment conditions are met, GMM estimates are consistent as the sample size increases.

4.2.3 Embedded Galton-Watson estimation approach

A single trajectory of a discretely-observed linear birth-death process (LBDP) with birth rate λ, death rate μ and equal inter-observation times Δt corresponds to a Galton-Watson (GW) process with offspring mean $m = m(\Delta t)$ and variance $\sigma^2 = \sigma^2(\Delta t)$ [75, 125]; where

$$\hat{m} = \frac{\sum\limits_{i=1}^{\mathcal{T}} Z_i}{\sum\limits_{i=1}^{\mathcal{T}} Z_{i-1}} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{\mathcal{T}} \sum_{i=1}^{\mathcal{T}} Z_{i-1} \left(\frac{Z_i}{Z_{i-1}} - \hat{m} \right)^2$$

with $Z_0 = Z(0)$ denoting the initial population size at time $t = 0$, the successive population sizes represented as $Z_1 = Z(\Delta t), Z_2 = Z(2\Delta t), \cdots, Z_{\mathcal{T}} = Z(\mathcal{T}\Delta t)$, and $\mathcal{T}$ being the final observed time. The GW process is thus the most basic (branching process) model for a population changing through time. It assumes that given an initial number of individuals or parasites ($Z_0 = 1$) at the zeroth generation (or at $t = 0$), each individual in the population at any time $t \geq 1$ can give birth (at a constant rate $\lambda > 0$) to offspring with the same probability distribution, independently of one another; whereas, each individual can die (at a constant rate $\mu > 0$) until population extinction may occur when the offspring mean $m < 1$. Also, analytical estimates of the birth and death rates (λ and μ) given a finite number of independent trajectories (or initial population size > 1) of a discretely-observed LBDP with equal inter-observation times conditioned on population non-extinction have been proposed in the general case by Davison et al. [75] as an embedded GW process (i.e., the more general or extended case with $Z_0 \geq 1$).

Now, let suppose $Z_{i,k}$ is the number of parasites on each kth surviving host at time t_i for $i = 1, 2, \cdots, 9$, $k = 1, 2, \cdots, s_r$, $r = 1, 2, \cdots, 100$ and equal inter-observation times $\Delta t = t_i - t_{i-1}$ (where s_r is the total number of surviving hosts for each rth simulation realisation). Suppose $Z_{i,k}$ with birth rate $\lambda > 0$ and death rate $\mu > 0$, is a GW process with mean $m(\Delta t) > 1$ and variance $\sigma^2(\Delta t) > 0$ (where $Z_{0,k} = 2$ in our case). Then, the analytical estimates of the birth and death rates for $k>1$ independent trajectories of the LBDP and equal Δt ($\Delta t = 2$ in our case) is given as

$$\hat{\lambda} = \frac{\log \hat{m}}{2\Delta t}\left\{\frac{\hat{\sigma}^2}{\hat{m}(\hat{m}-1)}+1\right\} \quad \text{and} \quad \hat{\mu} = \frac{\log \hat{m}}{2\Delta t}\left\{\frac{\hat{\sigma}^2}{\hat{m}(\hat{m}-1)}-1\right\}; \tag{4.16}$$

where

$$\hat{m} = \frac{\sum\limits_{k=1}^{s_r}\sum\limits_{i=1}^{9} Z_{i,k}}{\sum\limits_{k=1}^{s_r}\sum\limits_{i=1}^{9} Z_{i-1,k}} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{9s_r}\sum_{k=1}^{s_r}\sum_{i=1}^{9} Z_{i-1,k}\left(\frac{Z_{i,k}}{Z_{i-1,k}}-\hat{m}\right)^2.$$

From equation 4.16, it can be observed that when the mean $\hat{m}<1$ (case of subcritical process), $\hat{\lambda}<0$; and at $\hat{m}=1$ (case of critical process), $\hat{\lambda}$ and $\hat{\mu}$ are undefined. According to Davison et al. [75], $\hat{\lambda}=\hat{\mu}=\frac{\hat{\sigma}^2}{2\Delta t}$ when $\hat{m}=1$. In the non-critical case where the offspring mean $\hat{m}>1$, the birth and death rates of the B-D-C process are estimated based on equation 4.16 by conditioning on non-extinction or survival of host. To estimate the catastrophic rate (ρ) when $\hat{m}>1$, we estimate ρ using maximum likelihood estimation given by equation 4.18 based on the GW estimates of the birth and death rates ($\hat{\lambda}$ and $\hat{\mu}$ from equation 4.16). Suppose hosts die in the time interval $(t_{k_{i-1}}, t_{k_i}]$ at rate ρ, then the log-likelihood function is given as

$$l(\rho;\hat{\lambda},\hat{\mu}) = \sum_{k=1}^{n_r-s_r}\log\left[C(t_{k_i})-C(t_{k_{i-1}})\right] \quad \text{for} \quad r=1,2,\cdots,100; \tag{4.17}$$

where $C(t)$ is the theoretical probability of catastrophe given by

$$C(t) = 1-\left(\frac{k_1+k_2}{1-k_3}\right)^m,$$

with $k_1=\frac{\nu_0\nu_1(1-\sigma)}{\nu_1-\sigma\nu_0}$, $k_2=\frac{\nu_1\sigma-\nu_0}{\nu_1-\sigma\nu_0}$, $k_3=\frac{1-\sigma}{\nu_1-\sigma\nu_0}$, $\sigma=\mathrm{e}^{-\hat{\lambda}(\nu_1-\nu_0)t}$, $\nu_0=\frac{(\hat{\lambda}+\hat{\mu}+\rho)-\sqrt{(\hat{\lambda}+\hat{\mu}+\rho)^2-4\hat{\mu}\hat{\lambda}}}{2\hat{\lambda}}$, and $\nu_1=\frac{(\hat{\lambda}+\hat{\mu}+\rho)+\sqrt{(\hat{\lambda}+\hat{\mu}+\rho)^2-4\hat{\mu}\hat{\lambda}}}{2\hat{\lambda}}$. The MLE estimator of the catastrophe rate given the GW estimates of the birth and death rates is obtained such that

$$\hat{\rho} = \arg\max_{\rho\in\Theta} l(\rho;\hat{\lambda},\hat{\mu}). \tag{4.18}$$

Remark. The fastest computational time is expected to be achieved from the analytical estimation of the birth and death rates ($\hat{\lambda}$ and $\hat{\mu}$, respectively) using the GW estimator given by equation 4.16 together with the MLE estimator of the catastrophe rate ρ given by equation 4.18 compared to the computational time required for the MLE and GMM estimators given by equations equations 4.10 and 4.14, respectively. For instances where the offspring mean $\hat{m} \leq 1$ (for critical and subcritical cases), we estimated the B-D-C model parameters (λ, μ and ρ) using the GMM estimator given by equation 4.14 to obtain more consistent and less biased estimates within a fast computational time. Generally, GMM estimation is relatively faster in computational time than MLE. In this study, the GW estimation approach is employed if and only if the offspring mean $m > 1$ within the simulation experiments.

4.2.4 Comparison between the estimation methods

The best estimation methods are identified in this study by exploring the trade-off between estimation accuracy and computational speed. The MLE and GMM estimates were computed with their corresponding 95% confidence interval (C.I) for the three simulation experiments based on 50 independent hosts, respectively (Table 4.1). The GW estimation in this current study, only held for simulation Case 1 (where the true parameter values were set at $\lambda = 0.5$, $\mu = 0.3$ and $\rho = 0.001$) since it was the only instance where the offspring mean m from the Galton-Watson estimation approach > 1 (due to large parasite numbers obtained in simulation Case 1). Thus, GW estimation was not employed for simulation experiments two and three since the offspring mean $m \leq 1$. The GW estimates with their 95% confidence interval for simulation Case 1 are also presented in Table 4.1. The estimation methods' performance was compared by estimating the bias, variance, and mean square error of estimates (based on equations 4.19–4.21) for each parameter of the B-D-C process (Table 4.1). Given $\theta = [\lambda, \mu, \rho]^T$, suppose the true parameter values of the simulated data is $\theta_0 = [\lambda_0, \mu_0, \rho_0]^T$ and its corresponding estimate is $\hat{\theta} = [\hat{\lambda}, \hat{\mu}, \hat{\rho}]^T$; then the bias, variance and mean square error of estimates for each parameter θ_i for $i = 1, 2, 3$ are computed such that

$$bias(\hat{\theta}_i) = E(\hat{\theta}_i) - \theta_{0i} = \frac{1}{100}\sum_{r=1}^{100}\hat{\theta}_{ir} - \theta_{0i}, \tag{4.19}$$

$$Var(\hat{\theta}_i) = E[(\hat{\theta}_i - E(\hat{\theta}_i))^2], \tag{4.20}$$

and

$$MSE(\hat{\theta}_i) = E[(\hat{\theta}_i - \theta_{0i})^2] = Var(\hat{\theta}_i) + [bias(\hat{\theta}_i)]^2. \tag{4.21}$$

For simulation Case 1, the three estimation methods generally performed well based on their bias, variance and mean square error estimates (Table 4.1); however, the bias of MLE was consistently lower and relatively efficient (less variance) in estimating the B-D-C model parameters based on the three simulation experiments. Nevertheless, the Galton-Watson estimation approach was significantly faster in computational speed than MLE and GMM estimation methods (Table 4.4). The computational time required for MLE was significantly higher than the GMM estimation for simulation Cases 1 and 2, where parasite numbers were relatively higher than that of simulation Case 3 (Figure 4.6). This finding is due to the complexity of the B-D-C process's transition and log-likelihood functions. The consistency, efficiency and computational time of GW and GMM estimates were also investigated as the sample size increases for each simulation experiment at sample sizes of 50, 100 and 500 (Tables 4.2–4.4). The GW and GMM estimates were relatively efficient as the sample size increased from 50 to 500. The B-D-C process's mean behaviour was also bounded based on the MLE and GMM for all three *in silico* simulation experiments as goodness-of-fit plots (Figure 4.5). The mean trajectory based on the expected MLE and GMM estimates and actual parameter values perfectly averaged the mean trajectories evaluated at MLEs from the 100 realisations, respectively, across the three simulation cases. Based on the computational speed and accuracy measures of the respective estimation methods, the Galton-Watson estimation approach (given by equations 4.16 and 4.18) and the GMM estimator (given by equation 4.14) can be used together to obtain fast and reasonably accurate estimates of the B-D-C model parameters in refining the modified ABC algorithm for the sophisticated stochastic model. The estimation must be done such that the Galton-Watson method should be used when the offspring mean $m > 1$ and GMM employed when $m \leq 1$.

Table 4.1: Maximum likelihood, Generalised method of moments and Galton-Watson estimates based on the three different *in silico* simulation experiments (Cases 1-3) of 50 hosts respectively.

Simulation	Method	Parameters	$\hat{\theta}$	θ_0	bias($\hat{\theta}$)	Var($\hat{\theta}$)	MSE($\hat{\theta}$)	95% C.I
Case 1	MLE	λ	0.503	0.5	0.003	0.001	0.001	0.496-0.509
		μ	0.303	0.3	0.003	0.001	0.001	0.296-0.309
		ρ	0.001	0.001	7.29×10^{-5}	1.42×10^{-7}	1.48×10^{-7}	0.0009-0.0011
	GMM	λ	0.526	0.5	0.026	0.140	0.141	0.452-0.599
		μ	0.330	0.3	0.029	0.139	0.141	0.256-0.403
		ρ	0.001	0.001	-3.97×10^{-5}	1.28×10^{-7}	1.29×10^{-7}	0.0009-0.0011
	GW	λ	0.507	0.5	0.007	0.020	0.020	0.479-0.534
		μ	0.322	0.3	0.022	0.020	0.020	0.294-0.349
		ρ	0.0073	0.001	0.006	4.81×10^{-6}	4.46×10^{-5}	0.0068-0.0077
Case 2	MLE	λ	1.998	2	-0.002	0.031	0.031	1.964-2.033
		μ	1.009	1	0.001	0.031	0.031	0.975-1.044
		ρ	0.010	0.01	0.000	4.10×10^{-6}	4.13×10^{-6}	0.009-0.011
	GMM	λ	1.879	2	-0.121	0.474	0.489	1.744-2.014
		μ	0.932	1	-0.068	0.206	0.211	0.843-1.021
		ρ	0.010	0.01	2.06×10^{-4}	5.17×10^{-6}	5.21×10^{-6}	0.009-0.011
Case 3	MLE	λ	2.953	3	-0.047	0.222	0.224	2.861-3.045
		μ	1.969	2	-0.031	0.176	0.177	1.887-2.050
		ρ	0.107	0.1	0.007	0.001	0.001	0.101-0.112
	GMM	λ	2.728	3	-0.272	1.299	1.373	2.501-2.952
		μ	1.833	2	-0.167	0.700	0.728	1.669-1.997
		ρ	0.104	0.1	0.004	0.001	0.001	0.099-0.109

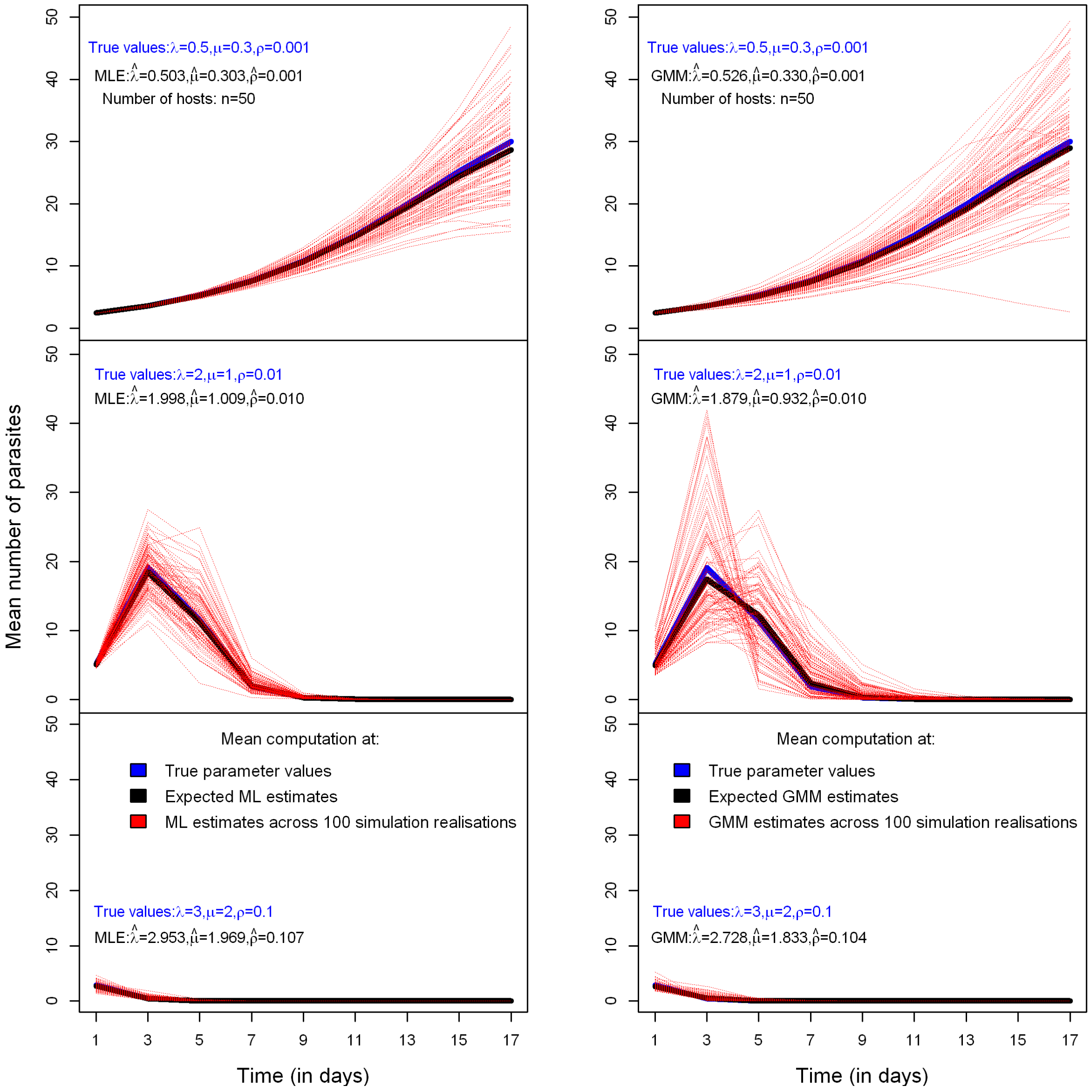

Figure 4.5: Bounded mean behaviour of the B-D-C process based on the MLE and GMM estimates over time across the three different *in silico* simulation experiments ($n = 50$ hosts).

Table 4.2: Effect of sample size on the GMM estimates based on the three different *in silico* simulation experiments (Cases 1-3).

Simulation	Sample size	Parameters	$\hat{\theta}$	θ_0	bias($\hat{\theta}$)	Var($\hat{\theta}$)	MSE($\hat{\theta}$)	95% C.I
Case 1	50	λ	0.526	0.5	0.026	0.140	0.141	0.452-0.599
		μ	0.330	0.3	0.029	0.139	0.141	0.256-0.403
		ρ	0.001	0.001	-3.97×10^{-5}	1.28×10^{-7}	1.29×10^{-7}	0.0009-0.0011
	100	λ	0.494	0.5	-0.006	0.026	0.026	0.462-0.526
		μ	0.294	0.3	-0.006	0.029	0.029	0.261-0.327
		ρ	0.001	0.001	-9.78×10^{-6}	7.39×10^{-6}	7.40×10^{-8}	0.0009-0.0011
	500	λ	0.496	0.5	-0.004	0.003	0.003	0.485-0.507
		μ	0.296	0.3	-0.004	0.003	0.003	0.285-0.307
		ρ	0.001	0.001	2.09×10^{-6}	1.18×10^{-8}	1.18×10^{-8}	0.0009-0.0011
Case 2	50	λ	1.879	2	-0.121	0.474	0.489	1.744-2.014
		μ	0.932	1	-0.068	0.206	0.211	0.843-1.021
		ρ	0.010	0.01	2.06×10^{-4}	5.17×10^{-6}	5.21×10^{-6}	0.009-0.011
	100	λ	1.808	2	-0.192	0.347	0.383	1.693-1.923
		μ	0.884	1	-0.116	0.140	0.154	0.810-0.957
		ρ	0.010	0.01	0.001	3.46×10^{-6}	3.69×10^{-6}	0.009-0.0011
	500	λ	1.927	2	-0.073	0.112	0.117	1.862-1.993
		μ	0.961	1	-0.039	0.046	0.048	0.919-1.003
		ρ	0.010	0.01	3.21×10^{-5}	3.59×10^{-7}	3.60×10^{-7}	0.009-0.011
Case 3	50	λ	2.728	3	-0.272	1.299	1.373	2.501-2.952
		μ	1.833	2	-0.167	0.700	0.728	1.669-1.997
		ρ	0.104	0.1	0.004	0.001	0.001	0.099-0.109
	100	λ	2.579	3	-0.421	0.628	0.806	2.423-2.734
		μ	1.728	2	-0.272	0.294	0.368	1.622-1.834
		ρ	0.100	0.1	-0.001	3.75×10^{-4}	3.76×10^{-4}	0.095-0.102
	500	λ	2.916	3	-0.084	0.243	0.250	2.820-3.013
		μ	1.953	2	-0.047	0.116	0.118	1.889-2.020
		ρ	0.100	0.1	-2.27×10^{-4}	8.06×10^{-5}	8.07×10^{-5}	0.098-0.102

Table 4.3: Effect of sample size on the Galton-Watson estimates (based on simulation Case 1).

Sample size	Parameters	$\hat{\theta}$	θ_0	bias($\hat{\theta}$)	Var($\hat{\theta}$)	MSE($\hat{\theta}$)	95% C.I
50	λ	0.507	0.5	0.007	0.020	0.020	0.479-0.534
	μ	0.322	0.3	0.022	0.020	0.020	0.294-0.349
	ρ	0.0073	0.001	0.006	4.81×10^{-6}	4.46×10^{-5}	0.0068-0.0077
100	λ	0.505	0.5	0.005	0.007	0.007	0.488-0.521
	μ	0.319	0.3	0.019	0.007	0.007	0.302-0.335
	ρ	0.0074	0.001	0.006	2.86×10^{-6}	4.36×10^{-5}	0.0070-0.0077
500	λ	0.528	0.5	0.028	0.001	0.002	0.520-0.535
	μ	0.341	0.3	0.041	0.001	0.003	0.334-0.348
	ρ	0.0067	0.001	0.001	3.58×10^{-7}	3.30×10^{-5}	0.0065-0.0068

Table 4.4: Computational time (in secs) between MLE, GMM and GW estimation based on simulation Case 1 (where true value of $\lambda = 0.5$, $\mu = 0.3$ and $\rho = 0.001$) at different sample sizes (n).

Estimation method	$n = 50$	$n = 100$	$n = 500$
Galton-Watson approach	9.736	19.046	88.531
Generalised method of moments	1086.836	1852.000	6202.532
Maximum likelihood	114584.588	-	-

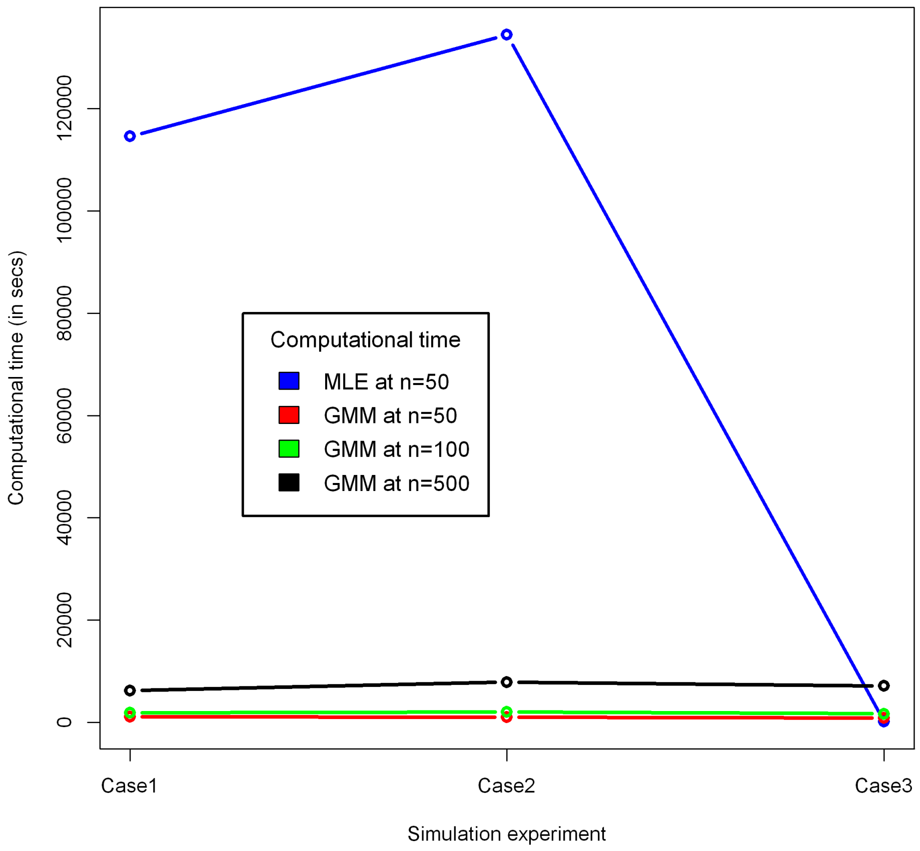

Figure 4.6: Comparison between MLE computational time and that of GMM at different sample sizes across the three simulation experiments (Cases 1-3).

4.3 B-D-C hybrid τ-leaping algorithm

Tau-leaping stochastic simulation is a fast approximate method of the exact stochastic simulation algorithm (SSA) originally proposed by Feller [100]. The $\tau-$leaping algorithm simulates a stochastic system such that all events are carried out during a time step before updating the event or transition rates [110]. The principle behind $\tau-$leaping is analogous to the standard Euler's method for solving deterministic systems (or differential equations). The only difference is that the one-step Euler's method makes use of fixed change in the system's state such that $x(t+\tau) = x(t) + \tau x'(t)$ where $x'(t) = f(t,x)$ and τ is the fixed step size; whereas $\tau-$leaping update the state using $x(t+\tau) = x(t) + P(\tau x'(t))$. Here, $P(\tau x'(t))$ is a Poisson random variable with fixed rate $\tau x'(t)$ and event or transition rate $x'(t)$ to approximate the number of transitions in the time interval $[t, t+\tau)$. However, the accuracy of $\tau-$leaping depends on how well the leap condition is satisfied during a time step, rising to different leaping methods [110, 111]. The leap condition requires a good choice of the leap size τ such that the state change is fixed or does not vary significantly within the time interval $[t, t+\tau)$ to avoid unexpected increment in the transition rates.

The size of τ determines the extent to which the system's history axis is leapt down. The choice of τ or not using τ $(\tau = 0)$ is all about the trade-off between speed and accuracy. When the population size of the system increases, the exact SSA is relatively slower and thus, the leap size $\tau > 0$ is chosen to allow multiple events while satisfying the leap condition during a time step. However, if only a small number of transitions are leapt over, then using the exact SSA is much preferred (i.e., when $\tau = 0$). Different choice of the leap size for accelerating stochastic simulations have been proposed in the literature [193, 249]; but this current study only focuses on leap-size selection methods proposed by Gillespie [110] and Gillespie and Petzold [111], respectively.

To successfully develop a tau-leaping method, we require to find the largest value of the leap size (τ) that satisfies the leap condition. Let $\mathbf{x}$ represent the state of the process $X(t)$, $\theta = [\lambda, \mu, \rho]^T$ denote the parameters of the B-D-C process, $a_j(\mathbf{x}) = \theta_j \mathbf{x}$ for $j = 1, 2, 3$ be the propensity or rate function, and v represent the state-change vector (which takes values -1, 0 or +1); then the leap condition assumes that the value of τ must be small enough such that the change in propensity function or event rates $|a_j(\mathbf{x} + \Lambda(\tau, \mathbf{x})) - a_j(\mathbf{x})|$ is bounded above by a pre-specified error control parameter ϵ $(0 < \epsilon \ll 1)$ of the sum of all event rates [110] such that

$$|a_j(\mathbf{x} + \Lambda(\tau, \mathbf{x})) - a_j(\mathbf{x})| \leq \epsilon a_0(\mathbf{x}), \quad \forall j = 1, 2, 3; \tag{4.22}$$

where $\Lambda(\tau, \mathbf{x}) \approx \sum_{j=1}^{3} P_j(a_j(\mathbf{x}), \tau) v_j$ (with $P_j(a_j(\mathbf{x}), \tau)$ denoting a Poisson random variable with rate $a_j(\mathbf{x})\tau$) and $a_0(\mathbf{x})$ is the total rate. However, since the state $\mathbf{x}$ cannot change by more than 1 within any infinitesimal time interval $[t, t+\tau)$ during the simulation of continuous-time Markov processes, we separately set-up the catastrophe event (where the entire population can hit 0 within the infinitesimal time interval) different from the birth and death events using standard Monte Carlo technique (based on a uniform random number). Hence, the name "B-D-C hybrid τ-leaping algorithm".

4.3.1 Procedure for selecting the Leap size

Two different procedures have been proposed by for selecting the leap size (τ) for processes with N-dimensional states such that the leap condition (equation 4.22) is satisfied [110, 111]. For the purpose of this study, the state $\mathbf{x}$ of the B-D-C process is one-dimensional and thus, the leap size selection stated in Lemma 2 based on τ-selection by Gillespie [110] and Lemma 3 based on a new τ-selection by Gillespie and Petzold [111] are defined for a Markov process with one-dimensional state.

Lemma 2. Let $a_0(\mathbf{x}) = \sum_{j=1}^{M} a_j(\mathbf{x})$, $\xi(\mathbf{x}) = \sum_{j=1}^{M} a_j(\mathbf{x})v_j$, and $b_j = \frac{da_j(\mathbf{x})}{dx}$ for $j = 1, 2, \cdots M$. According to Gillespie [110], a choice for τ satisfying the leap condition (equation 4.22) at a given value of ϵ is

$$\tau = \min_{j \in [1,M]} \left\{ \frac{\epsilon a_0(\mathbf{x})}{|\xi(\mathbf{x}) b_j|} \right\}, \tag{4.23}$$

where v_j is the state-change vector for $j = 1, 2, \cdots M$ and M is the total number of events the process $X(t)$.

Lemma 3. Given the M^2 functions $f_{jj'} = \frac{da_j(\mathbf{x})}{dx} v_{j'}$ for $j, j' = 1, 2, \cdots M$; suppose the $2M$ functions $\delta_j(\mathbf{x})$ and $\sigma_j^2(\mathbf{x})$, are defined such that

$$\delta_j(\mathbf{x}) = \sum_{j'=1}^{M} f_{jj'} a_{j'}(\mathbf{x}) \quad \text{and} \quad \sigma_j^2(\mathbf{x}) = \sum_{j'=1}^{M} f_{jj'}^2 a_{j'}(\mathbf{x}).$$

According to Gillespie and Petzold [111], the largest value for τ which satisfies the leap condition (equation 4.22) at a given value of ϵ is

$$\tau = \min_{j \in [1,M]} \left\{ \frac{\epsilon a_0(\mathbf{x})}{|\delta_j(\mathbf{x})|}, \frac{\epsilon^2 a_0^2(\mathbf{x})}{\sigma_j^2(\mathbf{x})} \right\}. \tag{4.24}$$

Remark. We propose a choice of τ value as presented in Theorems 3 and 4 (proposed for the first time in the current study) based on Lemmas 2 and 3, respectively, for the B-D-C hybrid τ-leaping algorithm by exploring the trade-off between speed and accuracy. Since we have set-up the catastrophe event differently from the birth and death events in the τ-leaping algorithm, the leap sizes (given by equations 4.25 and 4.26) are defined based on only the birth and death events of the B-D-C process. For easy comparison, the B-D-C hybrid τ-leaping algorithms motivated by Gillespie [110] and Gillespie and Petzold [111] studies are named "HTL2001" and "HTL2003", respectively, in subsequent sections (described fully in section 4.3.2).

Theorem 3. *Given the optimal leap size (defined by equation 4.23) and a pre-specified error bound (ϵ), the value of τ for simulating the B-D-C process (HTL2001) with birth rate λ and death rate μ is*

$$\tau_{HTL2001} = \frac{\epsilon(\lambda+\mu)}{|\lambda-\mu|\max(\lambda,\mu)}. \tag{4.25}$$

Proof of Theorem 3 .

Let λ and μ be the birth and death rates of the B-D-C process $X(t) = \mathbf{x}$. Suppose the propensity or even rate functions corresponding to birth and death events are given by: $a_1(\mathbf{x}) = \lambda\mathbf{x}$ and $a_2(\mathbf{x}) = \mu\mathbf{x}$, with the state-change vector $v = [+1, -1]$ and error bound ϵ. From Lemma 2, $a_0(\mathbf{x}) = (\lambda+\mu)\mathbf{x}$; $\xi(\mathbf{x}) = a_1(\mathbf{x}) - a_2(\mathbf{x}) = (\lambda-\mu)\mathbf{x}$; $b_1(\mathbf{x}) = \lambda$ and $b_2(\mathbf{x}) = \mu$.

Then from equation 4.23,

$$\begin{aligned}
\tau_{\text{HTL2001}} &= \min_{j\in[1,2]}\left\{\frac{\epsilon a_0(\mathbf{x})}{|\xi(\mathbf{x})b_j|}\right\} = \min\left\{\frac{\epsilon(\lambda+\mu)\mathbf{x}}{|(\lambda-\mu)\mathbf{x}\lambda|}, \frac{\epsilon(\lambda+\mu)\mathbf{x}}{|(\lambda-\mu)\mathbf{x}\mu|}\right\} \\
&= \min\left\{\frac{\epsilon(\lambda+\mu)}{|(\lambda-\mu)\lambda|}, \frac{\epsilon(\lambda+\mu)}{|(\lambda-\mu)\mu|}\right\} \\
&= \frac{\epsilon(\lambda+\mu)}{|(\lambda-\mu)|\max(\lambda,\mu)} \qquad \text{Q. E. D.}
\end{aligned}$$

as required by equation 4.25.

Theorem 4. *Given the optimal leap size (defined by equation 4.24) and a pre-specified error bound (ϵ), the value of τ for simulating the B-D-C process (HTL2003) with birth rate λ and death rate μ is*

$$\tau_{HTL2003} = \min\left\{ \frac{\epsilon(\lambda+\mu)}{|(\lambda-\mu)|\max(\lambda,\mu)}, \frac{\epsilon^2(\lambda+\mu)^2\mathbf{x}}{(\lambda+\mu)\max(\lambda^2,\mu^2)} \right\}. \tag{4.26}$$

Proof of Theorem 4 .

Let λ and μ be the birth and death rates of the B-D-C process $X(t) = \mathbf{x}$. Suppose the propensity or event rate functions corresponding to birth and death events are given by: $a_1(\mathbf{x}) = \lambda\mathbf{x}$ and $a_2(\mathbf{x}) = \mu\mathbf{x}$, with the state-change vector $v = [+1, -1]$, error bound ϵ and $a_0(\mathbf{x}) = (\lambda+\mu)\mathbf{x}$. Then from Lemma 3,

$$f_{jj'} = \begin{array}{c c} & \begin{array}{cc} 1 & 2 \end{array} \\ \begin{array}{c} 1 \\ 2 \end{array} & \begin{pmatrix} \lambda & -\lambda \\ \mu & -\mu \end{pmatrix} \end{array},$$

$$\begin{aligned}
\delta_1(\mathbf{x}) &= \sum_{j'=1}^{2} f_{1j'}a_{j'}(\mathbf{x}) = f_{11}a_1(\mathbf{x}) + f_{12}a_2(\mathbf{x}) \\
&= \lambda(\lambda\mathbf{x}) + (-\lambda)(\mu\mathbf{x}) = \lambda^2\mathbf{x} - \lambda\mu\mathbf{x} \\
&= (\lambda^2 - \lambda\mu)\mathbf{x}, \\
\delta_2(\mathbf{x}) &= \sum_{j'=1}^{2} f_{2j'}a_{j'}(\mathbf{x}) = f_{21}a_1(\mathbf{x}) + f_{22}a_2(\mathbf{x}) \\
&= \mu(\lambda\mathbf{x}) + (\mu)(\mu\mathbf{x}) = \lambda\mu\mathbf{x} - \mu^2\mathbf{x} \\
&= (\lambda\mu - \mu^2)\mathbf{x}, \\
\sigma_1^2(\mathbf{x}) &= \sum_{j'=1}^{2} f_{1j'}^2 a_{j'}(\mathbf{x}) = f_{11}^2 a_1(\mathbf{x}) + f_{12}^2 a_2(\mathbf{x}) \\
&= \lambda^2(\lambda\mathbf{x}) + \lambda^2(\mu\mathbf{x}) = (\lambda^3 + \lambda^2\mu)\mathbf{x},
\end{aligned}$$

and

$$\begin{aligned}
\sigma_2^2(\mathbf{x}) &= \sum_{j'=1}^{2} f_{2j'}^2 a_{j'}(\mathbf{x}) = f_{21}^2 a_1(\mathbf{x}) + f_{22}^2 a_2(\mathbf{x}) \\
&= \mu^2(\lambda\mathbf{x}) + \mu^2(\mu\mathbf{x}) = (\lambda\mu^2 + \mu^3)\mathbf{x}.
\end{aligned}$$

Then from equation 4.24,

$$
\begin{aligned}
\tau_{\text{HTL2003}} &= \min_{j\in[1,2]}\left\{\frac{\epsilon a_0(\mathbf{x})}{|\delta_j(\mathbf{x})|}, \frac{\epsilon^2 a_0^2(\mathbf{x})}{\sigma_j^2(\mathbf{x})}\right\} = \min\left\{\frac{\epsilon a_0(\mathbf{x})}{|\delta_1(\mathbf{x})|}, \frac{\epsilon^2 a_0^2(\mathbf{x})}{\sigma_1^2(\mathbf{x})}, \frac{\epsilon a_0(\mathbf{x})}{|\delta_2(\mathbf{x})|}, \frac{\epsilon^2 a_0^2(\mathbf{x})}{\sigma_2^2(\mathbf{x})}\right\} \\
&= \min\left\{\frac{\epsilon(\lambda+\mu)\mathbf{x}}{|\lambda^2-\lambda\mu|\mathbf{x}}, \frac{\epsilon^2(\lambda+\mu)^2\mathbf{x}^2}{(\lambda^3+\lambda^2\mu)\mathbf{x}}, \frac{\epsilon(\lambda+\mu)\mathbf{x}}{|\lambda\mu-\mu^2|\mathbf{x}}, \frac{\epsilon^2(\lambda+\mu)^2\mathbf{x}^2}{(\lambda\mu^2+\mu^3)\mathbf{x}}\right\} \\
&= \min\left\{\frac{\epsilon(\lambda+\mu)}{\max(|\lambda^2-\lambda\mu|, |\lambda\mu-\mu^2|)}, \frac{\epsilon^2(\lambda+\mu)^2\mathbf{x}}{\max(\lambda^3+\lambda^2\mu, \lambda\mu^2+\mu^3)}\right\} \\
&= \min\left\{\frac{\epsilon(\lambda+\mu)}{|(\lambda-\mu)|\max(\lambda,\mu)}, \frac{\epsilon^2(\lambda+\mu)^2\mathbf{x}}{(\lambda+\mu)\max(\lambda^2,\mu^2)}\right\} \qquad \text{Q. E. D.}
\end{aligned}
$$

as required by equation 4.26.

4.3.2 Pseudo-codes of the B-D-C hybrid τ-leaping algorithms

In the B-D-C τ-leaping algorithms (HTL2001 and HTL2003, proposed for the B-D-C process for the first time in this study), we forego the leaping method and switch to the exact SSA (Algorithm 1) whenever the leap size τ is very small such that, τ is at most a few multiples of the expected time to the next event in the exact SSA (1/total rate). The pseudo-codes of the τ-leaping algorithms are given by Algorithms 2 and 3, respectively (for R codes, see Appendix E). The main difference between the two B-D-C hybrid τ-leaping algorithms HTL2001 (Algorithm 2) and HTL2003 (Algorithm 3) is the choice of the optimal leap size estimator and its leap condition.

Algorithm 2: B-D-C hybrid τ-leaping algorithm (HTL2001 pseudo-code)

```
   Input: $X$, $\lambda$, $\mu$, $\rho$, $t$, $t_{\text{final}}$ $\epsilon$, and host survival status ($s$).
   Output: Parasite numbers and survival status (alive: $s=1$; dead: $s=2$)
           recorded at discrete times ($t=1,2,\cdots,t_{\text{final}}$).
1  while $t < t_{final}$ and $s = 1$ do
2      Set initial time $t=t_0$, state $X=X_0$ and $s=1$.
3      Calculate rates corresponding to birth ($a_1$), death ($a_2$) and catastrophe ($a_3$);
         such that $a_1 = \lambda X$, $a_2 = \mu X$ and $a_3 = \rho X$.
4      Compute the total rate, $a_0 = \sum_{j=1}^{3} a_j$, for $j = 1,2,3$ (from step 3).
5      Compute the leap size $\tau = \frac{\epsilon(\lambda+\mu)}{|(\lambda-\mu)|\max(\lambda,\mu)}$.
6      if $\tau > \frac{2}{a_0}$ then
7          set $t = t+\tau$ and choose a random number $r$ from $Uniform(0,1)$.
8          if $r < a_3\tau$ then
9              set $X=0$ and $s=2$ (catastrophe event occurs)
10         else
11             set $X = X + Poisson(a_1\tau) - Poisson(a_2\tau)$ (birth and death events
                 occur)
12         end
13     else
14         Execute exact SSA (Algorithm 1)
15     end
16
17     Record $(X,s)$ at the desired discrete times.
18 end
```

Algorithm 3: B-D-C hybrid τ-leaping algorithm (HTL2003 pseudo-code)

```
   Input: $X$, $\lambda$, $\mu$, $\rho$, $t$, $t_{\text{final}}$ $\epsilon$, and host survival status ($s$).
   Output: Parasite numbers and survival status (alive: $s = 1$; dead: $s = 2$)
           recorded at discrete times ($t = 1, 2, \cdots, t_{\text{final}}$).
 1 while $t < t_{final}$ and $s = 1$ do
 2     Set initial time $t = t_0$, state $X = X_0$ and $s = 1$.
 3     Calculate rates corresponding to birth ($a_1$), death ($a_2$) and catastrophe ($a_3$);
         such that $a_1 = \lambda X$, $a_2 = \mu X$ and $a_3 = \rho X$.
 4     Compute the total rate, $a_0 = \sum_{j=1}^{3} a_j$, for $j = 1, 2, 3$ (from step 3).
 5     Compute the leap size $\tau = \min\left\{\frac{\epsilon(\lambda+\mu)}{|(\lambda-\mu)|\max(\lambda,\mu)}, \frac{\epsilon^2(\lambda+\mu)^2\mathbf{x}}{(\lambda+\mu)\max(\lambda^2,\mu^2)}\right\}$.
 6     if $\tau > \frac{1}{10a_0}$ then
 7         set $t = t + \tau$ and choose a random number $r$ from $Uniform(0, 1)$.
 8         if $r < a_3\tau$ then
 9             set $X = 0$ and $s = 2$ (catastrophe event occurs)
10         else
11             set $X = X + Poisson(a_1\tau) - Poisson(a_2\tau)$ (birth and death events
                 occur)
12         end
13     else
14         Execute exact SSA (Algorithm 1)
15     end
16
17     Record $(X, s)$ at the desired discrete times.
18 end
```

4.3.3 Effects of the error bound on the accuracy and speed of the τ-leaping algorithms

We compared the two B-D-C hybrid τ-leaping algorithms by exploring a balance between simulation accuracy and computational speed at 15 different error bound values (for $0 \leq \epsilon \leq 0.1$) based on the three different *in silico* simulation experiments (Cases 1-3) in a full factorial design. For each simulation experiment, 3 million simulations were conducted. We quantified the simulation accuracy of the two τ-leaping algorithms (HTL2001 and HTL2003) by estimating the squared error loss between the true mean number of parasites and its corresponding predictions across all observed time points at each error bound (ϵ =0, 0.002, 0.004, 0.006, 0.008, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.09 and 0.1).

We also computed total standard error of the population mean (total deviation in the sampling distribution) and computational time (speed) at each ϵ value for comparison. We proposed an error threshold for the two τ-leaping algorithms based on the simulation speed and accuracy for the three simulation cases.

4.3.3.1 Relationship between the leap size, the state and the leap conditions

The relationship between the leap size, the state and the leap conditions were explored for the two tau-leaping algorithms and across the three different *in silico* experiments (Figure 4.7). It was consistently shown across the different simulation Cases that as the state variable x increases, the leap size of algorithm HTL2003 increases and converges towards the leap size of algorithm HTL2001. However, the leap condition decreases if the state value increases. Tau-leaping started in algorithm HTL2001 at state $x > 40$ for simulation Case 1, state $x > 50$ for simulation Case 2 and state $x > 30$ for simulation Case 3. For algorithm HTL2003, leaping started at state $x > 20$ for simulation Case 1, state $x > 30$ for simulation Case 2 and state $x > 20$ for simulation Case 3. Consequently, starting the tau-leaping algorithms earlier after state $x \geq 10$ or setting $\tau = 0$ for a small value of x, say, could improve the computational speed of the $\tau-$leaping algorithms (HTL2001 and HTL2003), but may have some cost of simulation accuracy.

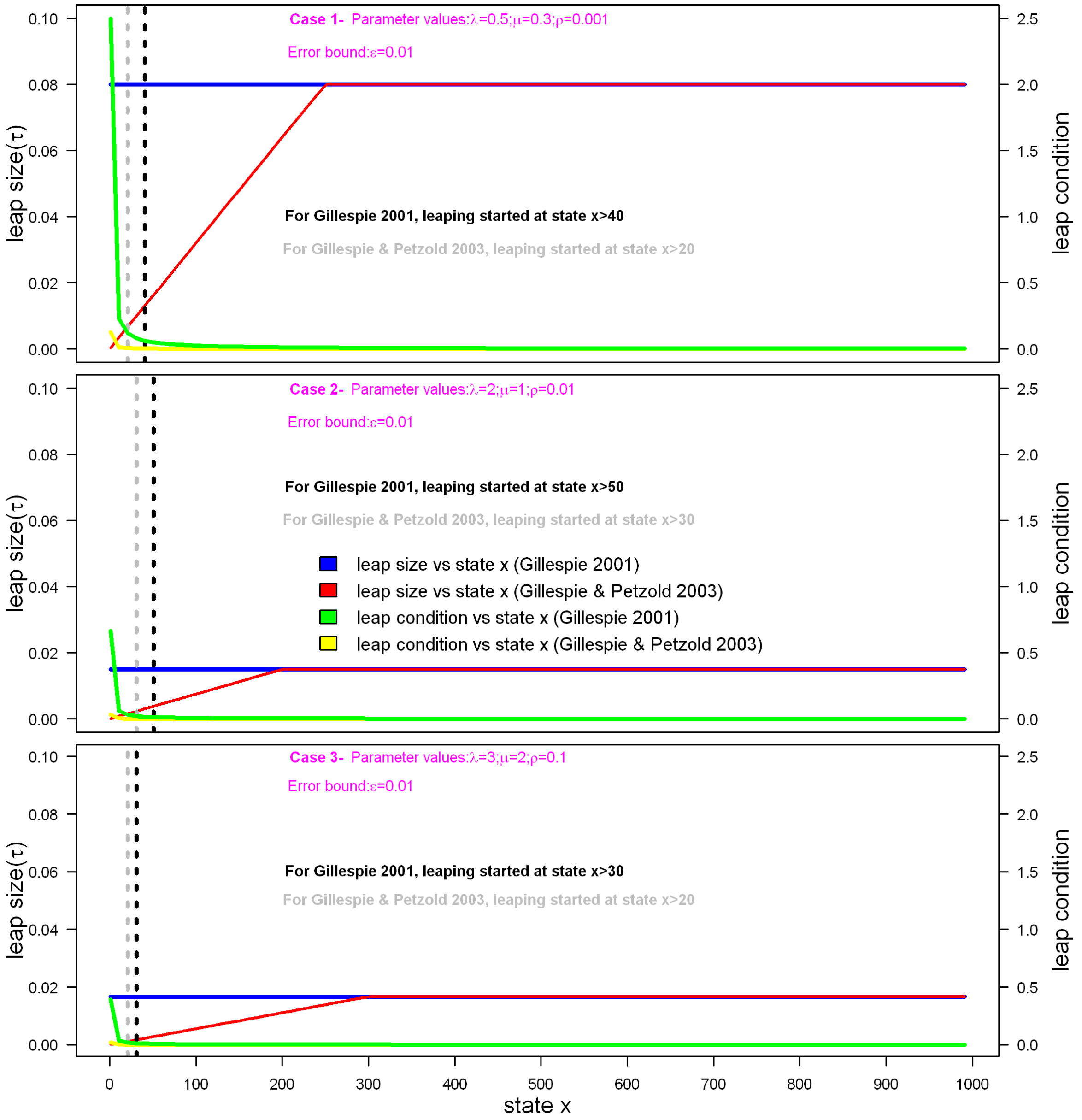

Figure 4.7: Relationship between the leap size, the state variable and the leap conditions.

4.3.3.2 Mean comparison at different error values

Figures 4.8–4.10 present the Monte Carlo estimates of the mean parasite numbers between the two B-D-C hybrid τ-leaping algorithms for $\epsilon \leq 0.01$ and their corresponding true mean values over time across the three simulation experiments based on 3 million simulations, respectively. Although 95% confidence intervals were obtained for the estimates, the interval width was very small in size due to small standard errors associated with the large number of simulations. Hence, the 95% confidence intervals are not presented here; nonetheless, the true mean was found in the estimated intervals over 90% of the time. The true mean and the Monte Carlo mean estimates were consistent across the observed time points (t=1-30 days) for both τ-leaping algorithms at $\epsilon \leq 0.01$. The τ-leaping methods were fairly accurate for small error values ($0 \leq \epsilon \leq 0.01$) compared to error values > 0.01 .

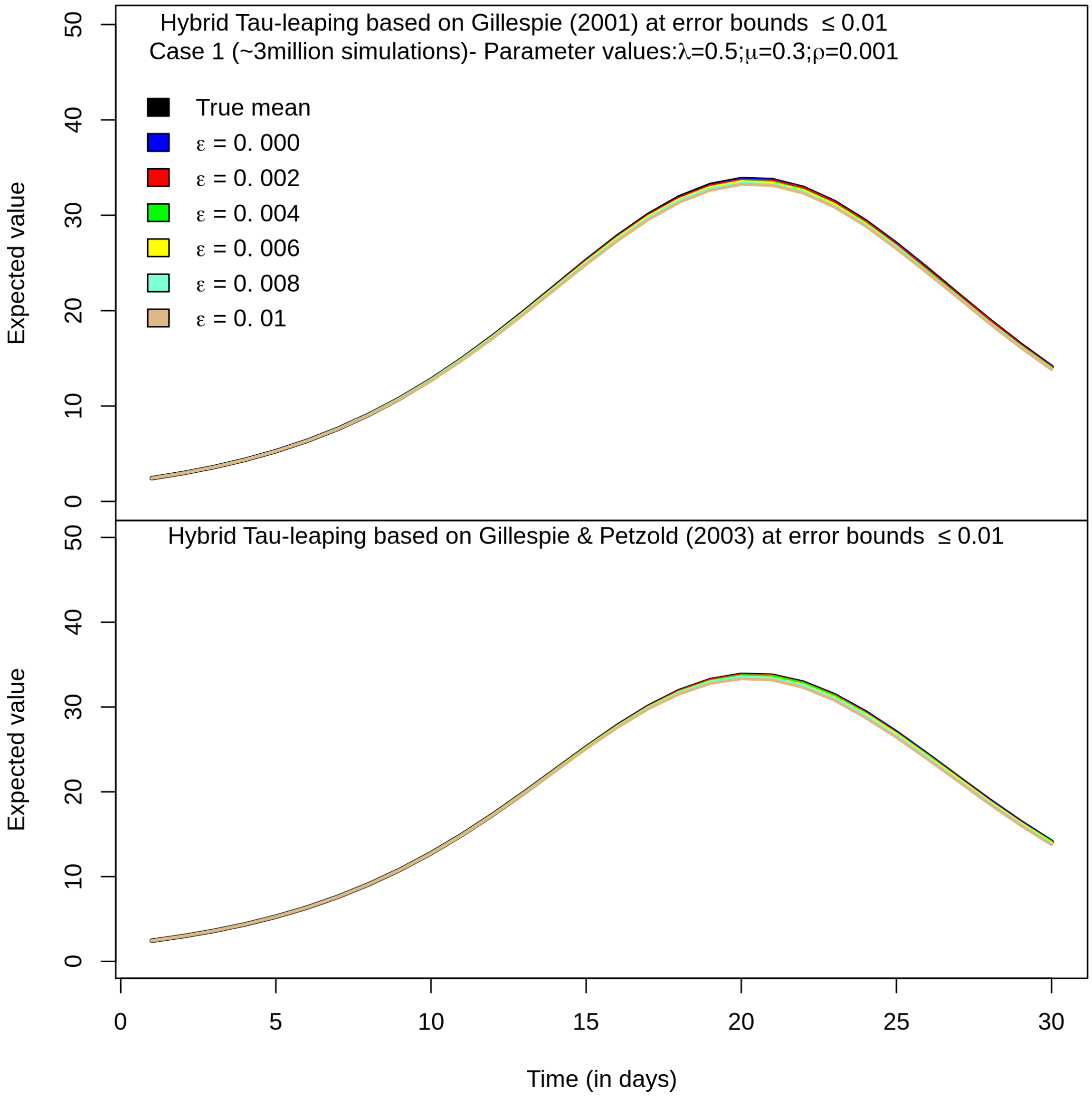

Figure 4.8: Comparing the mean behaviour of parasites between the two B-D-C hybrid τ-leaping algorithms at given parameter values ($\lambda = 0.5$, $\mu = 0.3$, $\rho = 0.001$) at $\epsilon \leq 0.01$ (Case 1).

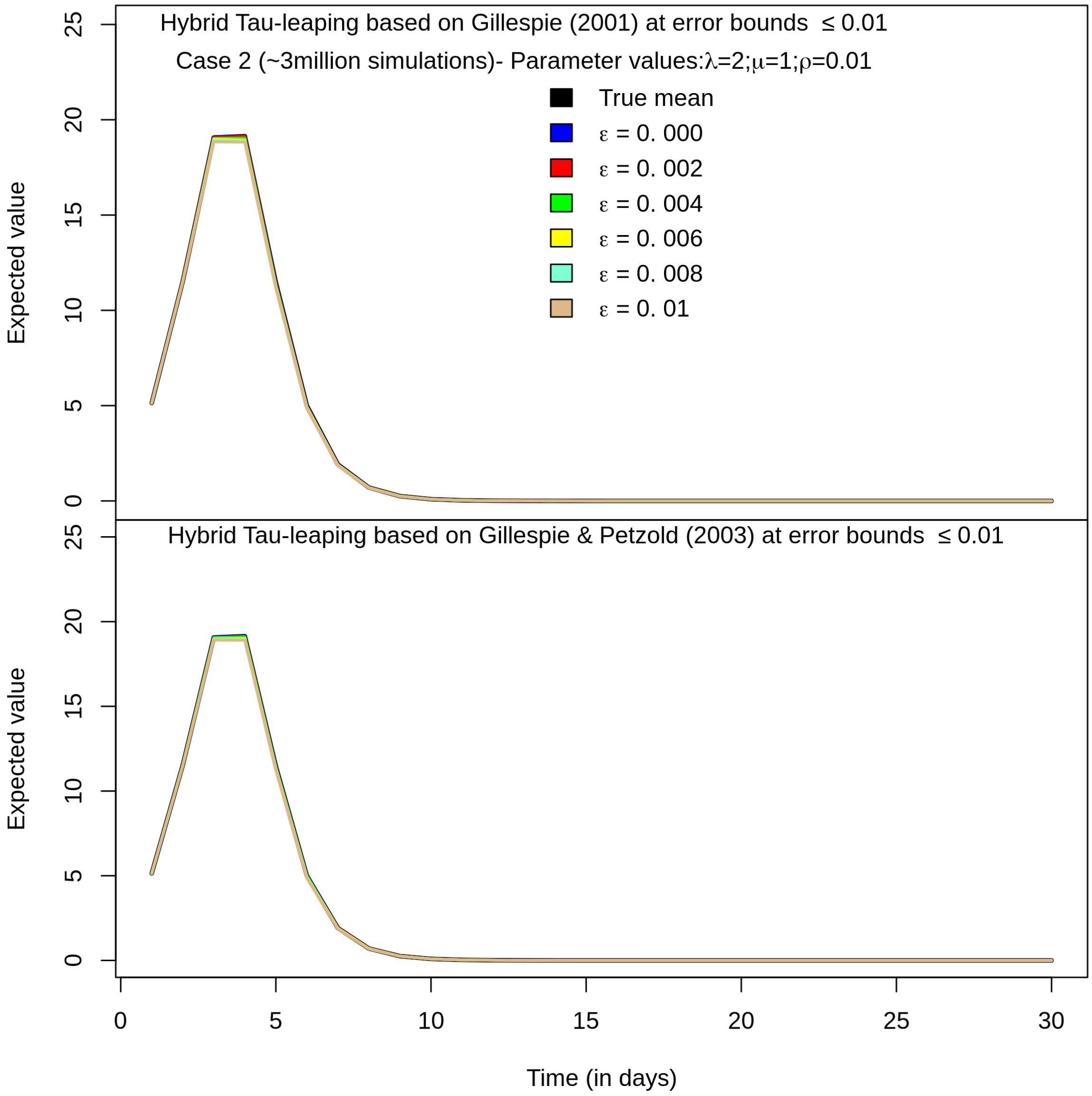

Figure 4.9: Comparing the mean behaviour of parasites between the two B-D-C hybrid τ-leaping algorithms at given parameter values ($\lambda = 2$, $\mu = 1$, $\rho = 0.01$) at $\epsilon \leq 0.01$ (Case 2).

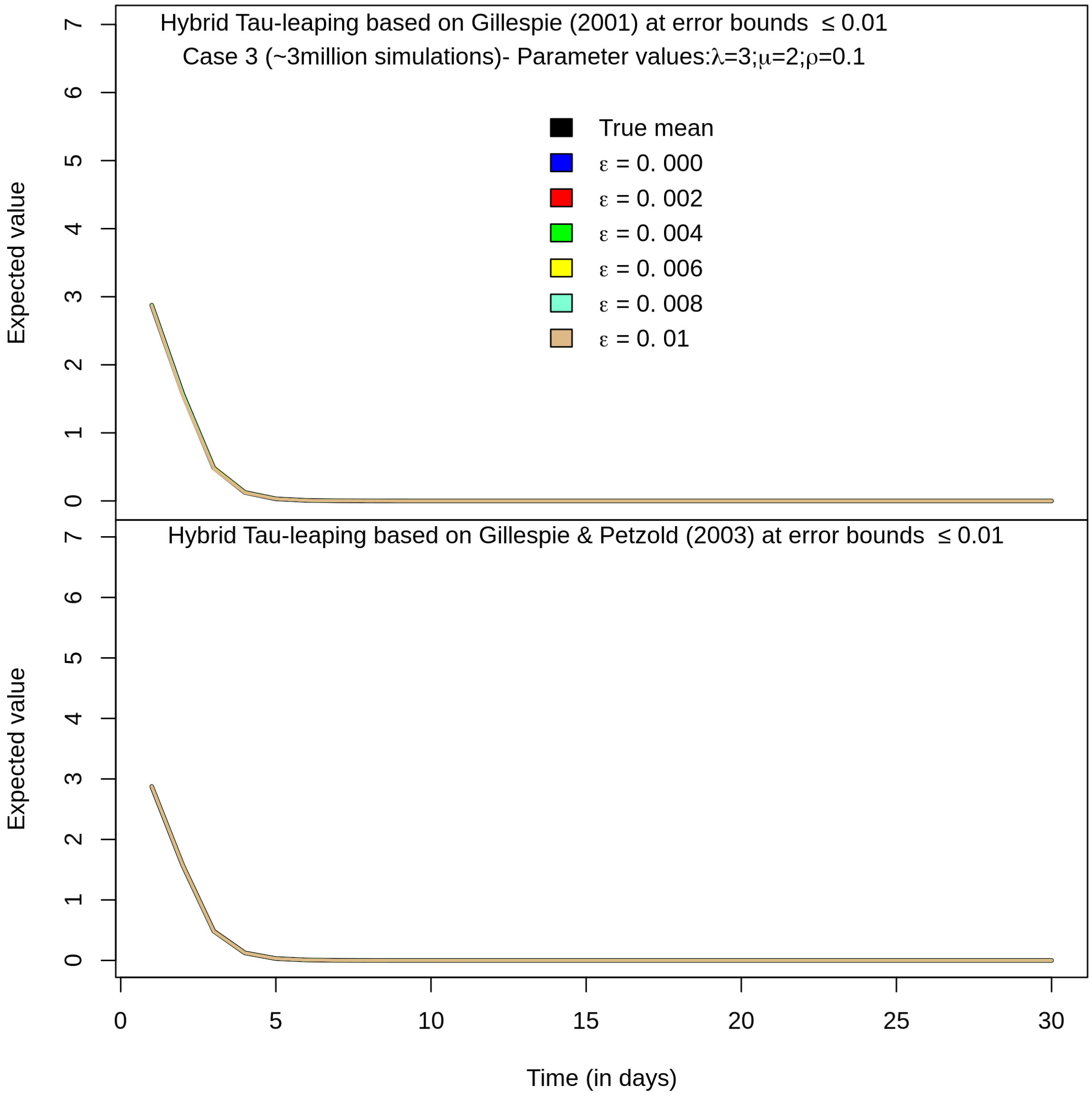

Figure 4.10: Comparing the mean behaviour of parasites between the two B-D-C hybrid τ-leaping algorithms at given parameter values ($\lambda = 3$, $\mu = 2$, $\rho = 0.1$) at $\epsilon \leq 0.01$ (Case 3).

4.3.3.3 Simulation accuracy and deviations in sampling distribution at different error values

We examined the simulation accuracy by estimating the squared error loss and the total standard error of the population mean at each value of ϵ across the three simulation experiments (Figures 4.11–4.13). The simulation accuracy decreased with increasing value of the error bound, while the small deviation in the sampling distribution of the simulated data was achieved for larger values of ϵ. In the first and second simulation experiments (where parasite numbers were relatively higher), there was no significant difference in the error loss and standard error estimates between the two τ-leaping algorithms across the different error bounds. However, simulation Case 3 (where parasite numbers were relatively low) revealed a significant difference in simulation accuracy and standard error at $\epsilon > 0.01$; the HTL2003 algorithm performed better than the HTL2001 algorithm, but the sampling variations from the HTL2003 algorithm was higher comparatively.

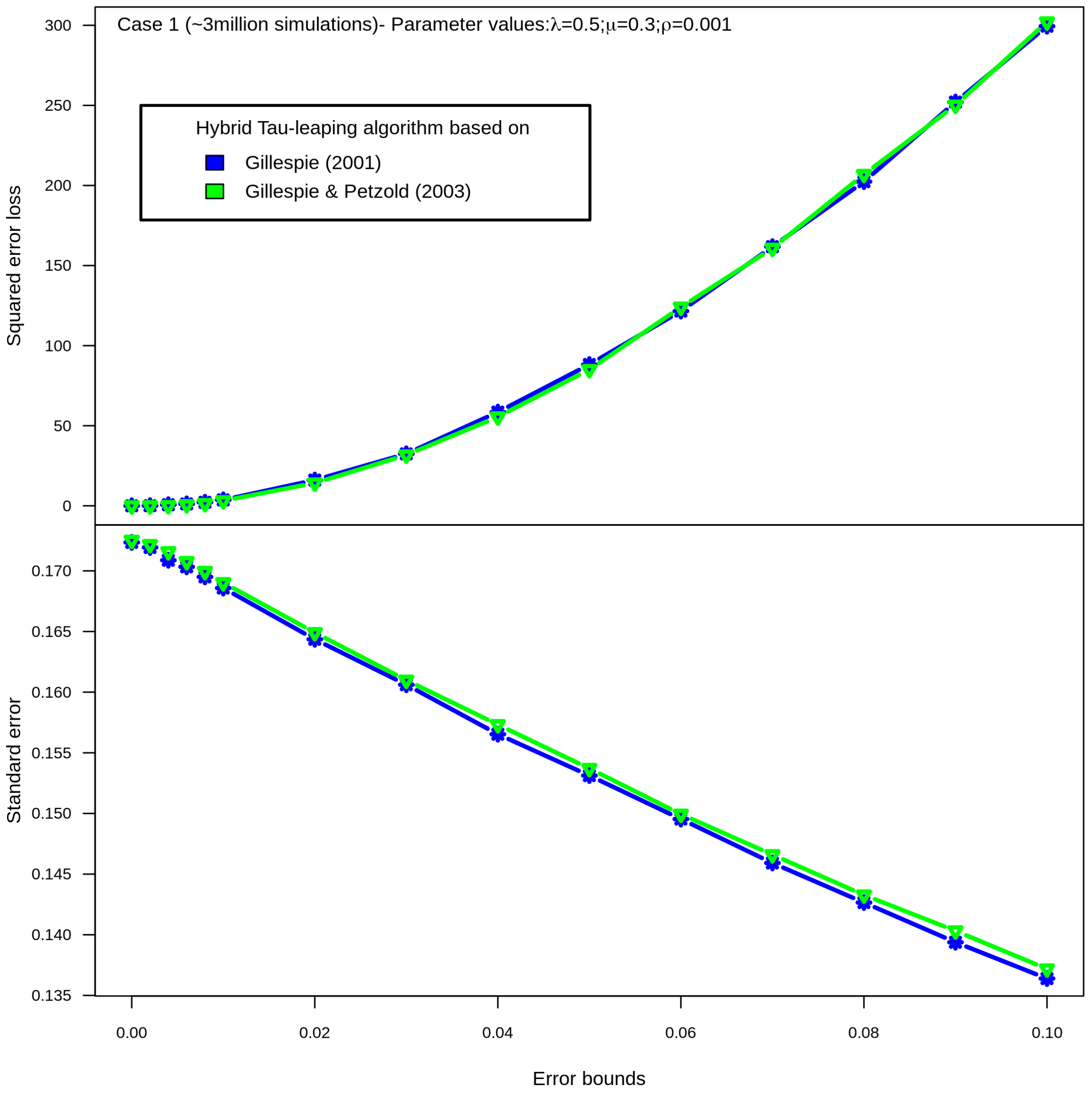

Figure 4.11: Comparing the simulation accuracy between the two B-D-C hybrid τ-leaping algorithms at given parameter values ($\lambda = 0.5$, $\mu = 0.3$, $\rho = 0.001$) across the error bound values (Case 1).

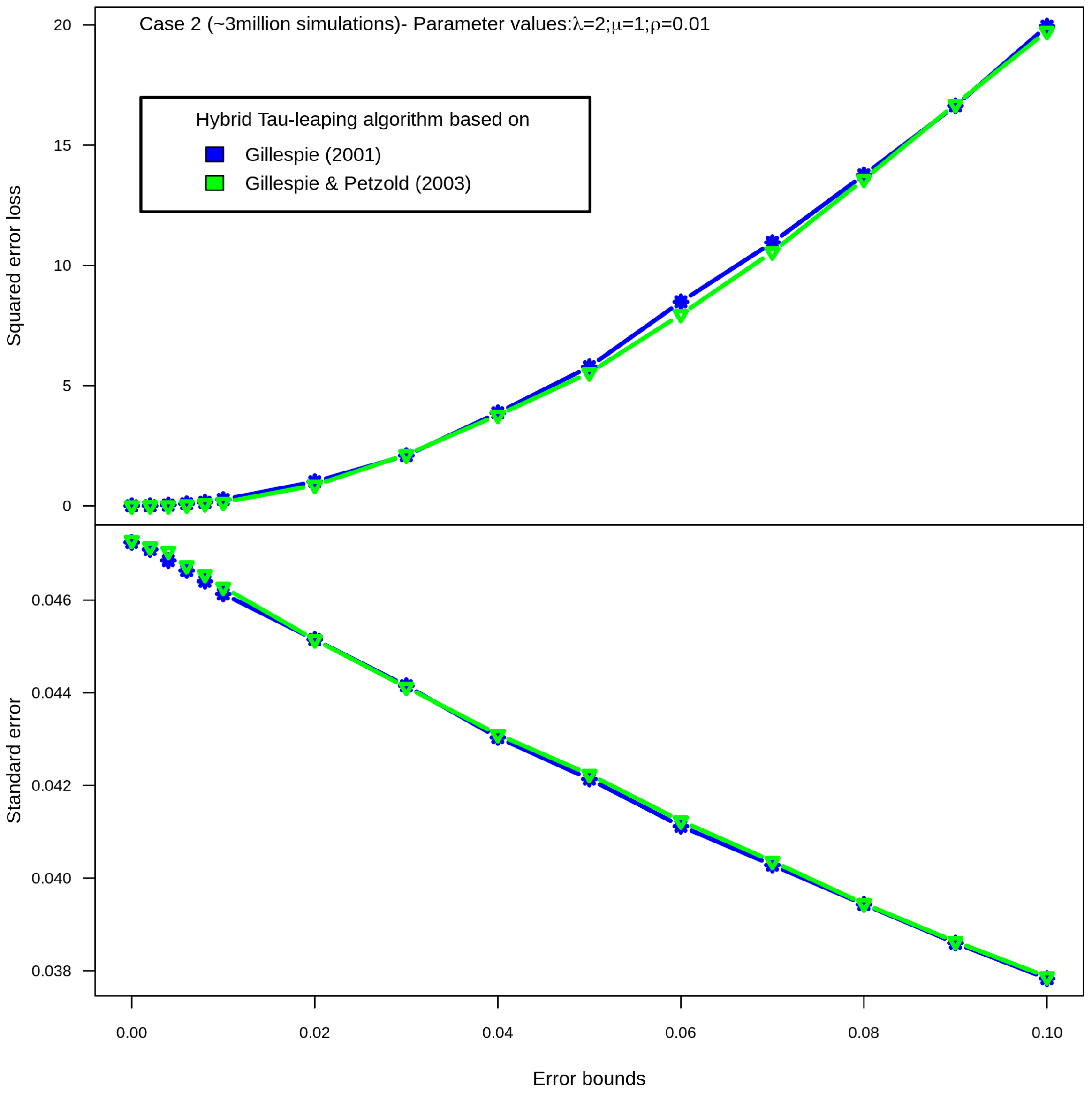

Figure 4.12: Comparing the simulation accuracy between the two B-D-C hybrid τ-leaping algorithms at given parameter values ($\lambda = 2$, $\mu = 1$, $\rho = 0.01$) across the error bound values (Case 2).

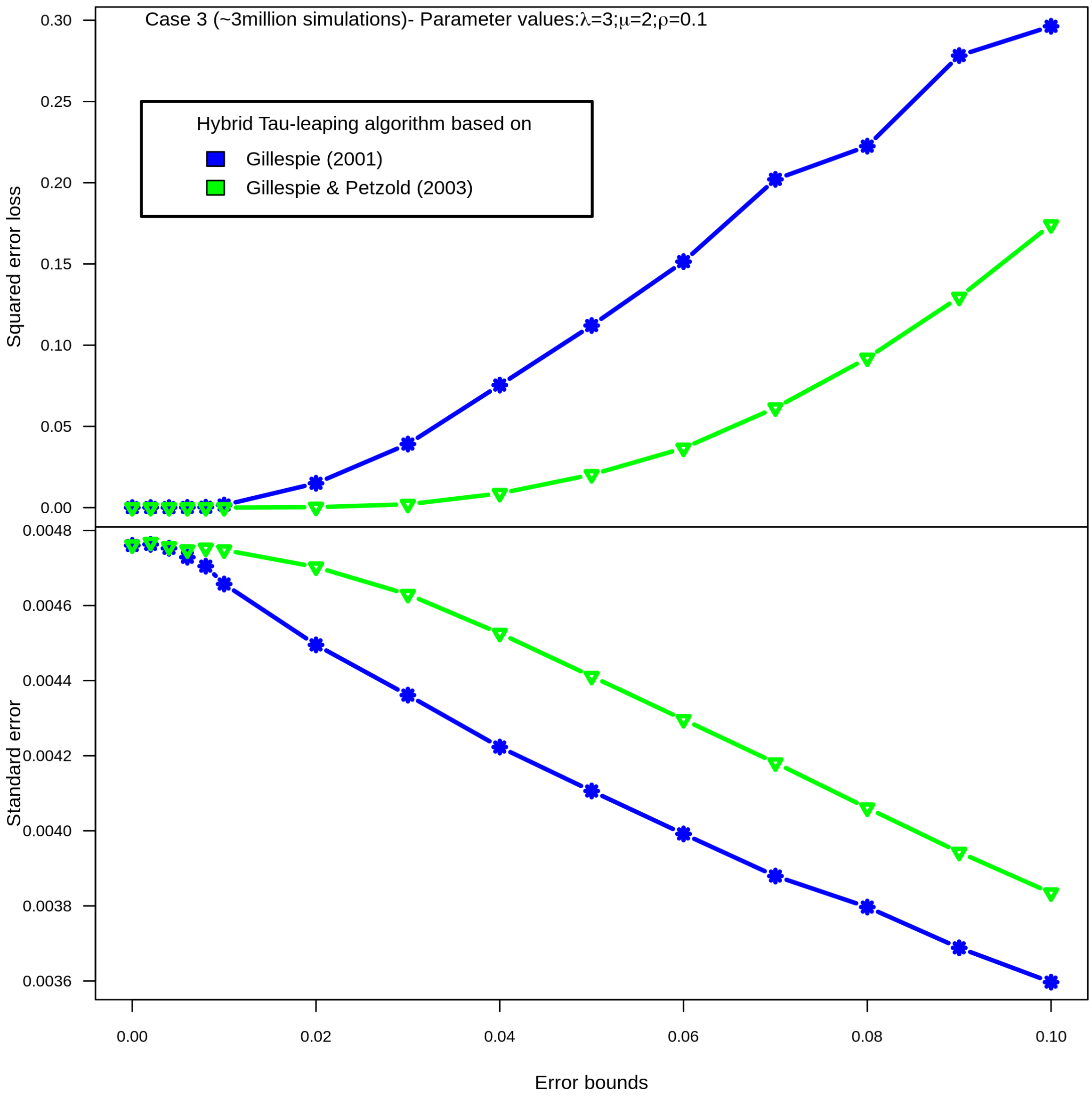

Figure 4.13: Comparing the simulation accuracy between the two B-D-C hybrid τ-leaping algorithms at given parameter values ($\lambda = 3$, $\mu = 2$, $\rho = 0.1$) across the error bound values (Case 3).

4.3.3.4 Computational speed versus simulation accuracy across different error values

We proposed a good choice of the error bound (based on the 15 different ϵ values) by exploring the trade-off between computational speed and simulation accuracy across the three simulation experiments for both τ-leaping algorithms (Figures 4.14–4.16). It can be observed from Figures 4.14–4.16 that $\epsilon = 0.01$ can be a good error bound choice for both $\tau-$leaping algorithms so as to achieve accurate simulations with relatively low computational time. Nonetheless, HTL2001 was much faster (with small computational time) than HTL2003 at any each ϵ value ($0 \leq \epsilon \leq 0.1$) across all the simulation experiments. From Figure 4.7, the computational speed of the algorithms can be improved further.

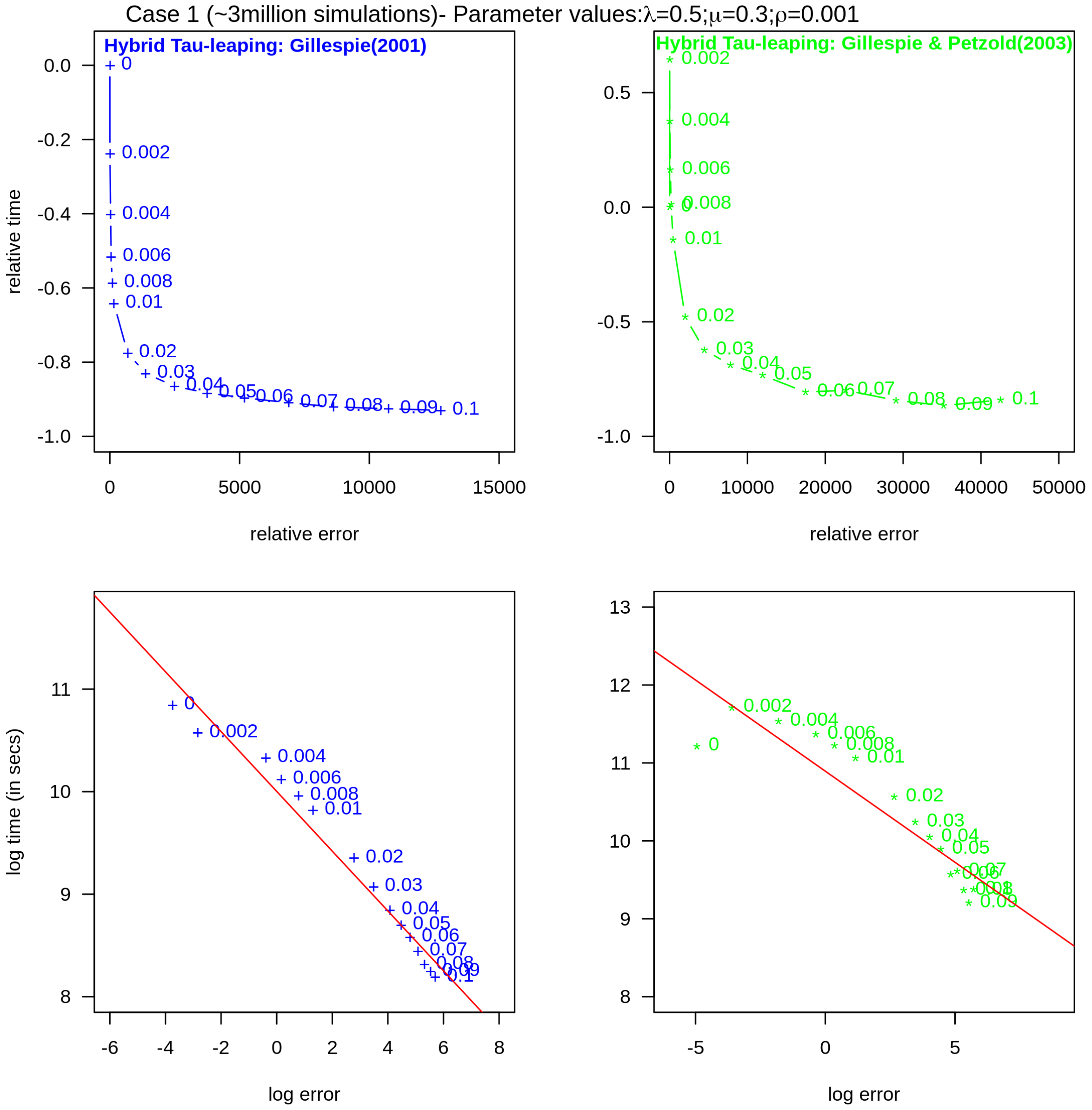

Figure 4.14: Plot of computational speed against simulation accuracy between the two B-D-C hybrid τ-leaping algorithms across the error bound values (Case 1).

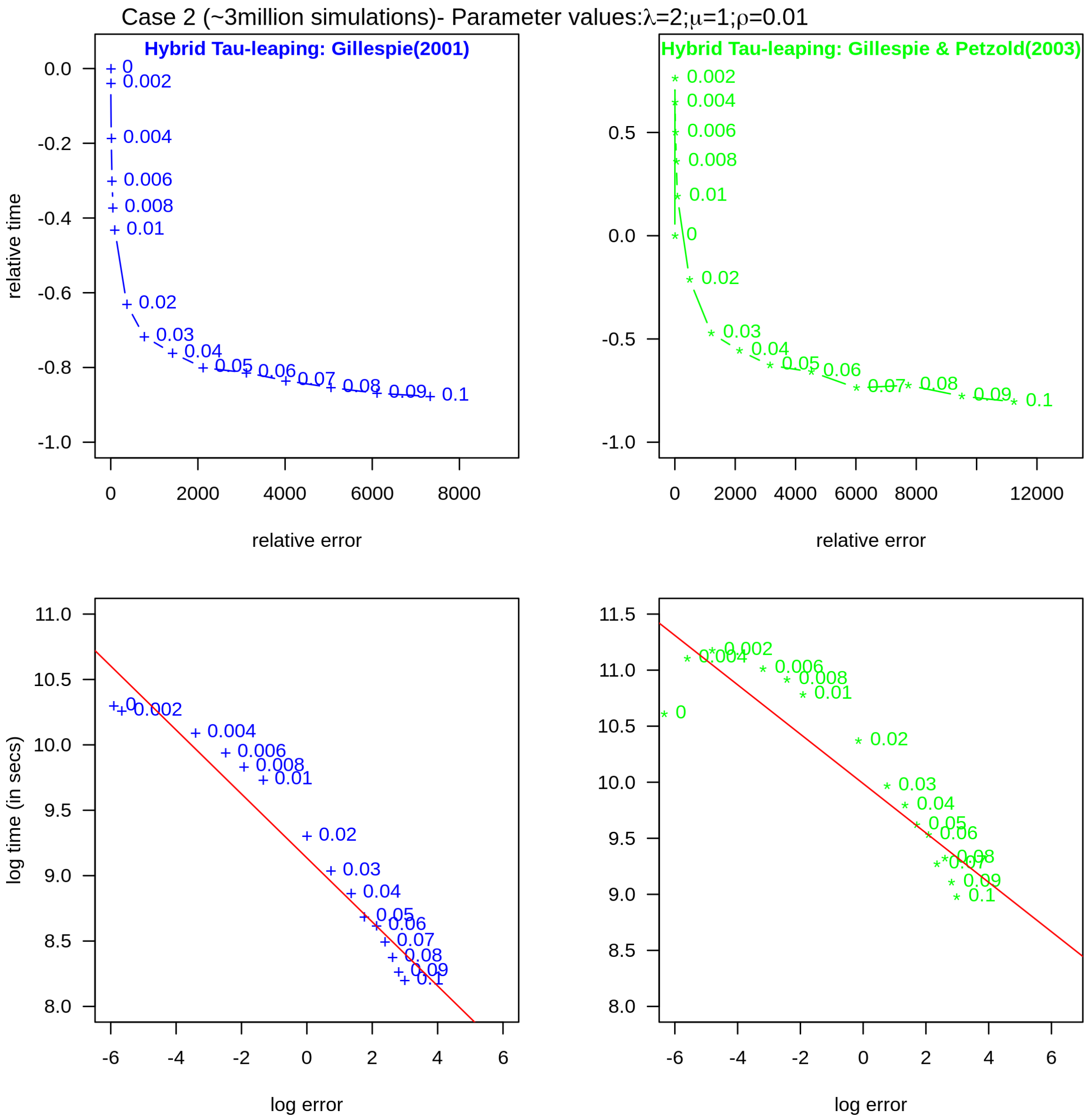

Figure 4.15: Plot of computational speed against simulation accuracy between the two B-D-C hybrid τ-leaping algorithms across the error bound values (Case 2).

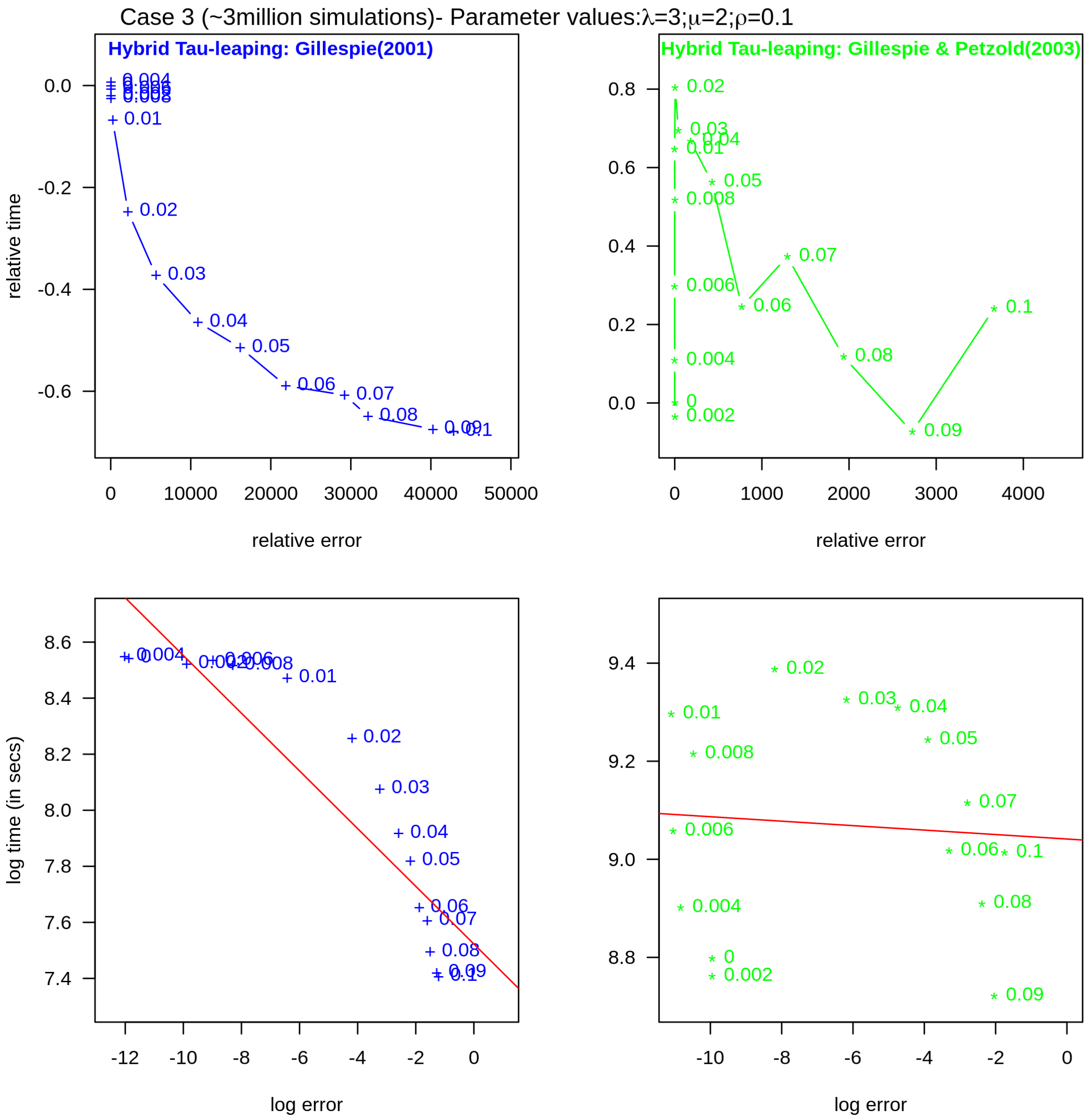

Figure 4.16: Plot of computational speed against simulation accuracy between the two B-D-C hybrid τ-leaping algorithms across the error bound values (Case 3).

Chapter 5

ABC for model calibration

5.1 Introduction

This thesis chapter first briefly reviews approximate Bayesian computation (ABC) algorithms for model calibration (section 5.2). The goal of section 5.2 is to summarise a few ABC algorithms and provide basic intuition behind some existing theoretical results (with relevant references for more formal definitions and theoretical works in ABC provided). Section 5.2.1 gives a more general overview of ABC, whereas section 5.2.2 summarises various levels of approximation in ABC. In section 5.2.3, we provide a short description of a few efficient ABC algorithms in the literature, such as Markov chain Monte Carlo ABC (ABC-MCMC), sequential Monte Carlo ABC (ABC-SMC) samplers, and existing regression-adjusted ABC methods.

Secondly, we present a modified ABC algorithm dubbed “weighted-iterative ABC” (based on sequential Monte Carlo, adaptive importance sampling, and a modified summary statistics weighting under section 5.3). Additionally, we propose a modified methodology for ABC post-processing or posterior mean adjustments (for multi-parameter models with a set of high-dimensional correlating summary statistics) using a weighted ridge regression estimator. The weighted-iterative ABC coupled with the extended local-linear regression (with $L2$ penalty term) is used to calibrate our novel stochastic simulation model (presented in Chapter 6).

Moreover, we propose an optimised linear regression function to project parasite numbers after host mortality till the end of the observation period to aid in computing summary statistics for ABC fitting (section 5.3.3). The proposed ABC post-processing regression (in section 5.3.4) is an extension of Beaumont et al. [27] local-linear regression with heteroscedastic errors and an $L2$ regularisation. With the help of a numerical experiment (where the true posterior is known), the fidelity of our proposed weighted-iterative ABC algorithm and the modified ABC post-processing regression with $L2$ regularisation are assessed, and the key findings are summarised in section 5.3.5.

5.2 Literature review of existing ABC algorithms

5.2.1 Overview of ABC

Estimating model parameters and accounting for uncertainty in both the parameters and model predictions when using mathematical models to investigate biological or physical events is essential. Nonetheless, due to the inability to compute the likelihood or explicitly define its exact form for a large class of mechanistic models, maximum likelihood and classical Bayesian estimation methods are challenging to implement. Consequently, a likelihood-free parameter estimation technique known as approximate Bayesian computation (ABC) has been proposed in the literature to overcome such estimation difficulties in the Bayesian setting. Explicitly, ABC is a collection of likelihood-free methods developed for implementing Bayesian analysis when the likelihood function $L(\theta) = f(y \mid \theta)$ [70] of a generative model $\mathcal{M}$ is either mathematically intractable or computationally expensive, by obtaining an approximation to the true posterior distribution, $p(\theta \mid y_{\text{obs}})$, given the observed data $y_{\text{obs}} \in \mathcal{Y}$ and prior beliefs of the underlying model parameter $\theta \in \Theta$ expressed through the prior distribution $\pi(\theta)$ [279, 282].

ABC has vast applications across different areas of science such as rainfall simulation [11], astronomy [66], vaccine assessment across populations [69], ecology [73], epidemic modelling [215], epidemiology [227], model selection in dynamical systems [295], and archaeology [319], amongst others. The idea of the ABC methodology was initially introduced by Rubin [264], but Tavaré et al. [291] pioneered a rejection-sampling method for posterior inference. The first major extension of the rejection-sampling method led to the development of the basic ABC rejection algorithm. The latter was generalised by Pritchard et al. [245] based on an approximation of the target (posterior) to arguably produce the first genuine basic ABC rejection algorithm. Extensions of the basic ABC rejection algorithm have also resulted in different but more improved versions concerning computational efficiency, convergence to the true posterior distribution, and its applicability. Beaumont et al. [27] also established ABC by further extending the methodology and examining its appropriateness for population genetics problems.

Mathematically, the posterior distribution $p(\theta \mid y_{\text{obs}}) \propto f(y_{\text{obs}} \mid \theta)\pi(\theta)$ contains all necessary information for predictive inference, decision making, models' goodness-of-fit and comparison; and $p(\theta \mid y_{\text{obs}})$ is simply derived using the Bayes' theorem such that

$$p(\theta \mid y_{\text{obs}}) = \frac{f(\theta, y_{\text{obs}})}{f(y_{\text{obs}})} = \frac{f(y_{\text{obs}} \mid \theta)\pi(\theta)}{\int\limits_{\Theta} f(y_{\text{obs}} \mid \theta)\pi(\theta)d\theta}. \tag{5.1}$$

During instances where the likelihood $f(y_{\text{obs}} \mid \theta)$ is challenging to evaluate, we forgo this stage and focus on our ability to simulate data from the model for ABC inference. Rather than explicitly evaluating $f(y_{\text{obs}} \mid \theta)$, ABC-based techniques employ systematic comparisons between observed and simulated data to approximate the true (but unachievable) posterior distribution $p(\theta \mid y_{\text{obs}})$. This entails defining an approximation to $p(\theta \mid y_{\text{obs}})$ so that only the capacity to sample from the model $f(\cdot \mid \theta)$ can provide a means to sample from this approximate posterior. To numerically evaluate the necessary integrals (possibly multidimensional integrals) in an ABC framework based on the prior distribution $\pi(\theta)$ and data from a model $f(\cdot \mid \theta)$, Monte Carlo numerical integration techniques (with or without importance sampling depending on the choice of ABC sampler) for Lebesgue-integrable functions [58] are commonly adopted by invoking the law of large numbers.

The idea behind ABC is to summarise the data first using low-dimensional summary statistics like sample means, autocovariances, or appropriate data quantiles (amongst others) to ease the comparison between high-dimensional simulated and observed data [190]. However, other ABC methods which directly compare data exist by adopting the

Kullback-Leibler divergence [163], the Wasserstein distance [31], or the energy statistic [224] (but they can be difficult to apply to network data and other high-dimensional data). If the dimension of summary statistics is too high, it can distort the posterior approximation due to a prohibitively low acceptance rate; whereas, if too few summaries are considered, it can result in data information loss [97, 243]. As a result, the balance between the low-dimension and informativeness of ABC summary statistics has remained a significant concern in the ABC framework. The general issue of the curse of dimensionality and selection strategies of ABC summary statistics (in the literature) are briefly discussed in section 5.2.2.2. The notion of sufficient summary statistics in the ABC setting appears to be an ideal choice in ABC but is frequently unavailable [243]. Thus, strategies for selecting insufficient low-dimensional summaries are required, and attempts have been made in previous studies [see works by 84, 99, 253, 260]. The quality of the posterior approximation depends on the choice of summary statistics, kernel function K_h (with bandwidth $h > 0$) or tolerance $\epsilon > 0$ (given a distance metric without the use of a kernel function), and the Monte Carlo sampler being implementing [190, 282]. It is important to note that some ABC samplers rely on a kernel function as a similarity metric (e.g., Algorithm L1 [282]) while others (such as the basic rejection ABC summarised below) do not in assessing the discrepancy between simulation and observed data. According to Beaumont et al. [27], the latter particularly corresponds to the use of a uniform kernel. In the case of implementing importance sampling in ABC (where particles are weighted), a proposal or importance density function is required (instead of repeatedly sampling particles from the prior).

The basic ABC rejection algorithm can be described in at most four steps. Let $\theta \in \mathbb{R}^n$, an n-dimensional real parameter vector, be the model parameters to be estimated; then the basic ABC rejection algorithm as follows:

Step 1. Draw sample parameter θ^* or particle from the prior distribution, $\theta^* \sim \pi(\theta)$.

Step 2. Simulate data y_{sim} from the forward or generative model, $y_{\text{sim}} \sim f(\cdot \mid \theta^*)$.

Step 3. Compare the discrepancy between the simulated data y^* and the observed data y_{obs}, and accept particle θ^* as a sample from the posterior distribution $p_\epsilon(\theta \mid y_{\text{obs}})$ defined by equation 5.2, if the simulated data y_{sim} is close to the observed data y_{obs} using a distance metric ρ (e.g., Euclidean distance) such that $\rho\left(S(y_{\text{sim}}), S(y_{\text{obs}})\right) \leq \epsilon$, where $S(\cdot) \in \mathbb{R}^m$ is a summary statistics of the data (possibly m-dimensional) and the sufficiently small $\epsilon > 0$ is a pre-specified tolerance level (measuring the proportion of accepted particles); else reject θ^*.

Step 4. Finally repeat steps 1-3 till the desired particle size (N) from the prior distribution is achieved.

The rejection-based ABC algorithm outlined above (from steps 1-4) leads to sampling from an approximate posterior $p_\epsilon(\theta \mid s_{\text{obs}})$ of the form,

$$p_\epsilon(\theta \mid s_{\text{obs}}) \propto \pi(\theta) \int f(s_{\text{sim}} \mid \theta) \mathbb{1}\left[\rho\left(s_{y_{\text{sim}}}, s_{y_{\text{obs}}}\right) \leq \epsilon\right] ds_{y_{\text{sim}}} \tag{5.2}$$

where $s_{y_{\text{sim}}} = S(y_{\text{sim}})$, $s_{y_{\text{obs}}} = S(y_{\text{obs}})$, and $y_{\text{sim}} \sim f(\cdot \mid \theta)$. We assume that $s_{y_{\text{sim}}} \approx s_{y_{\text{obs}}}$ implies $y_{\text{sim}} \approx y_{\text{obs}}$ (with a probability of one). From equation 5.2, let suppose $L(s_{y_{\text{obs}}} \mid \theta, \epsilon) = \int f(s_{\text{sim}} \mid \theta) \mathbb{1}\left[\rho\left(s_{y_{\text{sim}}}, s_{y_{\text{obs}}}\right) \leq \epsilon\right] ds_{y_{\text{sim}}}$. Then, ABC in this case can be thought of as a regular Bayesian analysis with an approximation to the likelihood defined by $L(s_{y_{\text{obs}}} \mid \theta, \epsilon)$. When $\epsilon \to 0$, we would expect equation 5.2 to converge to the posterior given s_{obs} for large value of the Monte Carlo sample size, N [99].

The indicator function, $\mathbb{1}\left[\rho\left(s_{y_{\text{sim}}}, s_{y_{\text{obs}}}\right) \leq \epsilon\right]$, in the classical rejection-based ABC samplers can lead to wasteful information loss since the algorithm does not distinguish between samples of θ which produce simulated data y_{sim} very close to the observed data y_{obs} and samples of θ for which the associated simulated data is the farthest away from y_{obs} [282]. Thus, a more continuous scale from 1 (i.e., when $y_{\text{sim}} \approx y_{\text{obs}}$) to 0 (when $\|y_{\text{sim}} - y_{\text{obs}}\| \approx \|s_{\text{sim}} - s_{\text{obs}}\|$ is large) is computationally preferred in most instances [282]. This is a motivation behind the use of the smoothing kernel density as a discrepancy metric. There exist other rejection-based ABC samplers that employ importance sampling and a positive smoothing kernel function $K_h(u)$ with bandwidth $h > 0$, where $\max_u K(u) = 1$ and $u = \|s_{\text{sim}} - s_{\text{obs}}\|$ (with $\|\cdot\|$ being a real norm; for instance, the Euclidean norm). K_h could be defined using any standard (symmetric) kernel such as the Gaussian, Epanechnikov, triangular, biweight or uniform kernels; where the kernel can be set such that $K_h(u) = K\left(\frac{u}{h}\right)$ [191]. For multivariate extension of these kernels in the ABC setting, see work by Sisson and Fan [281]. Phillips and Venkatasubramanian [240] have reviewed different class of kernel distance measures with more formal definitions. Algorithm L1 [282, p. 16] summarises how particles from an approximate posterior (defined by equation 5.5) is generated by incorporating importance sampling. Here, each particle $\theta \sim g(\theta)$ is sampled directly from an importance distribution with density g instead of the prior $\pi(\theta)$ with samples of θ accepted with probability $\propto \frac{\pi(\theta)}{g(\theta)}$. The ABC rejection algorithm with importance sampling defined by Algorithm L1 leads to sampling from an approximate posterior (as $h \to 0$) with marginal density defined as [190, 209]:

$$\begin{aligned} p_h(\theta \mid s_{\text{obs}}) &= \int p_h(\theta, s_{\text{sim}} \mid s_{\text{obs}}) ds_{\text{sim}} \\ &= \int \left[\frac{\pi(\theta) f(y_{\text{sim}} \mid \theta) K_h\left(\|s_{\text{sim}} - s_{\text{obs}}\|\right)}{\int_{\mathbb{R}^n \times \mathbb{R}^m} \pi(\theta) f(y_{\text{sim}} \mid \theta) K_h\left(\|s_{\text{sim}} - s_{\text{obs}}\|\right) d\theta ds_{\text{sim}}} \right] ds_{\text{sim}}. \end{aligned} \tag{5.3}$$

Although it is easy to implement the basic rejection ABC and other rejection-based ABC samplers, a major disadvantage of these samplers is a low acceptance rate if the prior distribution is significantly different from the posterior distribution. The trade-off between accuracy and computational capacity is determined by choice of the tolerance threshold (ϵ) or the kernel bandwidth (h) such that, if ϵ or h is too large, posterior estimates can be biased with wide credible intervals; and if ϵ or h decreases, computational cost increases [115]. Analytical results regarding the pointwise bias in the ABC for fixed θ (for both univariate and multivariate cases) as well as specific applications of instances where the ABC posterior was explicitly derived (e.g., for Algorithm L1) exist [see 281, pp. 21-26]. Li and Fearnhead [191] have examined the accuracy of estimators in relation to the Monte Carlo error of an importance sampling procedure that samples from the ABC posterior and a condition to reduce the asymptotic variance of the posterior mean

Algorithm L1: ABC rejection and importance sampling algorithm

```
Inputs:
• Prior distribution $\pi(\theta)$ where $\theta \in \mathbb{R}^n$ and a procedure for generating data ($y_{\text{sim}}$)
  under the model $f(\cdot \mid \theta)$.
• A proposal density function $g(\theta)$, with $g(\theta) > 0$ if $p(\theta \mid y_{\text{obs}}) > 0$.
• A pre-specified number of proposal draws $N > 0$, and a kernel function $K_h(\cdot)$ with
  scale parameter or bandwidth $h > 0$.
• Observed summary statistics $s_{\text{obs}} = S(y_{\text{obs}})$ computed from the observed data $y_{\text{obs}}$;
  where $s = S(\cdot) \in \mathbb{R}^m$ is a low-dimensional vector of summary statistics.

Sampling:
For $i = 1, \cdots, N$ :
1. Generate $\theta_i \sim g(\theta)$ from sampling density $g$.
2. Generate $y_{\text{sim}} \sim f\,(\cdot \mid \theta_i)$, and compute $s_{\text{sim}} = S(y_{\text{sim}})$.
3. Accept $\theta_i$ with probability $\frac{K_h\left(\|s_{\text{sim}} - s_{\text{obs}}\|\right)\pi(\theta_i)}{\mathcal{K} g(\theta_i)}$,
where $\mathcal{K} \geq K_h(0) \max_{\theta} \frac{\pi(\theta_i)}{g(\theta_i)}$. Else go to step 1.

Output:
A set particles $\{\theta_j\}_{1 \leq j \leq N_h} \sim p_h(\theta \mid s_{\text{obs}})$,
where $N_h$ is the number of accepted particles at given kernel scale parameter $h$.
```

(such that the dimension of summary statistics should be set equal to that of the model parameters). For a more extensive theoretical study on the ABC convergence rate and computational cost, see work by Barber and Voss [22]. There exist good theoretical and asymptotic properties of ABC posterior and its mean estimates (as the Monte Carlo sample size, N, increases) including the associated ABC error [see 104, 190, 191]. The Monte Carlo error is not dependent on only the ABC sampler but also by the choices of kernel bandwidth (or tolerance threshold) and the summary statistics [191]. Generally, higher dimension of the summary statistics, or a smaller value of ϵ or h, tend to increase the Monte Carlo error by reducing acceptance probability.

Other variants of the rejection and importance ABC algorithms have been proposed in the literature to improve the basic rejection ABC (e.g., ABC Rejection Control Importance Sampling, and ABC-KNN Importance Sampling) [34, 194, 239], Markov Chain Monte Carlo ABC (ABC-MCMC) samplers [50, 210], sequential Monte Carlo ABC (ABC-SMC) samplers [64, 79, 194], regression-adjusted ABC methods (for ABC post-processing).

Also, there exist methodologies for approximating the intractable likelihood function such as synthetic likelihoods, expectation-propagation ABC, and copula or regression-density estimation models [24, 97, 189]. Other adaptive ABC methodologies on tolerance selection and stopping rules have been proposed for sequential-type ABC methods [26, 62, 80, 279, 283, 295]. Nonetheless, these adaptive procedures for selecting tolerance and terminating iterative ABC algorithms after posterior convergence may not always be robust for calibrating other complex systems and high-dimensional models with more than five parameters [279]. For instance, common adaptive approaches for selecting the tolerance sequences for sequential ABC samplers are i) automatically choosing tolerance (ϵ_t) at step t based on quantiles of the distances corresponding to the accepted particles from iteration $t-1$ [66, 156, 280], ii) fixing the decreasing tolerance levels in advance [26, 283, 295], or iii) adaptively selecting ϵ_t based on some quantile of the effective sample size values [80, 227]. However, these approaches can result in inefficient sampling, and predetermined quantile can lead to the particle system remaining in local modes, if not chosen carefully [278].

The rest of section 5.2 is organised as follows. In Section 5.2.2, the impact of various levels of approximation in ABC (concerning the formulated data-generation process, summary statistics usage instead of the data itself, and summary statistics weighting) is presented. In Section 5.2.3, we briefly describe a few efficient ABC algorithms, such as ABC-MCMC (section 5.2.3.1) and ABC-SMC (section 5.2.3.2) samplers, and regression-adjusted ABC methods (section 5.2.3.3).

5.2.2 Various levels of approximation in ABC

The most challenging part of executing an ABC analysis is minimising the impact of the approximation while keeping the required computation to a minimum. As a result, a brief summary of the various stages or levels of approximation from model development to implementation of the ABC algorithms is imperative. Below are the key factors that can affect the ABC posterior accuracy based on the different approximation levels.

5.2.2.1 Approximations from data-generating processes or model

All simulation models or data-generating processes approximate a natural data-generating process of a system (e.g., biological systems). Simulations from stochastic models, for instance, can result in an ABC-specific issue if the suggested model is not robust enough to reproduce outputs that are very close to the empirical data. In this case, all simulated data would be far from that observed data, resulting in less accurate model calibrations via ABC estimation approaches [282]. Specifically, Frazier et al. [105] have shown with theoretical justifications that, given a misspecified simulation model, the ABC posterior and local-linear posterior adjustment (for ABC post-processing analysis) may not result in credible intervals with appropriate frequentist coverage (i.e., the minimum probability, for any parameter θ, that the credible interval will include the true θ value) as well as lead to a non-standard asymptotic behaviour. However, their theoretical results also revealed that some ABC sampling methods can still concentrate posterior mass on an appropriately defined pseudo-true parameter value under regularity conditions even if the model is misspecified. Ridgway [258] discovered in another theoretical and numerical study that some versions of ABC where posterior convergence are less quickly (e.g., sequential Monte Carlo ABC with importance sampling) are inherently robust to model misspecification at the cost of choosing either a larger kernel bandwidth (when a kernel is used in Monte Carlo methods) or tolerance threshold. Generally, worse ABC posterior approximation may also suggest a high chance of model misspecification. Consequently, inspecting the ABC posterior alone may not be sufficient to determine model adequacy since the ABC posterior may be a poor estimate of the true posterior, and less accurate simulation models may appear more likely than they are. Hence, the ABC posterior approximation's performance depends on the accuracy of the underlying simulation model replicating the system.

5.2.2.2 Summary statistics usage instead of full data

As previously highlighted, ABC algorithms typically reduce high-dimensional data to a low-dimensional user-chosen summary statistics and accept samples of the model parameter $\theta \in \mathbb{R}^n$ when the simulated summaries $s_{\text{sim}} \in \mathbb{R}^m$ are close to the observed summaries s_{obs}. Due to the general issue of the curse of dimensionality associated with the number of summaries of high-dimensional data for most ABC samplers (e.g., ABC rejection and MCMC samplers), low acceptance rates are often achieved. Thus, the tolerance is rather increased, leading to a high rate of distortion of the ABC approximation. The curse of dimensionality can be formally approached by considering how ABC approximation error relates to the Monte Carlo sample size, N [22]. Other studies have also shown that by adopting importance sampling, the Monte Carlo error in ABC can be made arbitrarily minimal [296]. It can be shown asymptotically that the rate at which ABC error decays becomes worse as N increases [243]. Under some regularity conditions and optimal ABC tuning, the mean squared error of a Monte Carlo estimate produced by ABC rejection sampling is shown to be $\mathcal{O}_n\left(N^{\frac{-2}{m+4}}\right)$ [22]; where m and n are the dimension of the summaries and the model parameters, respectively. Many authors have considered other definitions of error and several ABC algorithms, and qualitatively proved similar results (i.e., similar asymptotic results with slight modifications to the exponent of N) [see 34, 37, 99]. Although these asymptotic results do not precisely capture the behaviour for the kernel bandwidth h (typically a bias-variance trade-off), they have confirmed that high-dimensional summaries often result in poor posterior approximations. Hence, selection and dimensionality reduction techniques of summary statistics in ABC algorithms for parameter inference, model selection and model predictions have become vital research areas in recent times [2, 51, 136, 165].

Strategies for selecting summary statistics (to circumvent the curse of dimensionality problem) in the literature can be broadly grouped into three methods: subset selection, projection, and auxiliary likelihood (but each has its computational challenges). Firstly, given a candidate summary statistics $s_{\text{sim}} \in \mathbb{R}^m$, the subset selection method attempt to choose an informative representative subset of length $l < m$ via an iterative procedure to assess the significant impact of adding extra summary statistics based on its modification to the approximated posterior. This is done by: i) using a stepwise selection technique based on an approximate sufficiency test proposed by Joyce and Marjoram [165], ii) finding a subset of s_{sim} from either two-stage approach that first minimises the entropy of the approximated posterior and then minimises an error loss function (e.g., root mean square error) [228], or a subset that iteratively maximise the Kullback-Leibler divergence between two ABC posterior samples (based on existing summary subset and a newly created subset) in stepwise manner [23], iii) applying a regularisation method based on lasso regression [140, 272] or performing ABC regression post-processing and performing variable selection by an empirical Bayes approach [38]. These subset selection methods may result in high computational cost and scalability problems [243]. Also, a large ABC approximation error may significantly influence the judgements regarding the usefulness of a statistic at any point of the iterative procedure.

The projection methods begin with a set of data features and then find informative low-dimensional projections based on linear transformation. Thus, the subset selection method can be considered a particular case of the projection method. These projection techniques employ either the partial least squares regression approach by Wegmann et al. [313] (similar to the regularisation subset selection method) or a semi-automatic method (which minimises the quadratic loss of the parameter point estimates) exist [99]. However, Robert [259] revealed that the former lacks theoretical support compared to the latter. The projection method can avoid the computational costs associated with the subset selection methods and search for a more expansive space of summary statistics. [243]. Lastly, the auxiliary likelihood methods choose summaries for ABC of more complex models by using statistics known to inform a simpler related model [see 245]. The goal is to extract summary statistics (e.g., maximum likelihood estimates) from an auxiliary likelihood (given a simpler model) by obtaining an approximate and tractable likelihood for the data (from the complex model). For instance, in the current study, the linear B-D-C model (defined in Chapter 4) was considered as an auxiliary model for the complex stochastic simulation model (developed in Chapter 6) to inform some components of our

ABC summary statistics based on its parameter estimates (during ABC fitting of the complex model using our modified ABC methods proposed in section 5.3). Prangle [243] has reviewed specific types of auxiliary likelihood approaches. They can be computationally time-consuming, prone to numerical errors, and in some cases, unique MLE may not even exist [243]. This also suggests why the current study also investigated other B-D-C parameter estimation methods, such as a generalised method of moments and the Galton-Watson estimation approach (in addition to MLE), by exploring the trade-off between computational speed and estimation accuracy (in Chapter 4).

5.2.2.3 Weighting summary statistics and choice of tolerance in ABC analysis

Jung and Marjoram [166] demonstrated that utilising weighted summary statistics and a well-chosen tolerance in ABC analysis can tremendously result in enhanced performance when compared to unweighted analysis. The relevance of assigning weights to each summary statistic is to quantify their relative importance (where importance is used informally to denote the amount of information that a summary statistic carries regarding the parameter θ of interest). The weighting of summary statistics within a region of the observed summaries in ABC is typically specified at the start of the algorithm to standardise summaries (or to vary on similar scales). However, in iterative ABC procedures, where there is an update in the proposal distribution of model parameters, the weighting of summary statistics may not be necessary [244]. While adding more informative statistics, in theory, should enhance the degree of ABC posterior approximation, the consequence of doing so without weighting statistics can make the approximation worse [166]. Including uninformative statistics reduces the acceptance rate, and thus, the ABC's accuracy. Consequently, it is imperative to assign a high weight to a small but highly informative set of summary statistics; however, this is a non-trivial problem without a good intuition of which summary measures are adequately informative [166]. Before computing the discrepancy measure, several ways of determining the weights ($\omega \in \mathbb{R}^m$) for $m-$dimensional summary statistics ($s \in \mathbb{R}^m$) exist. A common way is to compute the weighted Euclidean distance ρ (defined by the sum of weighted squared distances) such that

$$\rho(y_{\text{sim}}, y_{\text{obs}}) = \left[\sum_{j=1}^{m} \omega_j \left(s_{\text{sim},j} - s_{\text{obs},j} \right)^2 \right]^{1/2}, \tag{5.4}$$

where $\omega_j = \frac{1}{\sigma_j^2}$, and $\sigma_j^2 = \text{Var}(s_{\text{sim},j})$ is the variance of the jth summary statistic (used as a scaling to normalise the summary measures) for each simulated data $y_{\text{sim}} \sim f(\cdot \mid \theta_i)$, $1 \leq i \leq N$ [27, 244]. The scaling in equation 5.4 prevents the most variable summary from dominating the distance metric. Other distance functions can be used in ABC (e.g., median absolute deviation, kernel functions and a chi-squared metric) [see 73, 215]. However, for iterative ABC algorithms (such as sequential Monte Carlo ABC or population Monte Carlo ABC), there is no assurance that ω would normalise the summary measures computed in subsequent iterations since particles are drawn from a different proposal distribution instead of the prior predictive distribution [244]. The kernel weights can be problematic in ABC (for $h > 0$) when θ is large since the vector of summary statistics must then be equivalently large for parameter identifiability [282]. Hence, the comparison $\|s_{\text{sim}} - s_{\text{obs}}\|$ will suffer from the curse of dimensionality. On the other hand, the choice of tolerance threshold ϵ, in theory, will yield a smaller distance for small $\epsilon \to 0$ but low acceptance rate; and otherwise for large ϵ.

The problem of optimally specifying this tolerance threshold has been the subject of extensive studies. Several adaptive methods of tolerance selections have been proposed in the literature for different iterative and non-iterative ABC sampling methods. In this current study, we adaptively compute summary statistics weights in our modified ABC algorithm based on the inverse of variance for the jth statistic across an entire host population for a given simulation and then sequentially update these weights at a particular ABC time step t in the ABC algorithm based on the summary statistics weights from the previous time step $t - 1$ (by finally computing and updating the summary statistics weights at time $t \geq 1$ based on the harmonic mean of summary weights at $t - 1$ and t; see section 5.3.4 for further details). With a decreasing sequence of tolerance thresholds carefully chosen (e.g., based on some heuristic principles such as using trial and error and good guesses from literature, amongst others), coupled with adaptive importance sampling, we also improve the fidelity of the modified ABC algorithm as implemented in sequential ABC samplers (see sections 5.2.3.2 and 5.3).

5.2.3 Efficient ABC sampling algorithms and adjustments

There are challenges associated with the classical ABC rejection samplers such that: the majority of the samples are taken from regions of parameter space with low posterior probability, and thus, there is a high rejection rate due to potential mismatch between simulated and observed data. As a result, several computational strategies have been proposed to improve the rejection-based ABC samplers' efficiency. These come in broadly three approaches: Regression-adjusted ABC samplers [27, 40, 290], Markov chain Monte Carlo (MCMC) ABC samplers [210, 251], and ABC implementing some variant of sequential importance sampling (SIS) or sequential Monte Carlo (SMC) [26, 283, 295]. The regression-adjusted ABC schemes can be combined with or integrated into other ABC samplers. Section 5.2.3 discusses the aforementioned efficient ABC samplers and regression adjustments that are improvements to the rejection-based ABC samplers. Section 5.2.3.1 describes Markov chain Monte Carlo ABC (ABC-MCMC) samplers; whereas section 5.2.3.2 reviews sequential Monte Carlo ABC (ABC-SMC) samplers (with SIS). Finally, existing regression methods for ABC posterior adjustments are presented in section 5.2.3.3.

5.2.3.1 Markov chain Monte Carlo ABC

Markov chain Monte Carlo ABC (ABC-MCMC) is one of the extensions of the standard rejection-based ABC to improve posterior approximations, especially if the prior distribution $\pi(\theta)$ of the model parameter θ is non-informative by minimising sampling from low posterior probability regions [209]. Marjoram et al. [210] thus introduced the ABC-MCMC sampler (based on the Metropolis-Hastings algorithm) by targeting the approximate posterior $p_\epsilon(\theta \mid s_{\text{obs}})$ from the joint density given by equation 5.5 (in the case where a kernel function is not considered, for instance); such that

$$p_\epsilon(\theta, s_{\text{sim}} \mid s_{\text{obs}}) = \frac{f(s_{\text{sim}} \mid \theta)\pi(\theta)\mathbb{1}_{A_{\epsilon,s_{\text{obs}}}}(s_{\text{sim}})}{\int_{A_{\epsilon,s_{\text{obs}}} \times \Theta} f(s_{\text{sim}} \mid \theta)\pi(\theta) ds_{\text{sim}} d\theta} \tag{5.5}$$

where $\mathbb{1}_{\mathcal{A}}(\cdot)$ represent the indicator function of the set $\mathcal{A}$, and

$$A_{\epsilon,s_{\text{obs}}} = \{s_{\text{sim}} \mid \rho(s_{\text{sim}}, s_{\text{obs}}) \leq \epsilon\},$$

with ϵ, s_{obs}, s_{sim} and ρ defined as before. Wegmann et al. [313] have provided some implementation tips for the ABC-MCMC sampler as well as proof of the central result in the former. ABC-MCMC samplers basically add a proposal chain with density g and a rejection step to generate a sample of θ [11].

Algorithms L2 [adapted from 11, 210] and L3 [adapted from 282] are two alternative implementations of ABC-MCMC sampling with or without a kernel distance metric (where h denote the kernel bandwidth as defined earlier). The main advantage of Algorithm L2 over Algorithm L3 is that the former avoids unnecessarily running the simulation step by simulating data based on only accepted proposals; whereas the latter has a minimal number of MCMC rejection steps comparatively with the acceptance probability directly dependent on the kernel distance (between s_{sim} and s_{obs}). The ABC-MCMC algorithms basically generate samples from $p(\theta \mid \rho(s_{\text{sim}}, s_{\text{obs}}) \leq \epsilon)$ or $p(\theta \mid K_h(\|s_{\text{sim}} - s_{\text{obs}}\|))$ (for small enough ϵ or h), adds a proposal chain (with density g) and a rejection scheme (also determined by α) to generate a sample θ_i for $i = 1, 2 \cdots, N$. The resulting posterior distribution is a stationary and limiting distribution of the chain under suitable regularity conditions [210]. The ABC-MCMC algorithm is most likely to converge as $\epsilon \to 0$ (or $h \to 0$), but determining when the Markov chain reaches the stationary regime is challenging; also, the chain may become trapped in local modes.

Algorithm L2: ABC-MCMC algorithm without a kernel distance metric

Inputs:

- Prior distribution $\pi(\theta)$ where $\theta \in \mathbb{R}^n$ and a procedure for generating data (y_{sim}) under the model $f(\cdot \mid \theta)$.
- A Markov proposal density function or transition kernel $g(\theta, \theta^*) = g(\theta^* \mid \theta)$.
- A pre-specified number of proposal draws $N > 0$, tolerance $\epsilon > 0$ and some distance function ρ.
- Observed summary statistics $s_{\text{obs}} = S(y_{\text{obs}})$ computed from the observed data y_{obs}; where $s = S(\cdot)$ is a low-dimensional vector of summary statistics.

Initialise:

Repeat:
1. Choose an initial parameter vector $\theta_0 \sim \pi(\theta)$ from the prior
2. Generate $y_{\text{sim},0} \sim f(\cdot \mid \theta_0)$ from the model and compute summary statistics $s_{\text{sim},0} = S(y_{\text{sim},0})$, until $\rho(s_{\text{sim},0}, s_{\text{obs}}) > \epsilon$.

Sampling:

For $i = 1, \cdots, N$:
1. Generate candidate vector $\theta^* \sim g(\cdot \mid \theta_{i-1})$ from the proposal density g.
2. Accept θ^* with probability $\alpha = \min\left\{1, \frac{\pi(\theta^*)g(\theta_i|\theta^*)}{\pi(\theta_{i-1})g(\theta^*|\theta_{i-1})}\right\}$,
and go to step 3, otherwise return to step 1.
3. Generate $y^*_{\text{sim}} \sim f(\cdot \mid \theta^*)$, and compute $s^*_{\text{sim}} = S(y^*_{\text{sim}})$.
4. If $\rho(s^*_{\text{sim}}, s_{\text{obs}}) \leq \epsilon$ then set $(\theta_i, s_{\text{sim},i}) = (\theta^*, s^*_{\text{sim}})$, where $s_{\text{sim},i} = S(y_{\text{sim},i})$
with $y_{\text{sim},i} \sim f(\cdot \mid \theta_i)$. Otherwise, set $(\theta_i, s_{\text{sim},i}) = (\theta_{i-1}, s_{\text{sim},i-1})$ and return to step 1.

Output:

A set of correlated parameter vectors $\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(N_\epsilon)} \sim p_h(\theta \mid s_{\text{obs}})$ from a Markov chain with stationary distribution $p_\epsilon(\theta \mid s_{\text{obs}})$, where N_ϵ is the number of accepted particles at tolerance threshold of ϵ.

Algorithm L3: ABC-MCMC algorithm with a kernel distance metric

Inputs:

- Prior distribution $\pi(\theta)$ where $\theta \in \mathbb{R}^n$ and a procedure for generating data (y_{sim}) under the model $f(\cdot \mid \theta)$.
- A Markov proposal density function or transition kernel $g(\theta, \theta^*) = g(\theta^* \mid \theta)$.
- A pre-specified number of proposal draws $N > 0$, and a kernel function $K_h(\cdot)$ with scale parameter or bandwidth $h > 0$.
- Observed summary statistics $s_{\text{obs}} = S(y_{\text{obs}})$ computed from the observed data y_{obs}; where $s = S(\cdot)$ is a low-dimensional vector of summary statistics.

Initialise:

Repeat:
1. Choose an initial parameter vector $\theta_0 \sim \pi(\theta)$ from the prior
2. Generate $y_{\text{sim},0} \sim f(\cdot \mid \theta_0)$ from the model and compute summary statistics $s_{\text{sim},0} = S(y_{\text{sim},0})$, until $K_h\left(\|s_{\text{sim},0} - s_{\text{obs}}\|\right) > 0$.

Sampling:

For $i = 1, \cdots, N$:
1. Generate candidate vector $\theta^* \sim g(\cdot \mid \theta_{i-1})$ from the proposal density g.

2. Generate $y^*_{\text{sim}} \sim f\,(\cdot \mid \theta^*)$, and compute $s^*_{\text{sim}} = S(y^*_{\text{sim}})$.
3. Accept θ^* with probability $\alpha = \min\left\{1, \frac{K_h\left(\|s^*_{\text{sim}} - s_{\text{obs}}\|\right)\pi(\theta^*)g(\theta_i|\theta^*)}{K_h\left(\|s_{\text{sim},i-1} - s_{\text{obs}}\|\right)\pi(\theta_{i-1})g(\theta^*|\theta_{i-1})}\right\}$, and set $(\theta_i, s_{\text{sim},i}) = (\theta^*, s^*_{\text{sim}})$ where $s_{\text{sim},i} = S(y_{\text{sim},i})$ with $y_{\text{sim},i} \sim f(\cdot \mid \theta_i)$. Otherwise, set $(\theta_i, s_{\text{sim},i}) = (\theta_{i-1}, s_{\text{sim},i-1})$.

Output:

A set of correlated parameter vectors $\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(N_h)} \sim p_h(\theta \mid s_{\text{obs}})$ from a Markov chain with stationary distribution $p_h(\theta \mid s_{\text{obs}})$, where N_h is the number of accepted particles at given kernel scale parameter h.

To avoid ABC-MCMC samplers getting stuck in low posterior regions, there has been developments of more improved versions in the literature, such as Augmented space ABC-MCMC samplers [e.g., 47, 250], Hamiltonian Monte Carlo ABC samplers [e.g., 216], and multi-try Metropolis ABC [e.g., 177]. Nevertheless, ABC-MCMC samplers (like rejection-based ABC samplers) also suffer from the curse of dimensionality concerning the number of summary statistics (a general limitation in MCMC methods for multidimensional problems), resulting in convergence issues in some instances [282, 313].

5.2.3.2 Sequential Monte Carlo ABC

Sequential-based Monte Carlo ABC (ABC-SMC) samplers are relatively robust for complex model calibration. They can overcome some of the shortcomings associated with rejection-based ABC and ABC-MCMC samplers, especially when coupled with sequential importance sampling (SIS) to generate particles in high posterior regions [26, 282]. ABC-SMC samples iteratively from a set of improving intermediate distributions that effectively converge to the target posterior distribution. Different variants of the SIS and SMC samplers, including the theoretical aspect of its posterior distribution and convergence, have then been introduced in the ABC framework by several authors [see 26, 79, 102, 284, 295].

Algorithm L4 summarises an approach of implementing ABC-SMC with SIS, where particles are assigned importance weights, $W_i^{(t)}$, for $1 \leq i \leq N$ and $1 \leq t \leq T$ (with T denoting the final ABC time step). Particles are initially sampled from the prior $\pi(\theta)$ and then iteratively propagated through a sequence of intermediate distributions, $p(\theta \mid \rho(s_{\text{obs}}, s_{\text{obs}}) \leq \epsilon_t)$ for $1 \leq t \leq T-1$, until the resulting posterior, $p(\theta \mid \rho(s_{\text{obs}}, s_{\text{obs}}) \leq \epsilon_T)$, is obtained. Instead choosing a single tolerance (as in the case of rejection-based ABC and ABC-MCMC samplers), a monotonically decreasing sequence of tolerances are chosen such that $\epsilon_1 > \cdots > \epsilon_T > 0$. Thus, the distribution of g_t gradually converge towards the target posterior distribution, $p_\epsilon(\theta \mid s_{\text{obs}}) \propto \pi(\theta) \int f(s_{\text{sim}} \mid \theta) \mathbb{1} \left[\rho\left(s_{y_{\text{sim}}}, s_{y_{\text{obs}}}\right) \leq \epsilon_t\right] ds_{y_{\text{sim}}}$.

Algorithm L4: Sequential Monte Carlo ABC (ABC-SMC) algorithm

Inputs:

- Prior distribution $\pi(\theta)$ where $\theta \in \mathbb{R}^n$ and a procedure for generating data (y_{sim}) under the model $f(\cdot \mid \theta)$.
- A perturbation kernel with sequence $\left\{K_h^{(t)}(\cdot \mid \cdot)\right\}_{1 \le t \le T}$ and bandwidth $h > 0$, which determines the importance or proposal distribution with density g.
- A pre-specified number of proposal draws $N > 0$, decreasing tolerances $\{\epsilon_t\}_{1 \le t \le T}$ (where T is the final ABC time step) and some distance function ρ.
- Observed summary statistics $s_{\text{obs}} = S(y_{\text{obs}})$ computed from the observed data y_{obs}; where $s = S(\cdot)$ is a vector of summary statistics.

Sampling:

```
 for all $1 \le t \le T$ do
1.  $i = 1$
   repeat
      if $t = 1$ then
         sample $\theta^* \sim \pi(\theta)$
      else
         sample $\theta$ from the previous population $\left\{\theta_i^{(t-1)}\right\}_i$ with importance weights
         $\left\{W_i^{(t-1)}\right\}_i$
         perturb $\theta^* \sim K_h^{(t)}(\cdot \mid \theta)$ such that $\pi(\theta^*) > 0$
      end if
2.       simulate $y_{\text{sim}} \sim f(\cdot \mid \theta^*)$
      if $\rho(s_{\text{sim}}, s_{\text{obs}}) \le \epsilon_t$ then
         $\theta_i^{(t)} \leftarrow \theta^*$
         $i \leftarrow i + 1$
      end if
   until $i = N + 1$
3.    calculate the importance weights: for all $1 \le i \le N$
   if $t \neq 1$ then
            $W_i^{(t)} = \frac{\pi\left(\theta_i^{(t)}\right)}{\sum_{j=1}^{N} W_j^{(t-1)} K_h^{(t)}\left(\theta_i^{(t)} | \theta_j^{(t-1)}\right)}$
   else $W_i^{(1)} = 1$
   end if
4.  normalise the importance weights over all $1 \le i \le N$.
```

Output:

A sample of weighted particles $\left\{\theta_j^{(T)}\right\}_{1 \le j \le N_{\epsilon_T}}$, where N_{ϵ_T} is the number of accepted particles at final time step T.

A modified version of Algorithm L4 is proposed in the current study in section 5.3 with a more detailed description of the step-by-step ABC-SMC sampling scheme. In principle, at a large Monte Carlo sample size (N), ABC-SMC samplers can avoid the problem of getting stuck in regions of low posterior probability as observed in most ABC-MCMC samplers [295]. There is extensive literature on the construction of efficient SMC and SIS as well as several other adaptive approaches and optimal kernels in ABC-SMC [64, 79, 80, 102, 295]. Some variants of ABC-SMC algorithms may incorporate the following: effective sample size computation (to determine the need for re-sampling particles) [e.g., 80, 89, 195], kernel distance metric [282, pp. 111-114], adaptive determination of tolerances and adaptive stopping rule to terminate the algorithm after posterior convergence [e.g., 279]. While ABC-SMC algorithms are computationally efficient, the overall computational burden is dependent not only on the model's complexity and the amount of data available, but also on the underlying components of the type of sequential scheme adopted [102].

The perturbation kernel $\left\{K_h^{(t)}(\cdot \mid \cdot)\right\}_{1\leq t\leq T}$, which is used to avoid particle degeneracy, can be flexibly chosen (and its bandwidth optimally determined). However, other studies have proposed several optimal perturbation kernels [102]. Generally, the optimal perturbation kernels can be a component-wise perturbation kernel (with independent normal densities or uniform densities) [26, 102, 215], multivariate normal perturbation kernel (take into account the potential correlations between weighted particles at any time t) [102], and local perturbation kernel (e.g., multivariate normal kernel with M-nearest neighbours or optimal local covariance matrix [102], or perturbation kernel based on the Fisher information [70]). For exact forms of the aforementioned optimal perturbation kernels and their specific theoretical properties, see work by Filippi et al. [102]. Some of the limitations associated with ABC-SMC samplers are that the computational demand involved in calculating the importance weights increases quadratically as a function of the number of particles (N), and these samplers can result in particle duplications depending on the magnitude of assigned importance weight [45]. If a large number of parameters must be estimated, ABC-SMC samplers may suffer from the dimensionality curse issues (as briefly discussed in section 5.2.2.2) [175]. ABC-SMC samplers have several advantages. Because the SMC sampler is population-based, complex posteriors which could be multi-modal, can be investigated more efficiently. Samples are drawn from a variety of proposal distributions $(g_1, g_2, \cdots, g_T)$ with varying tolerances $(\epsilon_1 > \epsilon_2 > \cdots > \epsilon_T)$, which allows for an a posteriori or dynamic study of the posterior's robustness or sensitivity to these thresholds. In contrast to the rejection-based sampler, major inefficiencies caused by mismatches between (initial) prior and target distributions are also circumvented [283].

5.2.3.3 Regression-adjusted ABC

The regression approaches in ABC fitting were originally proposed by Beaumont et al. [27] as post-processing step of the ABC output $p_\epsilon(\theta \mid s_{\text{obs}})$ to account for the imperfect match between the simulated $(s_{\text{sim}} \in \mathbb{R}^m)$ and observed $(s_{\text{obs}} \in \mathbb{R}^m)$ summary statistics [210, 283]. Thus, regression adjustment aims to improve the resulting approximation to the true posterior [190]. With the help of smooth weighting and regression adjustment, the rejection-based ABC samplers can be improved to handle a higher number of summary statistics and thereby deal with the curse of dimensionality [27]. This has led to coupling the rejection-based samplers with linear or non-linear regression adjustment [27, 40, 267, 282]. From similar intuition, the posterior distribution approximations from other efficient ABC samplers, such as ABC-MCMC and ABC-SMC, may be further corrected.

The existing regression adjustment methods are i) standard multiple linear regression with homoscedastic adjustment [27], ii) local-linear regression (LLR) [27] with heteroscedastic adjustment, and iii) a non-linear heteroscedastic conditional regression via a feed-forward neural network (FFNN) [40] in the neighbourhood of the observed summaries s_{obs}. According to Blum [39], the homoscedastic regression adjustments may not always be valid. This is because, when the Monte Carlo sample size (N) for ABC fitting is not very large due to computational constraints, local approximations which employ the homoscedastic assumption, are no longer valid because the neighbourhood corresponding to simulations for which $K_h(\|s_{\text{sim},i} - s_{\text{obs}}\|) \neq 0$, is too large for $1 \leq i \leq \eta$. The regression errors are likely to be heteroscedastic and thus, LLR and FFNN regression with heteroscedastic adjustments are improvements of the homoscedastic adjustment by weighting the accepted samples based on a kernel [27, 40]. To perform LLR with heteroscedastic adjustment on the η accepted samples of the model parameter $\theta \in \mathbb{R}^n$ (from ABC posterior approximations), Beaumont et al. [27] specified the regression model in the form

$$\theta_i = \alpha + (s_{\text{sim},i} - s_{\text{obs}})^\top \beta + \xi_i, \quad i = 1, 2, \cdots, \eta; \tag{5.6}$$

where α is the intercept, β is a vector of regression coefficients corresponding to predictors, ξ_i is uncorrelated error term with mean zero and non-constant variance (no other parametric assumptions are made about the distribution of ξ), η is the number of accepted samples, and $s_{\text{sim},i} = S(y_{\text{sim},i})$ for $y_{\text{sim},i} \sim f(\cdot \mid \theta_i)$. The weighted least-squares (WLS) estimates of (α, β) are obtained by minimising

$$\sum_{i=1}^{\eta} \{\theta_i - \alpha - (s_{\text{sim},i} - s_{\text{obs}})^\top \beta\}^2 K_h(\|s_{\text{sim}} - s_{\text{obs}}\|);$$

where $K_h(\cdot)$ is kernel function with bandwidth h (Epanechnikov kernel was chosen in [27], but other kernels such as a Gaussian kernel could be used due to its computational advantage). Then, the solution is

$$(\hat{\alpha}, \hat{\beta}) = (X^\top W X)^{-1} X^\top W \theta, \tag{5.7}$$

where $X = [1 \mid X_{*,1} \mid \cdots \mid X_{*,p}]$ is an $\eta \times (p+1)$ design matrix (with entries $X_{*,j} = s_{\text{sim}_{ij}} - s_{\text{obs}_j}$ for column $j = 2, 3, \cdots, p+1$; whereas, $s_{\text{sim}_{ij}}$ is the jth element of s_{sim_i}), p is the number of predictors, W is an $\eta \times \eta$ diagonal weighting matrix whose ith diagonal element is $K_h(\|s_{\text{sim},i} - s_{\text{obs}}\|)$, and $\hat{\alpha}$ is the adjusted posterior mean of θ given by $\frac{\sum_{i=1}^{\eta} K_h\left(\left\|s_{\text{sim}_i} - s_{\text{obs}}\right\|\right)\theta_i^*}{\sum_{i=1}^{\eta} K_h\left(\left\|s_{\text{sim}_i} - s_{\text{obs}}\right\|\right)}$ with $\theta_i^* = \hat{\alpha} + (s_{\text{sim},i} - s_{\text{obs}})^\top \hat{\beta}$. Then, the resulting adjusted posterior is the distribution of $\theta^* = \theta - (s_{\text{sim}} - s_{\text{obs}})^\top \hat{\beta}$ where $(\theta, s_{\text{sim}}) \sim p_\epsilon(\theta, s_{\text{sim}})$. The samples we get from the regression-adjusted ABC is basically a draw from $p_\epsilon^*(\theta^* \mid s_{\text{obs}})$

where the density of θ^* is

$$p^*_\epsilon(\theta^* \mid s_{\text{obs}}) = \int_{\mathbb{R}^m} p_\epsilon\{\theta^* + (s_{\text{sim}} - s_{\text{obs}})^\top \hat{\beta}, s_{\text{sim}} \mid s_{\text{obs}}\} ds_{\text{sim}}. \tag{5.8}$$

The variance of $p^*_\epsilon(\theta^* \mid s_{\text{obs}})$ is thus strictly small than $p_\epsilon(\theta \mid s_{\text{obs}})$ [190]. Li and Fearnhead [190] have provided asymptotic results on the convergence of $p^*_\epsilon(\theta^* \mid s_{\text{obs}})$, and it was shown that for appropriate choice of the kernel bandwidth (h), regression adjustment produces a posterior that correctly quantify uncertainty about θ.

The Nadaraya-Watson estimator of the posterior mean ($\hat{\alpha}$) still suffers from the dimensionality curse since the convergence rate of the estimator declines drastically as the dimension of the summary statistics increases [40]. On the other hand, Blum and François [40] non-linear FFNN with heteroscedastic adjustments can reduce the dimensionality of the collection of summary statistics using internal projections on lower-dimensional subspaces and also deal with collinearity of the design matrix [73]. However, dimension reduction of the summary statistics is not a focus in this study due to our choice of sequential Monte Carlo ABC sampler and weighted summary statistics coupled with a penalised regression adjustment. Also, the simulated summaries in the current study are likely to be multicollinear (in the vicinity of s_{obs}) due to the summary measures considered in this study and the design data matrix X can be high-dimensional or supercollinear (i.e., when the number of predictors exceeds the number of accepted particles after the sequential Monte Carlo ABC sampling). Thus, the term $X^\top W X$ (with dimension $(p+1) \times (p+1)$) in equation 5.7 is possibly not invertible (or linearly dependent) or close to singular, or estimated regression coefficients from the WLS estimator unreliable if $\text{rank}(X^\top W X) < p+1$ [308] (due to supercollinearity of the design matrix). Hence, another contribution of this study is to improve upon Beaumont et al. [27] LLR with heteroscedastic adjustment by using weighted ridge regression to obtain a robust estimator in the presence of multicollinearity, supercollinearity and non-normal residuals (see section 5.3.4 for the extended local-linear regression with $L2$ regularisation). The additional improvement is achieved by shrinking regression coefficients close to zero for predictors with less contribution to the approximated posterior samples (in the neighbourhood of the observed summaries), thereby minimising the standard errors corresponding to the regression coefficients. Consequently, the chance of model over-fitting due to model complexity (as a function of weights) and other levels of approximations in ABC can be minimised.

However, an important component of regression adjustment concerns posterior shrinkage since regression adjustment shrinkage can be severe, especially at low tolerance rates, and it can impact posterior calibration [39]. In a model of admixture, for instance, 95% credibility intervals obtained with regression adjustments were found to contain only 84% of the true values in less favourable situations [252]. Advantages of the weighted ridge regression (WRID) [originally proposed in 10] are that i) WRID is robust to sample size, non-normal errors and outliers [10, 149], ii) estimates from WRID improves as the predictors become more collinear [315], and can be used when the design data matrix is supercollinear [308]. A Gaussian kernel is used to estimate the diagonal elements of the weighting matrix W in the modified local-linear regression. Additionally, since the complex stochastic model simulates data for a host population, a matrix of summary statistics with dimension $M \times m$ (where M is the population size and m is the length of summaries per host) is returned in a single run or for each accepted particle. Hence, the design data matrix X [specified in 27]) is also modified (due to the high-dimensional summary statistics in the current study) as well as standardised in this study before posterior mean adjustment since the summary statistics are measured on different scales.

In section 5.3, we present a modified ABC-SMC algorithm and an extended regression-adjusted ABC method (with $L2$ regularisation) needed to calibrate our novel stochastic simulation model for the gyrodactylid-fish system.

5.3 The Weighted-iterative ABC

5.3.1 Introduction

The pseudo-code of the modified ABC algorithm dubbed "weighted-iterative ABC" for the complex model calibration is described in Section 5.3.2. The ABC algorithm developed in the current study is a modification of the ABC-SMC sampler described by Algorithm L4 in section 5.2.3.2. For our modified ABC-SMC algorithm, the set of summary statistics per host to extract relevant information from high-dimensional parasite population data is also weighted (for a detailed description of the empirical data, refer to section 2.2.1). The weighted-iterative ABC (Algorithm 4) is used to estimate the model parameters of a novel multidimensional individual-based stochastic model (with at least 23 model parameters) for the gyrodactylid-fish system (in Chapter 6). For a single simulation run, our stochastic model simulates multidimensional data or a set of M sample paths over time and space (i.e., across the host's body regions), corresponding to the entire observed fish with a population size of M (see Chapter 6 for specific details). To compute the summary statistics for both simulated and observed data (where the total observed fish infected at the beginning of the observation period is $M = 152$), the linear B-D-C process (defined in section 4.1.1 of Chapter 4) is considered as an auxiliary stochastic model to the complex simulation model to refine the summary statistics (by including the parameter estimates of the B-D-C process as additional summaries).

The set of summary statistics computed for a given host data is: i) log count of parasites across observed times (**9** summaries), ii) Wasserstein $1-D$ distance between host's body regions (**4** summaries), iv) the time before death (**1** summary) and v) parameter estimates of the B-D-C process based on all simulated sample paths using the Galton-Watson and GMM estimation approaches as recommended in section 4.2 (**3** summaries). Thus, for the entire host population or in a simulation run, a matrix with a dimension of 152×17 summary statistics is obtained for comparing the discrepancy between the simulated and observed data during ABC fitting of the complex stochastic model. A weighted sum of squares distance metric ρ, which extends the standard weighted Euclidean distance (given by equation 5.4), is considered the discrepancy metric. Additionally, an optimised linear regression function (presented in section 5.3.3) is developed to aid in computing the summary statistics after premature host mortality by projecting the infrapopulation of parasites till the end of the infection period.

A regression methodology for ABC posterior mean adjustment and post-processing based on the target posterior samples from the modified ABC algorithm is also presented in section 5.3.4. The post-processing analysis (which supplements the modified ABC algorithm as an independent final step) is employed to adjust further the resulting posterior distribution and the ABC posterior mean after model fitting. Specifically, we have proposed a penalised local-linear regression with heteroscedastic errors (described in section 5.3.4).

5.3.2 Description of the modified ABC algorithm

Algorithm 4 is the pseudo-code for the weighted-iterative ABC algorithm, which employs sequential Monte Carlo and sequential importance sampling. The main modifications in Algorithm 4 with respect to the previous ABC-SMC Algorithm L4 (described in section 5.2.3.2) are: i) adaptively integrating importance weights for importance proposal sampling and summary statistics weights (based on accepted simulations by computing the harmonic mean between previous and current summary statistics weights at time $t \geq 1$) to improve ABC posterior approximations, ii) inclusion of a weighted distance metric for comparing between multidimensional data of an entire population (in the case where summary statistics has bi-dimensional space), iii) adaptation of a computationally efficient multivariate normal perturbation kernel with bandwidth matrix optimally determined, and iv) a separate *post-hoc* step which entails a robust correction method to adjust the resulting ABC posterior approximation using a modified heteroscedastic local-linear regression with $L2$ regularisation. The modified ABC algorithm can briefly be explained as follows:

- Suppose we have a decreasing sequence of tolerances $\epsilon_1 > \epsilon_2 > \cdots > \epsilon_T$ (T being the final time step), the prior distribution $\pi(\cdot)$, a simulation model given by $f(\cdot \mid \theta)$, and a observed summary statistics $s_{\text{obs}} = S(y_{\text{obs}})$ (possibly multidimensional).

- At time $t = 1$, the weighted-iterative ABC algorithm draws proposals $\theta_i^{(1)} \sim \pi(\theta)$ (for $1 \leq i \leq N$) from the prior distribution $\pi(\theta)$ with equal importance weight of $W_i^{(1)} = \frac{1}{N}$; the accepted particles at the largest tolerance ($\epsilon_1 \leq 1$) is indicated as $p_{\epsilon_1}(\theta \mid s_{\text{obs}})$ (or p_{ϵ_1} for simplicity), and considered as the first intermediate prior distribution. Instead of commencing the rejection sampling with a smaller tolerance (as in the case of the standard rejection-based samplers), at $t = 1$, the algorithm is similar to the standard rejection ABC (but with a larger tolerance comparatively). The discrepancy between simulated and observed summary statistics, given $\theta_i^{(t)}$ at time $t \geq 1$, is computed using the scaled weighted sum of squares distance metric such that

$$\rho\left(s_{\text{sim}}, s_{\text{obs}}\right) = \sqrt{\frac{1}{M} \sum_{k=1}^{\mathbf{M}} \sum_{j=1}^{m} \mathbf{w}_j^{(t)} \left(s_{\text{sim}_{k,j}} - s_{\text{obs}_{k,j}}\right)^2}, \quad 1 \leq t \leq T. \tag{5.9}$$

where M is the total population size, m is the summary statistics length per simulation sample path or host ($m = 17$ in our case), $\mathbf{w}^{(t)}$ is a vector of the summary statistics weights at time t, and our summary statistics is assumed to have a bi-dimensional space (for a one-dimensional summary statistics, the standard weighted Euclidean distance defined by equation 5.4 can be used as the discrepancy measure instead). Prior to computing the weighted sum of squares distance metric $\rho(\cdot)$, the summary statistics weight $\mathbf{w}^{(t)}$ at time $t \geq 1$, is computed based on the harmonic mean of the current weight $w_{j'}^{(t)} = 1/\sigma_{j'}^2$ (based on accepted particles, where $\sigma_{j'}$ is the standard deviation of the j'th summary statistic) for $1 \leq j, j' \leq m$ and the previous weight $\mathbf{w}^{(t-1)}$; such that

$$w_j^{(t)} = \frac{2}{\frac{1}{w_j^{(t-1)}} + \frac{1}{w_{j'}^{(t)}}}.$$

According to Prangle [244], there is no assurance that the summary statistics weights $\mathbf{w}^{(t)}$ (meant to normalise the summary statistics at time step $t \geq 1$ for iterative ABC such as ABC-SMC) would actually normalise the summary statistics at subsequent iterations since particles or proposals are not sampled directly from the prior $\pi(\theta)$, but instead, from different proposal distributions $g_t(\theta)$ over time $t \geq 1$. Hence, the main motivations for adopting the harmonic mean of the previous and current summary statistics weights (based on the multiplicative inverse of the variance of the jth summary statistic of accepted particles) in this study (instead of strictly using the conventional approaches defined in [244]) are to i) minimise the degree of variability in the high-dimensional summary statistics weights at time $t \geq 1$ (based on averages across the entire host population as observed in the current study), and ii) control the potential high disparities between the summary statistics weights at the current ABC time step t and the previous time $t-1$ as well as improve normalisation of summary statistics weights due to direct particle sampling from different proposal distributions (at ABC time steps $t-1$ and t) instead of the (initial) prior.

- At $t \geq 2$, the algorithm works in steps (with $\epsilon_t < \epsilon_{t-1}$): instead of directly sampling from $\pi(\theta)$, we randomly draw weighted particles $\theta^* \sim p_{\epsilon_{t-1}}$ (for N different times) from the current intermediate prior $p_{\epsilon_{t-1}}$ with a probability equal to their corresponding normalised importance weight $W_i^{(t)}$ (estimated from equation 5.12). Following Filippi et al. [102], we then perturb particles $\theta_i^{(t)} \sim K_{H^{(t)}}(\cdot \mid \theta^*)$ at iterations $t \geq 2$ using a multivariate normal (MVN) perturbation kernel $K_{H^{(t)}}$ centred at or near θ^*, such that

$$K_{H^{(t)}}\left(\theta^{(t)} \mid \theta^*\right) = \frac{1}{\sqrt{(2\pi)^n \left(\det H^{(t)}\right)}} \exp\left\{-\frac{1}{2}\left(\theta^{(t)} - \theta^*\right)^\top \left(H^{(t)}\right)^{-1} \left(\theta^{(t)} - \theta^*\right)\right\}, \tag{5.10}$$

with an optimal bandwidth matrix

$$H^{(t)} = \sum_{i=1}^{N} \sum_{k=1}^{N_{\epsilon_{t-1}}} W_i^{(t-1)} \tilde{W}_k \left(\tilde{\theta}_k - \theta_i^{(t-1)}\right)\left(\tilde{\theta}_k - \theta_i^{(t-1)}\right)^\top; \tag{5.11}$$

where the quantity $\left\{\tilde{\theta}_k\right\}_{1\leq k\leq N_{\epsilon_{t-1}}}$ denote the set of accepted particles $\left\{\theta_i^{(t-1)}\text{s.t.} \quad \rho(s_{\text{sim}}, s_{\text{obs}}) \leq \epsilon_t, \quad 1 \leq i \leq N\right\}$, with their corresponding importance weight $\left\{\tilde{W}_k\right\}_{1\leq k\leq N_{\epsilon_{t-1}}}$ normalised over all $1 \leq k \leq N_{\epsilon_{t-1}}$. Filippi et al. [102] have shown that this choice of kernel bandwidth has good theoretical properties.

We then simulate data $y_{\text{sim}} \sim f(\cdot \mid \theta_i^{(t)})$ for $1 \leq i \leq N$, obtain N_{ϵ_t} accepted samples $p_{\epsilon_t}(\theta \mid s_{\text{obs}})$ accordingly, and repeat the process until we reach the final or target posterior $p_{\epsilon_T}(\theta \mid s_{\text{obs}})$ at the final time step $t=T$ (where $N \geq N_{\epsilon_1} > N_{\epsilon_2} > \cdots > N_{\epsilon_T}$). Here,

$$W_i^{(t)} = \frac{\pi\left(\theta_i^{(t)}\right)}{\sum\limits_{l=1}^{N} W_l^{(t-1)} K_{H^{(t)}}\left(\theta_i^{(t)} \mid \theta_l^{(t-1)}\right)}, \quad 2 \leq t \leq T \quad \text{and} \quad W_i^{(1)} = \frac{1}{N}. \tag{5.12}$$

- At time $t=1$, the initial prior density $\pi(\theta) \propto g_1(\theta)$ is considered as the first importance or proposal density $g_1(\theta)$; whereas at $t \geq 2$, the importance or proposal density $g_t(\theta)$ is derived from equation 5.13 such that

$$g_t(\theta) = \sum_{i=1}^{N} W_i^{(t-1)} K_t\left(\theta \mid \theta_i^{(t-1)}\right) / \sum_{i=1}^{N} W_i^{(t-1)}. \tag{5.13}$$

Finally, the approximate posterior distribution $p_{\epsilon_T}(\theta \mid s_{\text{obs}})$ at $t=T$ from the weighted-iterative ABC are adjusted using a robust regression adjustment method with $L2$ regularisation (proposed in section 5.3.4) to account for the imperfect mismatch between the simulated and observed data; while accounting for multicollinearity, supercollinearity, non-normal errors as well as employing shrinkage to improve the posterior mean of Beaumont et al. [27] local-linear regression.

Algorithm 4: Pseudo-code of the weighted-iterative ABC

```
Input: Initialise the sequence of decreasing tolerances $\epsilon_1 > \epsilon_2 > \cdots > \epsilon_T$; compute
       initial summary statistics weight $w^{(0)} = (w_1, w_2, \cdots, w_m)$; specify prior
       distribution $\pi(\theta)$; set number of proposal draws $N > 0$.
Output: Final unadjusted posterior $p_{\epsilon_T}(\theta \mid s_{\text{obs}}) = p\left(\theta \mid \rho(s_{\text{sim}}, s_{\text{obs}}) < \epsilon_T\right)$, and
        its adjusted posterior.
forall $1 \leq t \leq T$ do
    for $i = 1, 2, \cdots, N$ do
        if $t = 1$ then
            Draw particle $\theta_i^{(1)} \sim \pi(\theta)$
        else
            Randomly draw a particle $\theta^* \sim p_{\epsilon_{t-1}}(\theta \mid s_{\text{obs}})$ with a probability equal
              to their corresponding importance weight $W_i^{(t-1)}$, and further
              perturb $\theta_i^{(t)}$ from a MVN perturbation kernel $K_{H^{(t)}}(\cdot \mid \theta^*)$ (with
              optimal bandwidth matrix $H^{(t)}$ defined by equation 5.11) by
              sampling $\theta_i^{(t)}$ such that
                  $$\theta_i^{(t)} \sim K_{H^{(t)}}\left(\theta^{(t)} \mid \theta^*\right)$$
              to obtain a new proposal $\theta_i^{(t)}$ so that $\pi\left(\theta_i^{(t)}\right) > 0$
        end
        Simulate data $y_{\text{sim}} \sim f\left(\cdot \mid \theta_i^{(t)}\right)$
        Compute simulated and observed summary statistics such that
          $s_{\text{sim}} = S(y_{\text{sim}})$ and $s_{\text{obs}} = S(y_{\text{obs}})$
        Calculate weighted distance $d_i^{(t)} = \rho(s_{\text{sim}}, s_{\text{obs}})$ and accept $\theta_i^{(t)}$ if $d_i^{(t)} < \epsilon_t$
          to obtain accepted particles $p_{\epsilon_t}(\theta \mid s_{\text{obs}})$
        Calculate the $j'$th summary statistics weight $w_{j'}^{(t)} = 1/\sigma_{j'}^2$ based on the
          $N_{\epsilon_t} \leq N$ accepted particles; and update summary weight such that
          $w_j^{(t)} = \frac{2}{\frac{1}{w_j^{(t-1)}} + \frac{1}{w_{j'}^{(t)}}}$ (where $\sigma_{j'}^2$ is the variance of the $j'$th summary
          statistics at time $t$), and normalise $w_j^{(t)}$ over all $1 \leq j, j' \leq m$
    end
    if $t = 1$ then
        Set importance weight $W_i^{(1)} = \frac{1}{N}$ for all $1 \leq i \leq N$
    else
        Re-weight the importance weights at $t \neq 1$ for all $1 \leq i \leq N$ by setting
            $$W_i^{(t)} = \pi\left(\theta_i^{(t)}\right) / \sum_{l=1}^{N} W_l^{(t-1)} K_{H^{(t)}}\left(\theta_i^{(t)} \mid \theta_l^{(t-1)}\right),$$
          and normalise $W_i^{(t)}$ over all $1 \leq i \leq N$.
    end
end
Finally, adjust the target posterior $p_{\epsilon_T}(\theta \mid s_{\text{obs}})$ using the modified regression
  adjustment defined in section 5.3.4.2.
```

Remark. It can be inferred from Prangle [244] theoretical work on ABC-SMC convergence that as $t \to \infty$ and $\epsilon_t \to 0$, Algorithm 4 draws approximate samples from the ABC posterior with density

$$p_{\epsilon_t}(\theta \mid s_{\text{obs}}) = \int \left[\frac{f(s_{\text{sim}} \mid \theta)\pi(\theta)\mathbb{1}_{A_{\epsilon_t,s_{\text{obs}}}}}{\int_{\mathbb{R}^n \times \mathbb{R}^m} f(s_{\text{sim}} \mid \theta)\pi(\theta)\mathbb{1}_{A_{\epsilon_t,s_{\text{obs}}}} d\theta ds_{\text{sim}}} \right] ds_{\text{sim}},$$

where $\theta \sim g_t(\theta)$ and $\mathbb{1}_{A_{\epsilon_t,s_{\text{obs}}}}(\cdot) \to \{0,1\}$ is an indicator function of the Lebesgue-measurable set $A_{\epsilon_t,s_{\text{obs}}} = \{s_{\text{sim}} \mid \rho(s_{\text{sim}}, s_{\text{obs}}) \leq \epsilon_t\}$; whereas $\rho(\cdot)$ and $W^{(t)} \propto \frac{\pi(\theta)}{g_t(\theta)}$ are defined by equations 5.9 and 5.12, respectively. Given the MVN perturbation kernel $K_{H^{(t)}}$ with optimal bandwidth $H^{(t)}$ determined by equations 5.10 and 5.11, and for all $1 \leq i \leq N$, the proposal density $g_t(\theta)$ is set such that

$$g_t(\theta) = \begin{cases} \pi(\theta), & \text{if} \quad t = 1 \\ \dfrac{\sum\limits_{i=1}^{N} W_i^{(t-1)} K_{H^{(t)}}\left(\theta | \theta_i^{(t-1)}\right)}{\sum\limits_{i=1}^{N} W_i^{(t-1)}}, & \text{otherwise [244].} \end{cases}$$

5.3.3 Projection of parasite numbers after fish mortality

During realisations when an infected fish or host dies before the end of the observation period in the complex simulation model (developed fully in Chapter 6) or observed empirical data, we use an estimated linear regression function (given by equation 5.18) based on the parasite data prior to host mortality, to linearly project parasite numbers till the end of the observation period. Furthermore, to minimise the projection estimation error, we assign more weight to the most recent outcomes or data prior to host mortality based on equations 5.14–5.16. Thus, the parasite population projection after host mortality is used to aid in the summary statistics computation for ABC fitting of the complex stochastic model (described in section 5.3); specifically, during the computation of other components of the summary statistics such as the log counts of parasites over time till day 17, and weights of the Wasserstein 1-D distance metric between body regions of the host. Below is the newly proposed linear function to estimate missing values after fish mortality till the end of the observation period:

Let y_t= total parasites at time t, $z_t = \log(y_t)$, k be the time before fish mortality, α a regression parameter and $\gamma \in (0,1)$, a tuning parameter which can be optimized. Given $z_1, z_2, z_3, \cdots, z_k$, we want to estimate or project for $\hat{z}_{k+1}, \cdots, \hat{z}_9$. Now, let the proposed least squares regression equation, denoted by $m(t)$, be defined as given by equation 5.14;

$$m(t) = (k-t)\alpha + z_k \tag{5.14}$$

with the estimate of the regression parameter (α) determined such that

$$\hat{\alpha} = \min \sum_t (m(t) - z_t)^2 \gamma^{k-t}. \tag{5.15}$$

Thus, the prediction of z_i for $i = k+1, k+2, \cdots, 9$ can be estimated using equation 5.16

$$\hat{z}_{k+i} = m(k+i) = -i \times \hat{\alpha} + z_k, \tag{5.16}$$

where i is the number of predictions across time. Then,

$$y_{k+i} = \exp(\hat{z}_{k+i}). \tag{5.17}$$

Hence, also letting $y_{t,u}$ be parasite at location u and time t implies the expected projections can be made using equation 5.18; where,

$$\hat{y}_{k+i,u} = \hat{y}_{k+i} \times \frac{y_{k,u}}{y_k}. \tag{5.18}$$

Theorem 5. ***(Least squares estimate of the regression parameter)***
The exact least squares estimate of the regression parameter α based on equation 5.14 is given as:

$$\hat{\alpha} = \frac{\sum_{t=1}^{k-1} (z_t - z_k)(k-t)\gamma^{k-t}}{\sum_{t=1}^{k-1} (k-t)^2 \gamma^{k-t}}. \tag{5.19}$$

Proof of Theorem 5 .

Suppose $m(t) = (k-t)\alpha + z_k$, $\alpha = \sum\limits_t (m(t) - z_t)^2 \gamma^{k-t}$ whilst fixing γ, and the loss function

$$L(\alpha, \gamma) = \min_\alpha \sum_t (m(t) - z_t)^2 \gamma^{k-t}. \tag{5.20}$$

Differentiating equation 5.20 with respect to α and setting to zero gives

$$\frac{\partial L}{\partial \alpha} = 2 \sum_{t=1}^{k-1} \{(k-t)\alpha + z_k - z_t\} \gamma^{k-t} (k-t) = 0.$$

This implies

$$\sum_{t=1}^{k-1} (k-t)^2 \alpha \gamma^{k-t} + \sum_{t=1}^{k-1} (z_k - z_t) \alpha \gamma^{k-t} (k-t) = 0,$$

where

$$\alpha \sum_{t=1}^{k-1} (k-t)^2 \gamma^{k-t} = -\sum_{t=1}^{k-1} (z_k - z_t) \alpha \gamma^{k-t} (k-t).$$

Therefore, solving for α gives the required equation 5.19 such that

$$\hat{\alpha} = -\frac{\sum\limits_{t=1}^{k-1} (z_k - z_t)(k-t)\gamma^{k-t}}{\sum\limits_{t=1}^{k-1} (k-t)^2 \gamma^{k-t}} = \frac{\sum\limits_{t=1}^{k-1} (z_t - z_k)(k-t)\gamma^{k-t}}{\sum\limits_{t=1}^{k-1} (k-t)^2 \gamma^{k-t}}, \gamma \in (0,1). \qquad \text{Q. E. D.}$$

5.3.4 Weighted Ridge Regression for posterior adjustment

5.3.4.1 Introduction

The section presents the extended local-linear regression (previously discussed in section 5.2.3.3) for ABC post-analysis to improve further the posterior mean estimates by adjusting the ABC posterior distribution (for a population-based model with population size of $M > 1$, where the underlying summary statistics takes the form of a matrix). This regression mean adjustment method is a robust extension to Beaumont et al. [27] method (when $M = 1$ or the summary statistics are assumed to be one-dimensional) by considering heteroscedastic errors with $L2$ regularisation (in the neighbourhood of the observed summary statistics s_{obs}) to deal with potential problems of multicollinearity, supercollinearity, non-normal errors and outliers that may exist in the standard local-linear regression model for ABC post-processing analysis. Section 5.3.4.2 outlines the modified regression model for ABC posterior mean adjustment based on a weighted ridge regression (WRID) initially developed by Arkin et al. [10] for general regression problems. In the proposed regression adjustment model (presented in section 5.3.4.2), the dependent variables represent the posterior samples or approximate posterior distribution of the model parameters (on a logarithmic scale) from the modified ABC-SMC algorithm; and their predictors are the corresponding simulated summary statistics in the neighbourhood of the observed summaries. The proposed regression adjustment method can be adapted and compared to Beaumont et al. [27] local-linear regression for general modelling problems which are not population-based (e.g., if $M = 1$) or the summary statistics are assumed to one-dimensional.

5.3.4.2 Proposed ABC posterior mean adjustment

Given a set of η unadjusted posterior samples from the weighted-iterative ABC algorithm (described by Algorithm 4), let $\theta_i^{(r)}$ be the ith posterior sample (for $i = 1, 2, \cdots, \eta$) for the rth model parameter (for $r = 1, 2, \cdots, n$). Suppose $s_{\text{sim},i}$ are the accepted simulated summary statistics (with dimension $M \times m$) corresponding to the ith posterior sample; where the $M \geq 1$ corresponds to a population size, and $m \geq 1$ the number of summary statistics for each individual in the population model (to be simulated). The regression model in the vicinity of the observed summary statistics s_{obs} (with dimension $M \times m$) is given as

$$\theta_i^{(r)} = \alpha^{(r)} + \bar{\mathcal{S}}_i^\top \beta^{(r)} + \varsigma_i^{(r)}, \quad 1 \leq i \leq \eta \quad \text{and} \quad 1 \leq r \leq n \tag{5.21}$$

where $\bar{\mathcal{S}}_i = \frac{1}{M} \sum_{k=1}^{M} \left[s_{\text{sim}_{(k,m)},i} - s_{\text{obs}_{(k,m)}} \right]$ is an m-dimensional vector of mean differences between $s_{\text{sim},i}$ and s_{obs} across all M individuals for the ith posterior sample; $\alpha^{(r)}$ is the intercept (whose estimate represent the required adjusted posterior mean), $\beta^{(r)}$ is a vector of regression coefficients corresponding to the m predictors (in the neighbourhood of

s_{obs}), and $\varsigma_i^{(r)}$ are the regression error terms with mean 0 and heteroscedastic variance, corresponding to the rth model parameter. If $M = 1$, $\bar{\mathcal{S}}_i = s_{\text{sim},i} - s_{\text{obs}}$ as in the case of Beaumont et al. [27] regression adjustment methods (where $s_{\text{sim},i}$ and s_{obs} are assumed to a one-dimensional array or vector of length m, respectively).

Given equation 5.21, the robust weighted ridge regression estimates of $\left(\alpha^{(r)}, \beta^{(r)}\right)$ are derived by minimising the loss function $\mathcal{L}_{\text{ridge}}^{(r)}$ for each rth model parameter such that

$$\mathcal{L}_{\text{ridge}}^{(r)} = \sum_{i=1}^{\eta} \left\{ \theta_i^{(r)} - \alpha^{(r)} - \sum_{j=1}^{m} \bar{\mathcal{S}}_{i,j} \beta_j^{(r)} \right\}^2 K_\delta(\|s_{\text{sim},i} - s_{\text{obs}}\|) + \lambda \left\|\beta^{(r)}\right\|_2^2; \tag{5.22}$$

where $K_\delta(\cdot)$ is a Gaussian kernel with bandwidth or scale parameter δ given as

$$K_\delta(\|s_{\text{sim},i} - s_{\text{obs}}\|) = \omega_i = \frac{1}{\sqrt{2\pi\delta}} \mathrm{e}^{\frac{-1}{2\delta^2} \|s_{\text{sim},i} - s_{\text{obs}}\|^2}, \tag{5.23}$$

and $\|s_{\text{sim},i} - s_{\text{obs}}\| = \rho(s_{\text{sim},i}, s_{\text{obs}})$ is the weighted distance (computed using equation 5.9) between $s_{\text{sim},i}$ and s_{obs}; and the penalty term $\lambda \left\|\beta^{(r)}\right\|_2^2 = \lambda \sum_{j=1}^{m} \beta_j^{(r)2}$ is the $L2$ regularisation element, with λ representing the biasing or penalty parameter. To solve equation 5.22, we rewrite it using the transformed variables given by equation 5.24. The estimates of $\beta^{(r)}$ and $\alpha^{(r)}$ are obtained separately (by initially ignoring the intercept $\alpha^{(r)}$ in equation 5.21) since the predictors and the dependent variables are respectively mean centred and re-scaled using $\sqrt{\omega_i}$ to obtain set of variables with similar scaling (where the latter is motivated by [220]); such that for $1 \leq i \leq \eta$ and $1 \leq j \leq m$:

$$\theta_i^{(r)*} = \sqrt{\omega_i}\left(\theta_i^{(r)} - \bar{\theta}^{(r)}\right) \qquad \text{and} \qquad \bar{\mathcal{S}}_{ij}^* = \sqrt{\omega_i}\left(\bar{\mathcal{S}}_{ij} - \bar{\bar{\mathcal{S}}}_j\right), \tag{5.24}$$

where $\bar{\theta}^{(r)}$ is the weighted mean of $\theta_i^{(r)}$, and $\bar{\bar{\mathcal{S}}}_j$ is the weighted mean of the jth predictor. To obtain an expression for the intercept $\alpha^{(r)}$ in equation 5.21, we rely on Theorem 6 by reverse transformation of $\theta_i^{(r)*}$ and $\bar{\mathcal{S}}_{ij}^*$ into their respective original scales after model fitting. The reason for the use of the re-scaled variables is that since ridge regression regularises the linear regression by imposing a penalty based on the size or magnitude of the regression coefficients, it requires the variables (predictors and posterior samples) to have similar measurement scales in order to assess their contributions to the penalised term fairly, while maintaining the information content of the variables after re-scaling. Hence, equation 5.22 is transformed (without the intercept) such that

$$\begin{aligned}
\mathcal{L}^{(r)*}_{\text{ridge}} &= \sum_{i=1}^{\eta} \left\{ \left[\sqrt{\omega_i} \left(\theta_i^{(r)} - \bar{\theta}^{(r)} \right) \right] - \sum_{j=1}^{m} \left[\sqrt{\omega_i} \left(\bar{\mathcal{S}}_{ij} - \bar{\bar{\mathcal{S}}}_j \right) \right] \beta_j^{(r)*} \right\}^2 \omega_i + \lambda \left\| \beta^{(r)*} \right\|_2^2 \\
&= \sum_{i=1}^{\eta} \left\{ \theta_i^{(r)*} - \sum_{j=1}^{m} \bar{\mathcal{S}}^{*}_{i,j} \beta_j^{(r)*} \right\}^2 \omega_i + \lambda \left\| \beta^{(r)*} \right\|_2^2,
\end{aligned} \tag{5.25}$$

where $\beta_j^{(r)*}$ are the regression coefficient corresponding to the scaled predictors. The estimate of $\beta_j^{(r)*}$ which minimises the loss function given by equation 5.25 such that

$$\hat{\beta}^{(r)*} = \underset{\beta^{(r)*} \in \mathbb{R}^m}{\arg\min} \, \mathcal{L}^{(r)*}_{\text{ridge}},$$

is given as

$$\hat{\beta}^{(r)*}_{m\times 1} = (X^{\top}_{m\times\eta} W_{\eta\times\eta} X_{\eta\times m} + \lambda I_{m\times m})^{-1} X^{\top}_{m\times\eta} W_{\eta\times\eta} \theta^{(r)*}_{\eta\times 1} \quad 1 \le r \le n; \tag{5.26}$$

where $I_{m\times m}$ is an $m \times m$ identity matrix, W is a diagonal weighting matrix with the ith diagonal element given by

$$\omega_i = W_{ii} = K_{\delta}(\|s_{\text{sim},i} - s_{\text{obs}}\|), \quad 1 \le i \le \eta,$$

$$X = \begin{bmatrix} \bar{\mathcal{S}}^{*}_{1,1} & \bar{\mathcal{S}}^{*}_{1,2} & \cdots & \bar{\mathcal{S}}^{*}_{1,m} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{\mathcal{S}}^{*}_{\eta,1} & \bar{\mathcal{S}}^{*}_{\eta,2} & \cdots & \bar{\mathcal{S}}^{*}_{\eta,m} \end{bmatrix}, \quad \theta^{(r)*} = \begin{bmatrix} \theta_1^{(r)*} \\ \vdots \\ \theta_{\eta}^{(r)*} \end{bmatrix}, \quad \bar{\theta}^{(r)} = \frac{\sum\limits_{i=1}^{\eta} \omega_i \theta_i^{(r)}}{\sum\limits_{i=1}^{\eta} \omega_i}, \quad \text{and} \quad \bar{\bar{\mathcal{S}}}_j = \frac{\sum\limits_{i=1}^{\eta} \omega_i \bar{\mathcal{S}}_{ij}}{\sum\limits_{i=1}^{\eta} \omega_i}.$$

Theorem 6. *Let suppose a weighted ridge regression model which passes through the origin such that*

$$Y_i^* = \beta_1^* X_{i1}^* + \beta_2^* X_{i2}^* + \cdots + \beta_m^* X_{im}^* + \epsilon_i^*, \qquad i = 1, 2, \cdots, \eta \tag{5.27}$$

where Y^* is the dependent variable, β_j^* denote the corresponding regression coefficient of the jth predictor (X_j^*, $1 \le j \le m$), ϵ_i^* denote the errors with mean 0 (and heteroscedastic variance), and η is the sample size. Let also assume the dependent variable (Y^*) and design matrix (X^*) are in their standardised form such that

$$Y_i^* = \sqrt{\omega_i}(Y_i - \bar{Y}) \qquad \textit{and} \qquad X_{ij}^* = \sqrt{\omega_i}(X_{ij} - \bar{X}_j),$$

where Y is the unstandardised dependent variable, X is unstandardised design matrix, $\bar{Y}$ is the weighted mean of Y, $\bar{X}_j$ is the weighted mean corresponding to the jth predictor (X_j), and ω_i is a weight corresponding to ith sample. Suppose that the corresponding fitted regression model of equation 5.27 is given as

$$\hat{Y}_i^* = \hat{\beta}_1^* X_{i1}^* + \hat{\beta}_2^* X_{i2}^* + \cdots + \hat{\beta}_m^* X_{im}^*, \tag{5.28}$$

where the estimated regression coefficients, $\hat{\beta}^* = \underset{\beta^* \in \mathbb{R}^m}{\arg\min} \sum_{i=1}^{\eta} \left\{ Y_i^* - \sum_{j=1}^{m} X_{i,j}^* \beta_j^* \right\}^2 \omega_i + \lambda \|\beta^*\|_2^2$ (with λ denoting the ridge penalty parameter). Then, reverting the variables in equation 5.28 to their original scales (Y and X) result in a fitted linear model with an estimated intercept term defined as

$$\hat{\alpha} = \bar{Y} - \sum_{j=1}^{m} \hat{\beta}_j^* \bar{X}_j. \tag{5.29}$$

Proof of Theorem 6 .

Let suppose the weighted ridge regression model (given by equation 5.27) is fitted such that equation 5.28 holds, where $\hat{Y}_i^* = \hat{\beta}_1^* X_{i1}^* + \hat{\beta}_2^* X_{i2}^* + \cdots + \hat{\beta}_m^* X_{im}^*, \qquad 1 \le i \le \eta$. Given that $\hat{Y}_i^* = \sqrt{\omega_i}(\hat{Y}_i - \bar{Y})$ and $X_{ij}^* = \sqrt{\omega_i}(X_{ij} - \bar{X}_j)$ implies that for $1 \le i \le \eta$,

$$\sqrt{\omega_i}(\hat{Y}_i - \bar{Y}) = \beta_1^* \left[\sqrt{\omega_i}(X_{i1} - \bar{X}_1)\right] + \beta_2^* \left[\sqrt{\omega_i}(X_{i2} - \bar{X}_2)\right] + \cdots + \beta_m^* \left[\sqrt{\omega_i}(X_{im} - \bar{X}_m)\right]. \tag{5.30}$$

Now, multiplying $\frac{1}{\sqrt{\omega_i}}$ to both sides of equation 5.30 and applying a simple algebraic re-arrangement gives

$$\begin{aligned}
\hat{Y}_i &= \bar{Y} + \hat{\beta}_1^* \left(X_{i1} - \bar{X}_1\right) + \hat{\beta}_2^* \left(X_{i2} - \bar{X}_2\right) + \cdots + \hat{\beta}_m^* \left(X_{im} - \bar{X}_m\right) \\
&= \bar{Y} - \hat{\beta}_1^* \bar{X}_1 - \hat{\beta}_2^* \bar{X}_2 - \cdots - \hat{\beta}_m^* \bar{X}_m + \hat{\beta}_1^* X_{i1} + \hat{\beta}_2^* X_{i2} + \cdots + \hat{\beta}_m^* X_{im} \\
&= \bar{Y} - \sum_{j=1}^{m} \hat{\beta}_j^* \bar{X}_j + \sum_{j=1}^{m} \hat{\beta}_j^* X_{ij} \\
&= \hat{\alpha} + \sum_{j=1}^{m} \hat{\beta}_j^* X_{ij}.
\end{aligned} \tag{5.31}$$

Hence, equation 5.31 corresponds to a fitted linear model with an estimated intercept expressed as

$$\hat{\alpha} = \bar{Y} - \sum_{j=1}^{m} \hat{\beta}_j^* \bar{X}_j \qquad \text{Q. E. D.}$$

Our proposed Theorem 6 strictly focuses on how we can derive an expression for the estimate of the intercept, $\alpha^{(r)}$, after fitting the regression model (equation 5.21) without the intercept term by minimising the loss function (given by equation 5.25) with a ridge penalty based on the transformed variables defined in equation 5.24. This gives us an estimate of $\alpha^{(r)}$ in our modified local-linear regression for each rth model parameter as

$$\hat{\alpha}^{(r)} = \bar{\theta}^{(r)} - \sum_{j=1}^{m} \hat{\beta}_j^{(r)*} \bar{X}_j, \tag{5.32}$$

where $\bar{X}_j = \frac{\sum\limits_{i=1}^{\eta} \omega_i X_{ij}}{\sum\limits_{i=1}^{\eta} \omega_i}$, $\bar{\theta}^{(r)}$ is the weighted mean of $\theta^{(r)}$ and $\hat{\beta}_j^{(r)*}$ is the estimate of the regression coefficient corresponding to the jth transformed predictor. $\hat{\alpha}^{(r)}$ is a quantity denoting the adjusted posterior means on a logarithmic scale in the current study (since our unadjusted posterior samples were on a logarithmic scale). Hence, the required posterior mean adjustment of the rth model parameter is estimated by taking inverse of its logarithmic form (given by equation 5.32) such that

$$\hat{\alpha}^{(r)}_{\text{adjust}} = \mathrm{e}^{\hat{\alpha}^{(r)}}, \quad r = 1, 2, \cdots, n. \tag{5.33}$$

It is imperative to note that the exponential transform of the estimate of $\hat{\alpha}^{(r)}$ in equation 5.33 holds since the current study assumes the unadjusted posterior samples were obtained on a logarithmic scale. An exponential transformation is unnecessary for other studies where the unadjusted posterior samples were obtained based on their original scales. In addition, the adjusted posterior distribution $\theta^{(r)}_{\text{adjust}}$ (on logarithmic scale) for the rth model parameter is derived from equation 5.34 such that

$$\theta^{(r)}_{\text{adjust},i} = \theta^{(r)}_{i} - \sum_{j=1}^{m} \hat{\beta}^{(r)*}_{j} \bar{\mathcal{S}}_{ij}, \quad i = 1, 2, \cdots, \eta. \tag{5.34}$$

Remark. The glmnet package in R [139] was used to obtain the optimal value of the penalty parameter λ via cross-validation (among a range of values from 0.01 to 100) that achieve the least predictive error before posterior adjustments. Also, optimal value of the bandwidth or smoothing parameter δ of the Gaussian kernel $K_{\delta}(\cdot)$ (given by equation 5.23) was adaptively estimated (based on the weighted distances between the simulated and observed summary statistics) via a cross-validation procedure (which minimises the asymptotic mean integrated squared error) using the kedd package in R [120]. In this study, 95% credible intervals of posterior mean estimates were estimated based on the Equal-Tailed Interval (ETI) of posterior distributions using the bayestestR package in R [204].

Before fitting the complex simulation model (described in Chapter 6), the proposed regression adjustment methodology, as well as the weighted-iterative ABC with SMC and SIS, were first tested based on a simple modelling problem (with multivariate normal likelihood) where the exact posterior distribution is known (using conjugate priors for a multivariate normal distribution with unknown mean vector and known covariance matrix). For the simple numerical experiment (using an artificial data whose true model parameters are known), the robustness of posterior approximations from the weighted-iterative ABC at different draws ($N = 500$, 1000, 2000, 3000, 4000 and 5000) from the proposal distribution (on a logarithmic scale) and the pre-specified set of decreasing tolerances are also assessed (see section 5.3.5). It helped to justify that once the decreasing tolerance thresholds and adaptive importance weights are appropriately set up in the weighted-iterative ABC algorithm with SMC and SIS, the number of particles drawn from the proposal distribution (N) does not significantly affect the fidelity of the target posterior. In other words, it will help determine whether the resulting approximate posterior is independent of N or result in mutually compatible approximations at different values of N.

Consequently, these numerical experiments (based on a toy model with a multivariate normal likelihood function) will also help determine the minimal number of draws from the proposal or prior distribution needed to obtain good posterior approximations when calibrating the complex stochastic simulation model due to the potentially high computational costs of simulating from the stochastic model as well as fitting the model via sequential Monte Carlo ABC (whose computational cost increases quadratically as a function of N). All the results from the numerical experiments (based on the toy model), including the ABC results, are presented in section 5.3.5. Results on ABC fitting of the complex stochastic simulation model are presented in section 6.3. For more detailed R codes of the modified ABC algorithm and the proposed ABC post-processing methodology (including other supporting functions), see Appendix F.

5.3.5 Assessing the modified ABC and regression adjustment using a numerical experiment

5.3.5.1 Introduction

Section 5.3.5 presents a simple numerical experiment with multivariate normal likelihood function (where the presumed true model parameter values and the exact form of the true posterior are known) to assess the modified ABC approximations (using Algorithm 4) and the proposed ABC posterior adjustment methodology (described in section 5.3.4.2) at a different number of proposal draws (i.e., $N = 500$, 1000, 2000, 3000, 4000, and 5000 samples); where we further explore whether the resulting approximated posterior is independent of N, and mutually compatible ABC approximations are achieved at the different values of N). Here, the modified regression adjustment is also compared with the standard local-linear regression adjustment with heteroscedastic errors proposed by Beaumont et al. [27]. Findings from the numerical experiments are also used to determine the minimal number of proposal draws needed when fitting the complex stochastic simulation model (formally described in Chapter 6) due to the high computational costs involved in i) model simulation, ii) estimation of the multidimensional summary statistics for the entire host population (especially the summary component which estimates the B-D-C model parameters during realisations of parasite population explosion as discovered in section 4.2), and iii) implementing sequential Monte Carlo ABC methods (whose computational cost increases quadratically as a function of N).

5.3.5.2 Description of the toy model and modelling problem

For the numerical experiment based a toy model defined below, artificial multivariate data $\mathbf{X} \sim \mathcal{N}_k(\boldsymbol{\theta}, \Sigma)$ was simulated from a multivariate normal (MVN) distribution for a k-dimensional random variables $\mathbf{X} = (X_1, X_2, \cdots, X_k)^\top$ with k-dimensional mean vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_k)^\top$ and $k \times k$ covariance matrix Σ. For simplicity, we set $k = 6$ and assume that the true mean vector (or population mean) of $\mathbf{X}$ (in the toy model) is $\boldsymbol{\theta} = (0.5, 1.0, 1.5, 2.0, 2.5, 3.0)^\top$ with a positive-definite symmetric covariance matrix $\Sigma = \text{Var}(\mathbf{X})$ (which was randomly generated in R for the toy model) also known. Specifically, we randomly assumed that

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{pmatrix} \sim \mathcal{N}_6 \left[\begin{pmatrix} 0.5 \\ 1.0 \\ 1.5 \\ 2.0 \\ 2.5 \\ 3.0 \end{pmatrix}, \begin{pmatrix} 24.5134 & 11.6042 & 8.2851 & 15.6787 & 19.6029 & 18.4657 \\ 11.6042 & 36.1535 & 16.9813 & 9.1931 & 12.6557 & 33.0837 \\ 8.2851 & 16.9813 & 24.7937 & 5.0379 & 18.2924 & 16.5758 \\ 15.6787 & 9.1931 & 5.0379 & 16.1338 & 11.7926 & 9.8223 \\ 19.6029 & 12.6557 & 18.2924 & 11.7926 & 35.7758 & 18.0006 \\ 18.4657 & 33.0837 & 16.5758 & 9.8223 & 18.0006 & 35.2091 \end{pmatrix} \right]. \tag{5.35}$$

Toy model:

Let suppose $\mathbf{X}$ (with n number of observations) is a multivariate data randomly generated from a 6-dimensional MVN distribution with density function $f_{\mathbf{X}}(\cdot \mid \boldsymbol{\theta}, \Sigma)$ given by equation 5.36, where the population mean vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_6)^\top$ (with 6 unknown model parameters) and known covariance matrix Σ (as specified in equation 5.35); such that

$$f_{\mathbf{X}}\left(X_1, X_2, \cdots, X_6; \boldsymbol{\theta}, \Sigma\right) = \frac{1}{\sqrt{(2\pi)^6 \left(\det \Sigma\right)}} \exp\left\{-\frac{1}{2}\left(\mathbf{X} - \boldsymbol{\theta}\right)^\top \Sigma^{-1} \left(\mathbf{X} - \boldsymbol{\theta}\right)\right\}. \qquad (5.36)$$

Assuming that $\boldsymbol{\theta} \in \mathbb{R}^6$ is also a random variable, let suppose the prior distribution of $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta}) \propto \mathcal{N}_6\left(\boldsymbol{\mu}_0, \Sigma_0\right)$ is MVN with mean $\boldsymbol{\mu}_0$ and covariance matrix Σ_0.

Modelling problem:

Given the simulated MVN data $\mathbf{X}$ (whose population mean vector is assumed to be unknown), and the MVN prior density $\pi(\boldsymbol{\theta}) \propto \mathcal{N}_6\left(\boldsymbol{\mu}_0, \Sigma_0\right)$; we want to estimate the posterior predictive distribution $p(\boldsymbol{\theta} \mid \mathbf{X}, \Sigma)$ using the weighted-iterative ABC algorithm (outlined in Algorithm 4) as well as perform regression adjustment using both the proposed posterior correction method (defined in section 5.3.4.2) and standard local-linear regression adjustment with heteroscedastic errors proposed by Beaumont et al. [27]. Here, we assume that the true likelihood $f(\mathbf{X} \mid \boldsymbol{\theta}, \Sigma)$ is unknown (for the sake of ABC fitting and assessment). The accuracy of the posterior estimates from the ABC fitting and ABC post-processing analyses (based on the toy model) at different values of N (where N is the number of proposal draws) are also compared to the true hyperparameter posterior mean estimator of $\boldsymbol{\theta} \in \mathbb{R}^6$ defined in accordance with Lemma 4 and the true parameter values using some standard accuracy measures (i.e., the bias, variance and mean square error of the posterior estimates as well as their corresponding 95% credible intervals).

Lemma 4. Suppose that $\mathbf{X} \sim \mathcal{N}_k\left(\boldsymbol{\theta}, \Sigma\right)$ is a multivariate data (with sample size of n) generated from a MVN distribution with unknown mean vector $\boldsymbol{\theta} \in \mathbb{R}^k$ and known covariance matrix Σ. Let assume that the prior distribution of $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta}) \propto \mathcal{N}_k\left(\boldsymbol{\mu}_0, \Sigma_0\right)$ is multivariate normal with mean $\boldsymbol{\mu}_0 \in \mathbb{R}^k$ and covariance matrix Σ_0. Then, given the MVN data $\mathbf{X}$, the resulting posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{X}, \Sigma)$ and prior $\pi(\boldsymbol{\theta})$ are conjugate distributions; such that $p(\boldsymbol{\theta} \mid \mathbf{X}, \Sigma) \propto \mathcal{N}_k\left(\hat{\boldsymbol{\theta}}_n, \Sigma_n\right)$ is MVN with the exact posterior hyperparameter mean estimator given as

$$\hat{\boldsymbol{\theta}}_n = \Sigma_n\left(\Sigma_0^{-1}\boldsymbol{\mu}_0 + n\Sigma^{-1}\bar{\mathbf{X}}\right), \tag{5.37}$$

where the covariance matrix $\Sigma_n = (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}$, $\bar{\mathbf{X}} \in \mathbb{R}^k$ is the sample mean vector, and n is the sample size of the observed data [223].

Remark. For the numerical experiment, pseudo-observed data with a sample size of $n = 1000$ was simulated (in R software) from MVN with true mean and covariance matrix specified per equation 5.35 (where the true population mean was already known). This pseudo-observed data was considered as the observed data to be used for ABC fitting, where we assumed the true parameter mean vector $\boldsymbol{\theta} \in \mathbb{R}^6$ of the pseudo-observed data is unknown with 6 model parameters (as a form of an inverse problem), but the covariance matrix Σ is known. For simplicity, we also set the covariance matrix of the prior $\Sigma_0 = \Sigma$ (since Σ is not to be estimated, and thus, not of interest). We further assume that the prior mean vector $\boldsymbol{\mu}_0 \in \mathbb{R}^6$ is the only hyperparameter to consider in the prior distribution $\pi(\boldsymbol{\theta})$ during the ABC analysis and evaluation of the exact posterior hyperparameter mean estimator given by equation 5.37 under Lemma 4. Additionally, since the sample mean vector $\bar{\mathbf{X}}$ is known to be a sufficient summary statistics for the population mean of MVN distribution (with known covariance matrix), the sample mean vector was considered as summary statistics for both pseudo-observed and simulated data from the toy model during ABC fitting (based on the prior or proposal samples). The weighted Euclidean distance metric (given by equation 5.4) was used as the discrepancy measure in the proposed weighted-iterative ABC algorithm to compare between the pseudo-observed and simulated data. The main results from the numerical experiment (based on the toy model) are presented in section 5.3.5.3.

5.3.5.3 Summary of results from the numerical experiment

The MVN model (with 6 model parameters described in section 5.3.5.2) was fitted using the modified weighted-iterative ABC with sequential Monte Carlo and adaptive importance sampling (given by Algorithm 4). The modified ABC algorithm was set-up to have a fixed number of iterations or time steps at $T = 10$ (i.e., a total of 10 time steps), and a set of monotonically decreasing tolerances (ϵ_t, $t = 1, 2, \cdots, 10$) at each ABC time step t was carefully pre-specified based on the total number of proposal draws or prior samples (N) according to the following: if $N < 1000$, $\epsilon_t = 0.5$, 0.43, 0.4, 0.35, 0.3, 0.2, 0.1, 0.08, 0.06, 0.02; whereas if $N \geq 1000$, $\epsilon_t = 0.5$, 0.3, 0.2, 0.1, 0.08, 0.07, 0.06, 0.03, 0.02, 0.01. To examine the robustness of the modified weighted-iterative ABC based on the choice of N and pre-specified tolerances, the ABC fitting of the toy model was done at different values of N: $N = 500$, 1000, 2000, 3000, 4000, and 5000, respectively; and the resulting posterior distributions were further adjusted to estimate the posterior mean of the model parameter $\boldsymbol{\theta}$ using the proposed posterior correction method (defined in section 5.3.4.2) and standard local-linear regression adjustment with heteroscedastic errors proposed by Beaumont et al. [27] (for comparison purposes) across the 6 model parameters.

Figure 5.1 is a comparative plot showing the variability in the unadjusted posterior mean estimates of the model parameters with their respective 95% credible intervals at the different values of N. Figure 5.1 shows that the ABC approximations from the weighted-iterative ABC with SMC and SIS (Algorithm 4) resulted in mutually compatible approximations at the different values of N based on the pre-specified tolerance and ABC time steps. Thus, it can be inferred (from Figure 5.1) that the resulting posterior from the modified ABC-SMC algorithm is independent of N (for $500 \leq N \leq 5000$), and the degree of variability in the posterior distributions are not significantly different irrespective of the number of proposal draws from the importance distribution with density g defined by equation 5.13 (from $N = 500$ to $N = 5000$). Nonetheless, Figure 5.2 indicates that the computational cost (or cost in time) increases quadratically as the number of proposal draws increases from $N = 500$ to $N = 5000$ during ABC fitting of the toy simulation model (run in parallel using over 20 CPU cores of a multi-core processor). Based on Figures 5.1 and 5.2, it will be cost-effective to fit the complex stochastic simulation (described in section 6.2.1) at $N = 500$ due to the high computational cost associated with model simulation from the complex model (especially during realisations of parasite population explosion), estimation of the multidimensional summary statistics (across the entire host population) and the potential cost of implementing the modified ABC algorithm at higher values of $N \gg 500$.

After ABC fitting of the toy model, the approximate posterior mean was estimated and its posterior distribution adjusted using the modified regression adjustment with $L2$ regularisation and the standard local-linear regression adjustment (with heteroscedastic errors) at the different values of N. Figures 5.3–5.8 are goodness-of-fit density plots at the different values of N, which graphically show the unadjusted and adjusted posterior distributions against the prior distributions (of the fitted toy simulation model). It can be seen from the goodness-of-fit density plots that the unadjusted posterior based on the modified ABC algorithm as well as the adjusted posterior using the modified regression adjustment (with heteroscedastic errors and $L2$ regularisation) performed well when compared to the prior distribution across all model parameters at the different values of N. However, at $1000 \leq N \leq 4000$, the posterior adjustments from Beaumont et al. [27] standard local-linear regression (with heteroscedastic errors) resulted in very flat and poorly adjusted for a few model parameters (relative to the prior distribution). Hence, the modified regression adjustment, which can deal with potential multicollinearity in the regression predictors (in the neighbourhood of the observed summaries), supercollinearity and shrink regression coefficients of predictors with less contribution, appears to be more robust in adjusting the posterior distribution than the standard correction method. At certain simulation realisations with very high multicollinearity, the standard local-linear regression could not be implemented at all or performed poorly since the design matrix X or the term $X^{\top}WX$ (where W was the diagonal weighting matrix) was either singular or close to being singular. The unadjusted and adjusted posterior means were computed and compared with the true parameter values and the true posterior mean estimates at the different values of N.

Additionally, the bias, variance, and mean square error (MSE) of the posterior mean estimates, as well as their corresponding 95% credible intervals, were estimated to compare the performance of approximations from the weighted-iterative ABC and the two regression adjustments numerically (Tables 5.1–5.3). Generally, there was no significant difference in the degree of accuracy between the unadjusted and adjusted posterior mean estimates at the different values of N; and the true posterior mean estimates (based on equation 5.37) was found in their respective estimated credible intervals. However, the MSE of the adjusted posterior mean based on the proposed regression correction (with $L2$ regularisation) resulted in relatively smaller MSE and credible interval width most of the time (especially at $N \leq 1000$). Hence, it can be inferred from Tables 5.1–5.3 that the proposed regression adjustment is relatively robust in estimating the posterior mean compared to the standard local-linear regression of Beaumont et al. [27]. In addition, it can further be adapted to estimate the posterior mean after ABC fitting of the complex stochastic simulation model in the presence of high multicollinearity (since the multidimensional summary statistics to be used for calibrating the complex model appears to be highly correlated). Also, since the number of predictor variables exceed the number of posterior samples at $N = 500$ (a condition which results in suppercollinearity), the proposed regression adjustment will be more suitable for posterior correction. Hence, the standard local-linear regression may not be possible to be implemented or result in incorrect adjustments in these aforementioned instances.

Remark. Based on findings from the numerical experiment, it is recommended to fit the complex simulation model based on $N = 500$ proposal draws from the importance density, using the weighted-iterative ABC algorithm since it will be cost-effective to calibrate the complex model at $N = 500$; whereas the adjusted posterior of the resulting ABC posterior (based on the modified ABC regression correction) is considered for subsequent analyses including hypotheses testing.

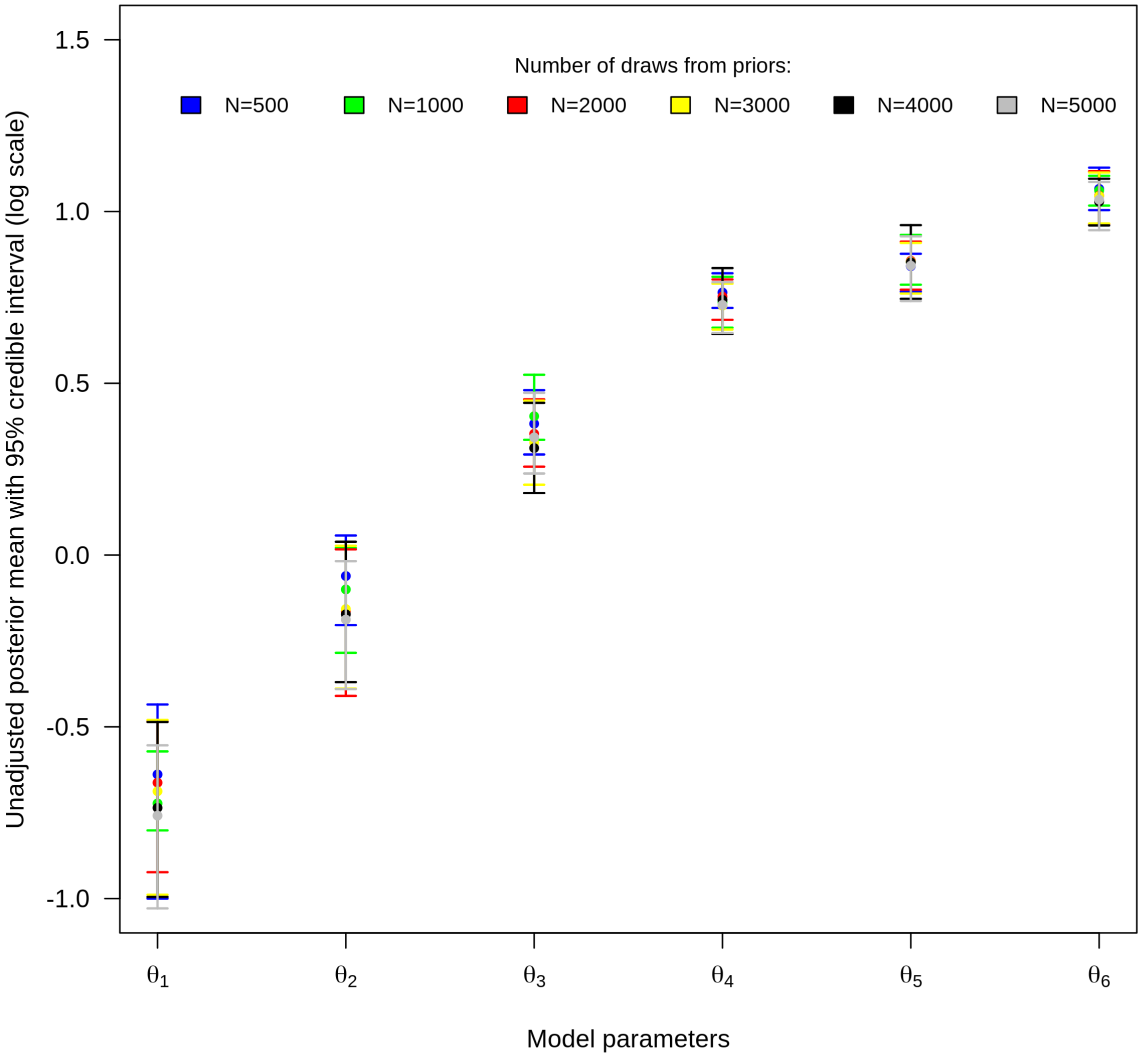

Figure 5.1: Comparative plot of the unadjusted posterior mean estimates of the toy model parameters with their respective 95% credible intervals (on logarithmic scale) at different values of N ($N = 500$, 1000, 2000, 3000, 4000, and 5000 proposal samples).

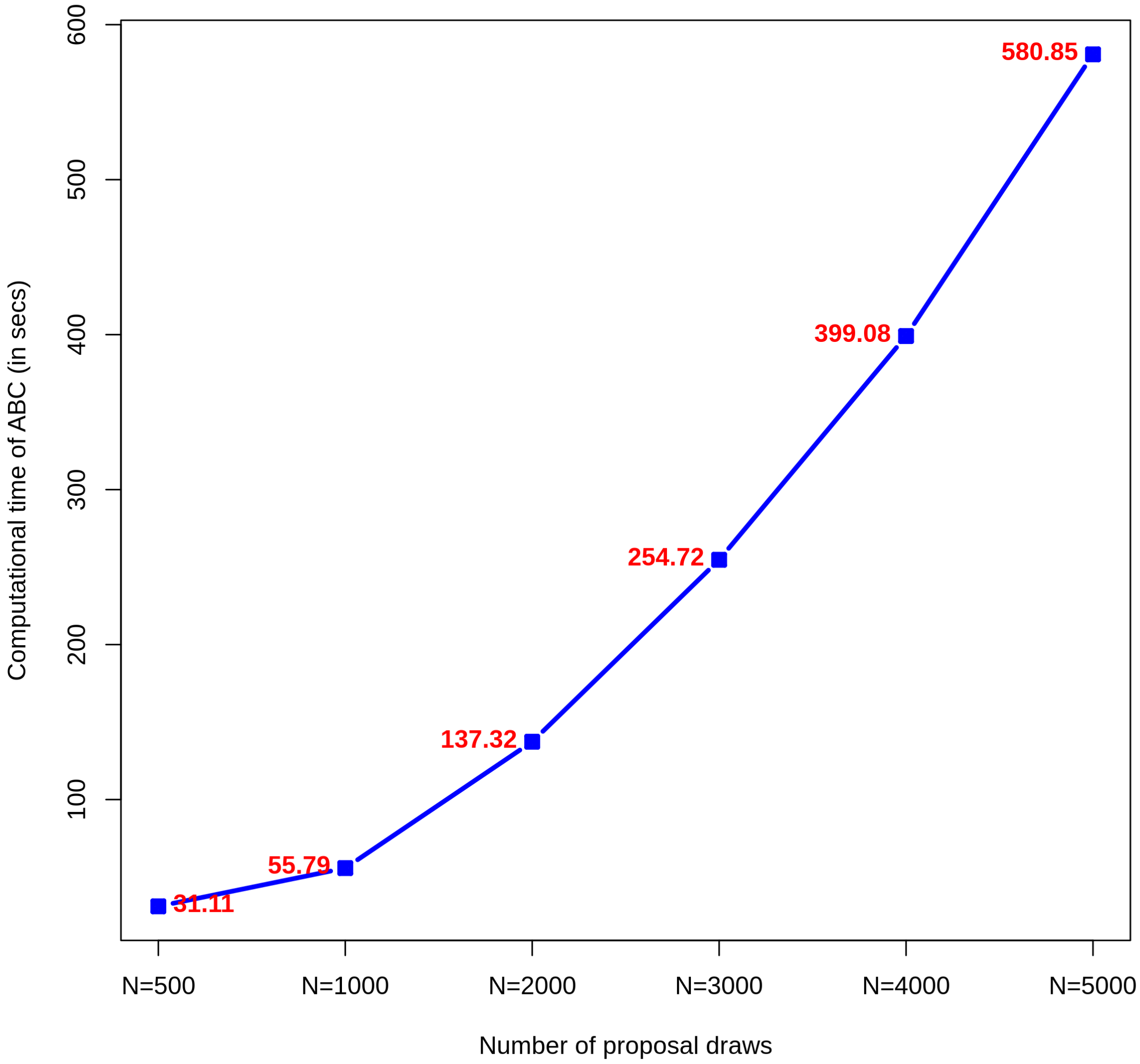

Figure 5.2: Computational times (in secs) of ABC fitting of the toy model at different values of N ($N = 500$, 1000, 2000, 3000, 4000, and 5000 proposal samples).

Table 5.1: Comparison between unadjusted ABC posterior mean estimates ($\hat{\boldsymbol{\theta}}_{\text{unadj}}$), true parameter values ($\hat{\boldsymbol{\theta}}_0$) and conjugate posterior mean estimates ($\hat{\boldsymbol{\theta}}_{\text{conjugate}}$) of the 6 parameters of the toy model across different values of N (from $N = 500$ to $N = 5000$).

Parameters	$\hat{\boldsymbol{\theta}}_{\text{unadj}}$	$\hat{\boldsymbol{\theta}}_0$	$\hat{\boldsymbol{\theta}}_{\text{conjugate}}$	bias($\hat{\boldsymbol{\theta}}_{\text{unadj}}$)	Var($\hat{\boldsymbol{\theta}}_{\text{unadj}}$)	MSE($\hat{\boldsymbol{\theta}}_{\text{unadj}}$)	95% Cred. Int.
N=500							
θ_1	0.5280	0.5	0.5063	0.02802	0.0086	0.0094	0.3690—0.6477
θ_2	0.9408	1.0	0.8980	-0.0592	0.0065	0.0100	0.8162—1.0587
θ_3	1.4656	1.5	1.4664	-0.0344	0.0099	0.0111	1.3403—1.6164
θ_4	2.1475	2.0	2.1252	0.1475	0.0062	0.0279	2.0530—2.2712
θ_5	2.3186	2.5	2.4327	-0.1814	0.0083	0.0412	2.1536—2.4037
θ_6	2.9038	3.0	2.9467	-0.0962	0.0162	0.0255	2.7292—3.0889
N=1000							
θ_1	0.4853	0.5	0.5058	-0.0147	0.0015	0.0017	0.4487—0.5657
θ_2	0.9047	1.0	0.8975	-0.0953	0.0076	0.0166	0.7542—1.0208
θ_3	1.4979	1.5	1.4629	-0.0021	0.0097	0.0097	1.3987—1.6916
θ_4	2.0873	2.0	2.1241	0.0874	0.0103	0.0179	1.939—2.2482
θ_5	2.3485	2.5	2.4320	-0.1515	0.0108	0.0338	2.1973—2.5405
θ_6	2.8892	3.0	2.9465	-0.1108	0.0083	0.0206	2.7659—3.0162
N=2000							
θ_1	0.5155	0.5	0.5074	0.0155	0.0045	0.0047	0.3977—0.6163
θ_2	0.8487	1.0	0.8964	-0.1513	0.0102	0.0331	0.6647—1.0167
θ_3	1.4233	1.5	1.4584	-0.0767	0.0068	0.0127	1.2936—1.5739
θ_4	2.1219	2.0	2.1216	0.1219	0.0056	0.0205	1.9836—2.2309
θ_5	2.355	2.5	2.4305	-0.1450	0.0087	0.0297	2.1661—2.4906
θ_6	2.8469	3.0	2.9465	-0.1531	0.0191	0.0425	2.6208—3.0568
N=3000							
θ_1	0.5028	0.5	0.5070	0.0028	0.0055	0.0055	0.3732—0.6188
θ_2	0.8546	1.0	0.8971	-0.1454	0.0113	0.0325	0.6780—1.0287
θ_3	1.3874	1.5	1.4592	-0.1126	0.0099	0.0225	1.2278—1.5689
θ_4	2.0671	2.0	2.1228	0.0671	0.0072	0.0117	1.9282—2.2036
θ_5	2.3484	2.5	2.4321	-0.1516	0.0113	0.0342	2.1416—2.4808
θ_6	2.8438	3.0	2.9476	-0.1562	0.0160	0.0404	2.6251—3.0456
N=4000							
θ_1	0.4791	0.5	0.5070	-0.0209	0.0046	0.0051	0.3697—0.6150
θ_2	0.8417	1.0	0.8968	-0.1583	0.0077	0.0328	0.6909—1.0394
θ_3	1.3654	1.5	1.4595	-0.1345	0.0119	0.0300	1.1978—1.5574
θ_4	2.1020	2.0	2.1231	0.1020	0.0132	0.0236	1.9029—2.3057
θ_5	2.3448	2.5	2.4320	-0.1552	0.0191	0.0431	2.1084—2.6126
θ_6	2.7976	3.0	2.9473	-0.2024	0.0103	0.0513	2.6104—2.9908
N=5000							
θ_1	0.4683	0.5	0.5072	-0.0317	0.0039	0.0049	0.3577—0.5750
θ_2	0.8288	1.0	0.8965	-0.1712	0.0073	0.0366	0.6769—0.9823
θ_3	1.4084	1.5	1.4590	-0.0915	0.0109	0.0193	1.2680—1.6036
θ_4	2.0716	2.0	2.1221	0.0716	0.0081	0.0132	1.9085—2.2131
θ_5	2.3222	2.5	2.4319	-0.1778	0.01541	0.0470	2.0948—2.5289
θ_6	2.8138	3.0	2.9468	-0.1862	0.0135	0.0482	2.5742—2.9624

Table 5.2: Comparison between the standard adjusted posterior mean estimates ($\hat{\boldsymbol{\theta}}_{\text{adj}}$), true parameter values ($\hat{\boldsymbol{\theta}}_0$) and conjugate posterior mean estimates ($\hat{\boldsymbol{\theta}}_{\text{conjugate}}$) of the 6 parameters of the toy model across different values of N (from $N = 500$ to $N = 5000$).

Parameters	$\hat{\boldsymbol{\theta}}_{\text{adj}}$	$\hat{\boldsymbol{\theta}}_0$	$\hat{\boldsymbol{\theta}}_{\text{conjugate}}$	bias($\hat{\boldsymbol{\theta}}_{\text{adj}}$)	Var($\hat{\boldsymbol{\theta}}_{\text{adj}}$)	MSE($\hat{\boldsymbol{\theta}}_{\text{adj}}$)	95% Cred. Int.
				N=500			
θ_1	0.5502	0.5	0.5063	0.0502	0.0071	0.0097	0.4418—0.7090
θ_2	0.9568	1.0	0.8980	-0.0432	0.0336	0.0354	0.6571—1.1809
θ_3	1.4172	1.5	1.4664	-0.0828	0.0141	0.0210	1.1992—1.5399
θ_4	2.1245	2.0	2.1252	0.12447	0.0017	0.0172	2.0658—2.1730
θ_5	2.3436	2.5	2.4327	-0.1564	0.0079	0.0324	2.2000—2.4866
θ_6	2.9435	3.0	2.9467	-0.0565	0.0178	0.0210	2.7030—3.0890
				N=1000			
θ_1	0.6350	0.5	0.5058	0.1350	0.1554	0.1736	0.1522—1.3605
θ_2	1.066	1.0	0.8976	0.0660	0.0930	0.0973	0.5970—1.5427
θ_3	1.6321	1.5	1.4629	0.13205	0.0629	0.0803	1.1401—1.8758
θ_4	2.0795	2.0	2.1241	0.0795	0.0129	0.0193	1.9184—2.2594
θ_5	2.4111	2.5	2.4320	-0.0889	0.0667	0.0745	2.0394—2.6897
θ_6	3.1265	3.0	2.9465	0.1266	0.2522	0.2683	2.1989—3.7414
				N=2000			
θ_1	0.5200	0.5	0.5074	0.0200	0.0050	0.0054	0.3874—0.6491
θ_2	0.8283	1.0	0.8964	-0.1717	0.0076	0.0371	0.6665—0.9776
θ_3	1.4294	1.5	1.4584	-0.0706	0.0202	0.0252	1.2091—1.6672
θ_4	2.0865	2.0	2.1216	0.0866	0.0091	0.0166	1.9378—2.2277
θ_5	2.3017	2.5	2.4305	-0.1983	0.1059	0.1452	1.8252—2.8712
θ_6	2.8308	3.0	2.9464	-0.1692	0.0174	0.0460	2.5979—3.0677
				N=3000			
θ_1	0.5561	0.5	0.5070	0.0561	0.0099	0.0131	0.3366—0.6926
θ_2	1.0135	1.0	0.8971	0.0136	0.0302	0.0304	0.6207—1.2354
θ_3	1.3542	1.5	1.4592	-0.1458	0.4966	0.5181	0.7311—3.3913
θ_4	2.0331	2.0	2.1228	0.0331	0.1064	0.1075	1.6225—2.7373
θ_5	2.3965	2.5	2.4321	-0.1035	0.0295	0.0402	1.9518—2.6140
θ_6	3.0575	3.0	2.9476	0.0575	0.1713	0.1746	2.1120—3.6480
				N=4000			
θ_1	0.4122	0.5	0.5070	-0.0878	0.1041	0.1118	0.2259—0.9687
θ_2	0.9166	1.0	0.8968	-0.0834	0.0295	0.0365	0.5314—1.1814
θ_3	1.4103	1.5	1.4595	-0.0896	0.0125	0.02061	1.1861—1.6170
θ_4	2.0267	2.0	2.1231	0.0268	0.0497	0.0504	1.7632—2.4752
θ_5	2.5443	2.5	2.4320	0.0443	0.1350	0.1370	1.6821—3.1107
θ_6	2.8656	3.0	2.9473	-0.1344	0.0204	0.0385	2.5604—3.0843
				N=5000			
θ_1	0.4411	0.5	0.5072	-0.0589	0.0067	0.0103	0.3259—0.5829
θ_2	0.7651	1.0	0.8965	-0.2349	0.0113	0.0665	0.6276—1.0433
θ_3	1.4439	1.5	1.4590	-0.0561	0.0095	0.0126	1.2722—1.6720
θ_4	2.0309	2.0	2.1221	0.0310	0.0115	0.0124	1.8560—2.2761
θ_5	2.3640	2.5	2.4319	-0.1360	0.0124	0.0309	2.1593—2.5994
θ_6	2.7485	3.0	2.9468	-0.2515	0.0036	0.0669	2.6816—2.8948

Table 5.3: Comparison between the proposed adjusted posterior mean estimates ($\hat{\boldsymbol{\theta}}_{\text{adj}}$), true parameter values ($\hat{\boldsymbol{\theta}}_0$) and conjugate posterior mean estimates ($\hat{\boldsymbol{\theta}}_{\text{conjugate}}$) of the 6 parameters of the toy model across different values of N (from $N = 500$ to $N = 5000$).

Parameters	$\hat{\boldsymbol{\theta}}_{\text{adj}}$	$\hat{\boldsymbol{\theta}}_0$	$\hat{\boldsymbol{\theta}}_{\text{conjugate}}$	bias($\hat{\boldsymbol{\theta}}_{\text{adj}}$)	Var($\hat{\boldsymbol{\theta}}_{\text{adj}}$)	MSE($\hat{\boldsymbol{\theta}}_{\text{adj}}$)	95% Cred. Int.
N=500							
θ_1	0.5092	0.5	0.5063	0.0092	0.0082	0.0083	0.3690—0.6381
θ_2	0.9691	1.0	0.8980	-0.0309	0.0064	0.0074	0.8160—1.0586
θ_3	1.4422	1.5	1.4664	-0.0578	0.0099	0.0132	1.3403—1.6164
θ_4	2.1352	2.0	2.1252	0.1352	0.0031	0.0214	2.0358—2.1947
θ_5	2.2658	2.5	2.4327	-0.2342	0.0057	0.0605	2.1927—2.4078
θ_6	2.9442	3.0	2.9467	-0.0557	0.0174	0.0204	2.6854—3.1162
N=1000							
θ_1	0.5059	0.5	0.5058	0.0059	0.0011	0.0011	0.4460—0.5507
θ_2	0.8756	1.0	0.8976	-0.1244	0.0076	0.0230	0.7542—1.0208
θ_3	1.4595	1.5	1.4629	-0.0405	0.0097	0.0114	1.3988—1.6918
θ_4	2.0740	2.0	2.1241	0.0739	0.0087	0.0142	1.9511—2.2306
θ_5	2.3327	2.5	2.4320	-0.1673	0.0106	0.0386	2.2014—2.5397
θ_6	2.8354	3.0	2.9464	-0.1645	0.0083	0.0354	2.7659—3.0162
N=2000							
θ_1	0.5209	0.5	0.5074	0.0209	0.0036	0.0041	0.4083—0.5990
θ_2	0.8851	1.0	0.8964	-0.1149	0.0102	0.0234	0.6648—1.0167
θ_3	1.4342	1.5	1.4584	-0.0657	0.0059	0.0102	1.3191—1.5620
θ_4	2.1169	2.0	2.1216	0.1169	0.0046	0.0183	2.0053—2.2160
θ_5	2.3697	2.5	2.4304	-0.1303	0.0071	0.0241	2.1784—2.4740
θ_6	2.8744	3.0	2.9465	-0.1256	0.0176	0.0334	2.5819—3.0153
N=3000							
θ_1	0.50178	0.5	0.5070	0.0017	0.0052	0.0052	0.3784—0.6144
θ_2	0.8439	1.0	0.8971	-0.1561	0.0108	0.0351	0.6908—1.0490
θ_3	1.4079	1.5	1.4592	-0.0921	0.0082	0.0167	1.2420—1.5525
θ_4	2.0646	2.0	2.1228	0.0646	0.0043	0.0085	1.9617—2.1635
θ_5	2.3612	2.5	2.4321	-0.1388	0.0084	0.0277	2.1541—2.4752
θ_6	2.8406	3.0	2.9476	-0.1594	0.0132	0.0387	2.6553—3.0584
N=4000							
θ_1	0.4830	0.5	0.5070	-0.0170	0.0038	0.0041	0.3742—0.5953
θ_2	0.8494	1.0	0.8968	-0.1506	0.0077	0.0304	0.6909—1.0394
θ_3	1.3655	1.5	1.4595	-0.1344	0.0119	0.0299	1.1978—1.5574
θ_4	2.1047	2.0	2.1231	0.1047	0.0106	0.0216	1.9211—2.2789
θ_5	2.3459	2.5	2.4320	-0.1541	0.0129	0.0366	2.1378—2.5463
θ_6	2.7938	3.0	2.9473	-0.2062	0.0088	0.0513	2.6270—2.9766
N=5000							
θ_1	0.4757	0.5	0.5072	-0.0243	0.0030	0.0035	0.3735—0.5520
θ_2	0.8180	1.0	0.8965	-0.1810	0.0056	0.0384	0.6950—0.9689
θ_3	1.3951	1.5	1.4589	-0.1049	0.0098	0.02078	1.2522—1.6369
θ_4	2.0777	2.0	2.1221	0.0777	0.0052	0.0112	1.9547—2.2093
θ_5	2.3147	2.5	2.4319	-0.1853	0.0091	0.0435	2.1605—2.5080
θ_6	2.8034	3.0	2.9468	-0.1966	0.0094	0.0480	2.6093—2.9506

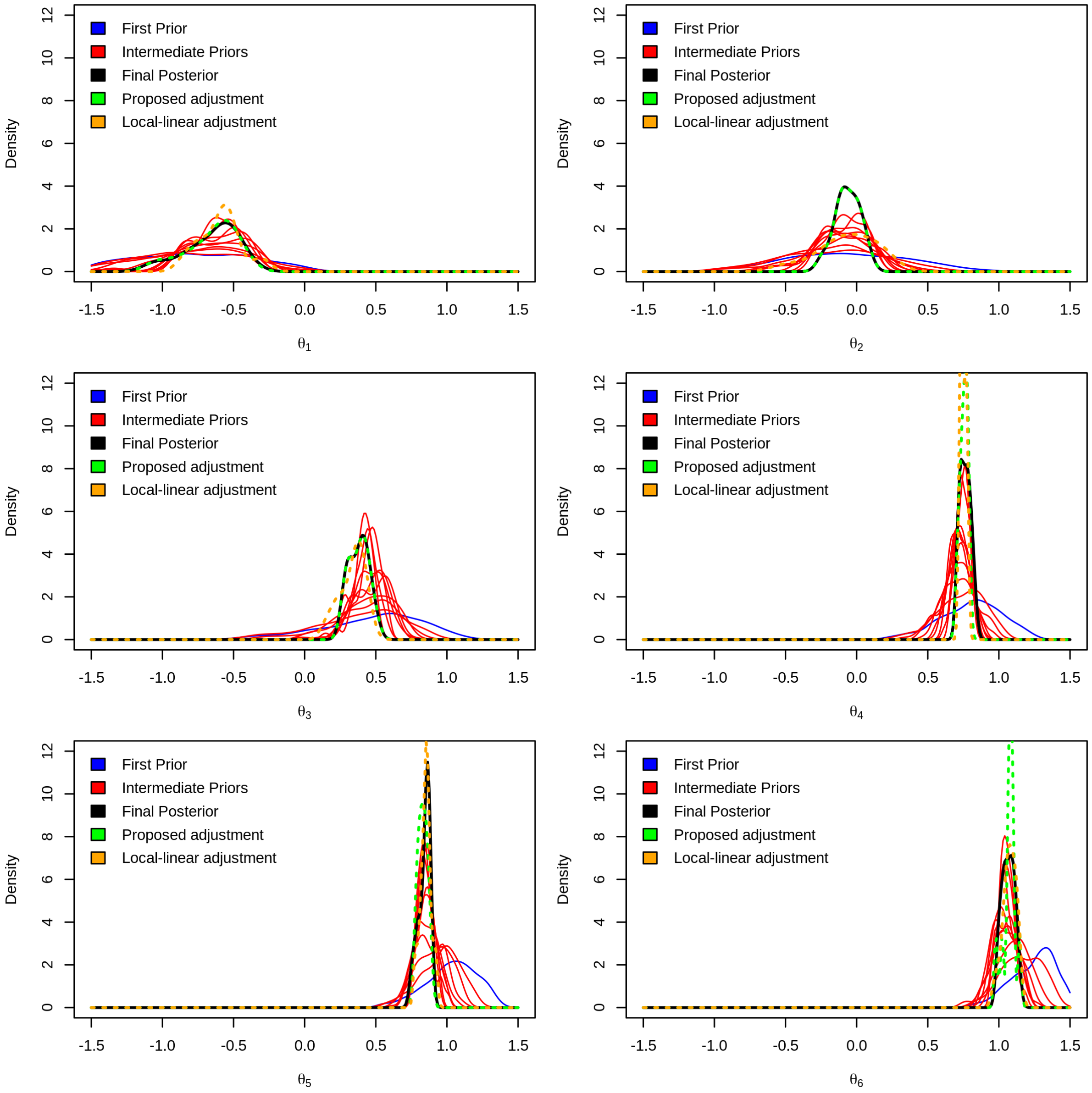

Figure 5.3: Goodness-of-fit density plots of the (unadjusted) approximate posterior distribution (in black) for the 6 parameters of the toy model against the sequentially improving prior distributions and the adjusted posterior from the two regression adjustments at $N = 500$ (on logarithmic scale).

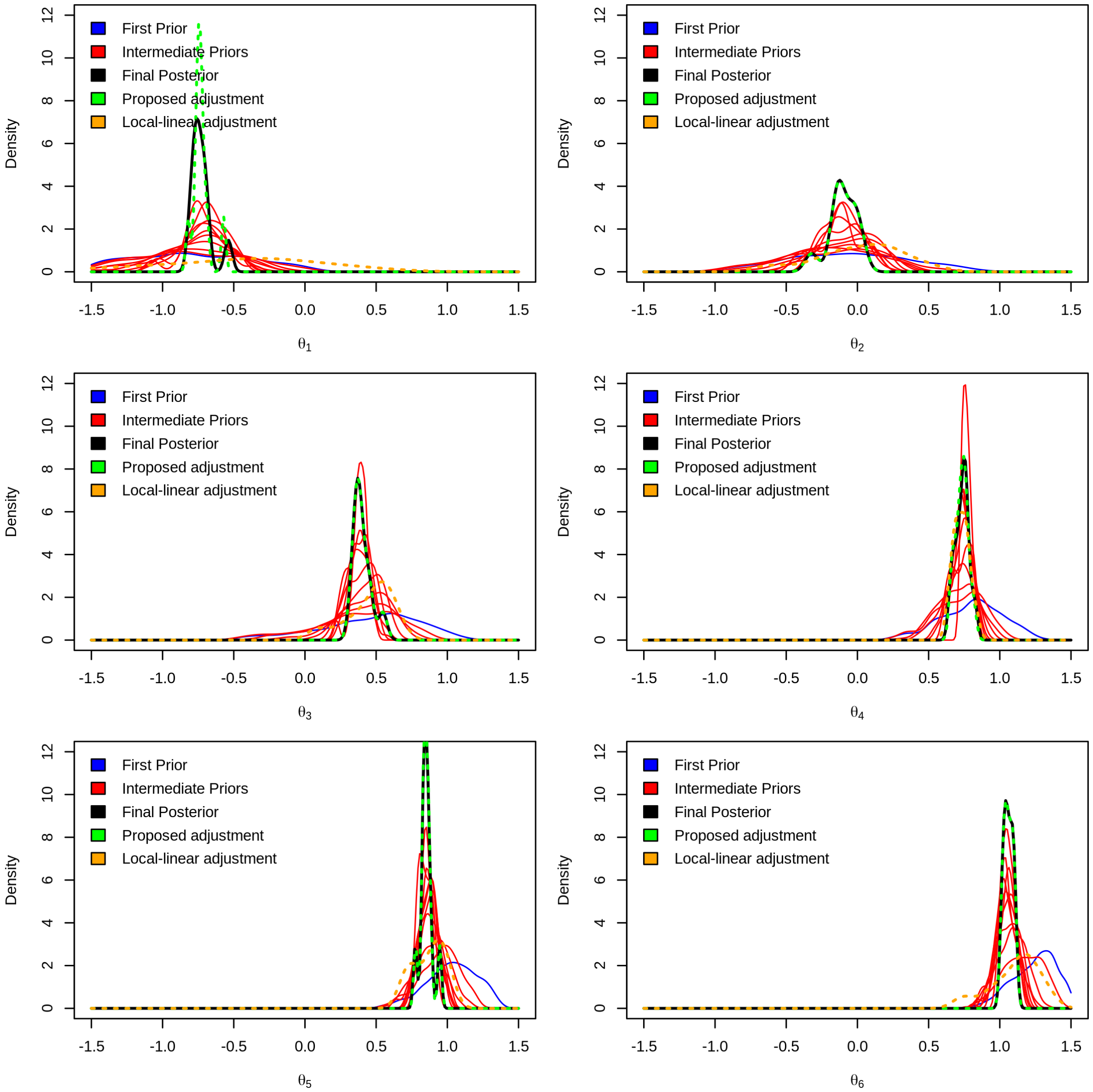

Figure 5.4: Goodness-of-fit density plots of the (unadjusted) approximate posterior distribution (in black) for the 6 parameters of the toy model against the sequentially improving prior distributions and the adjusted posterior from the two regression adjustments at $N = 1000$ (on logarithmic scale).

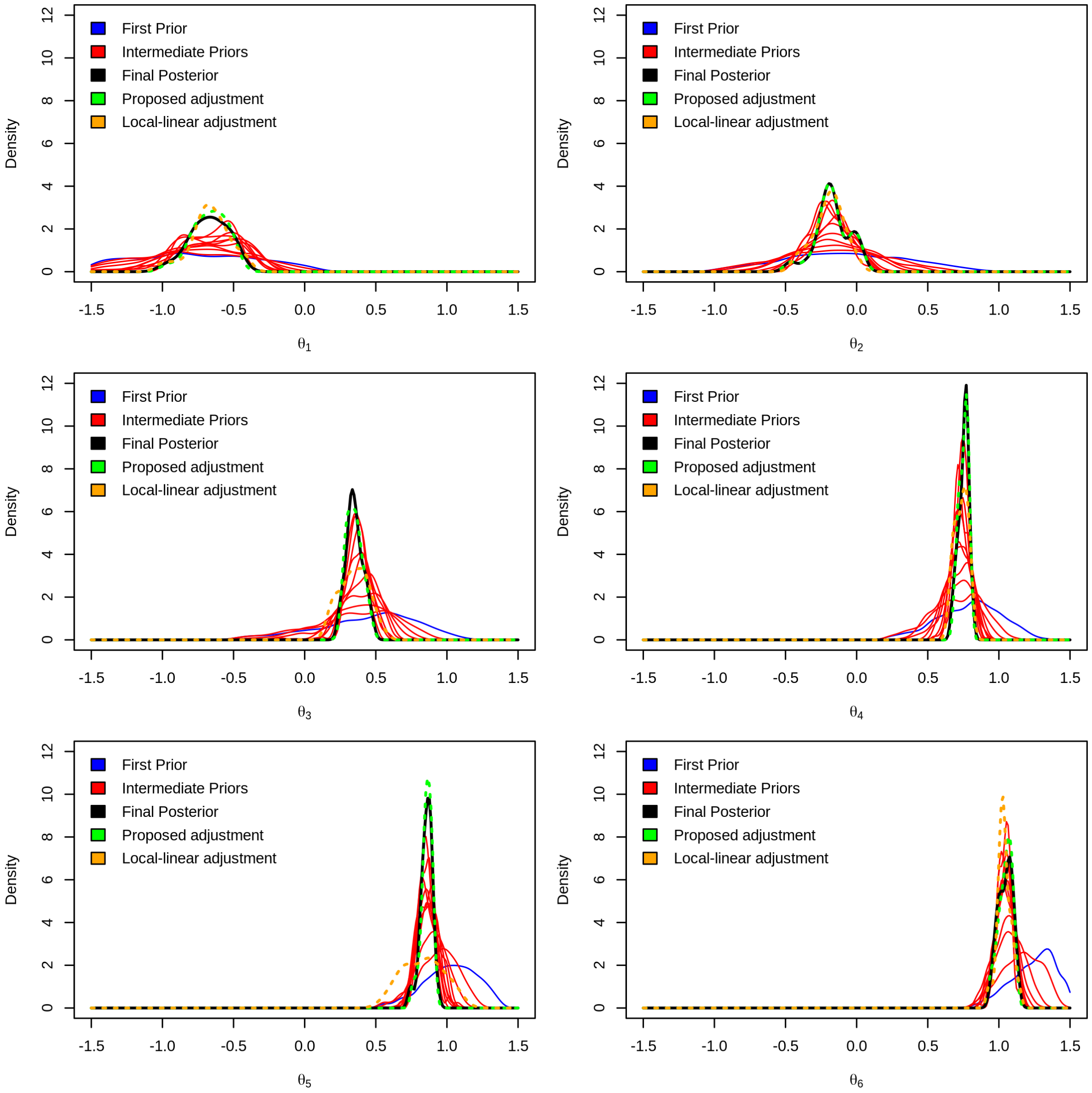

Figure 5.5: Goodness-of-fit density plots of the (unadjusted) approximate posterior distribution (in black) for the 6 parameters of the toy model against the sequentially improving prior distributions and the adjusted posterior from the two regression adjustments at $N = 2000$ (on logarithmic scale).

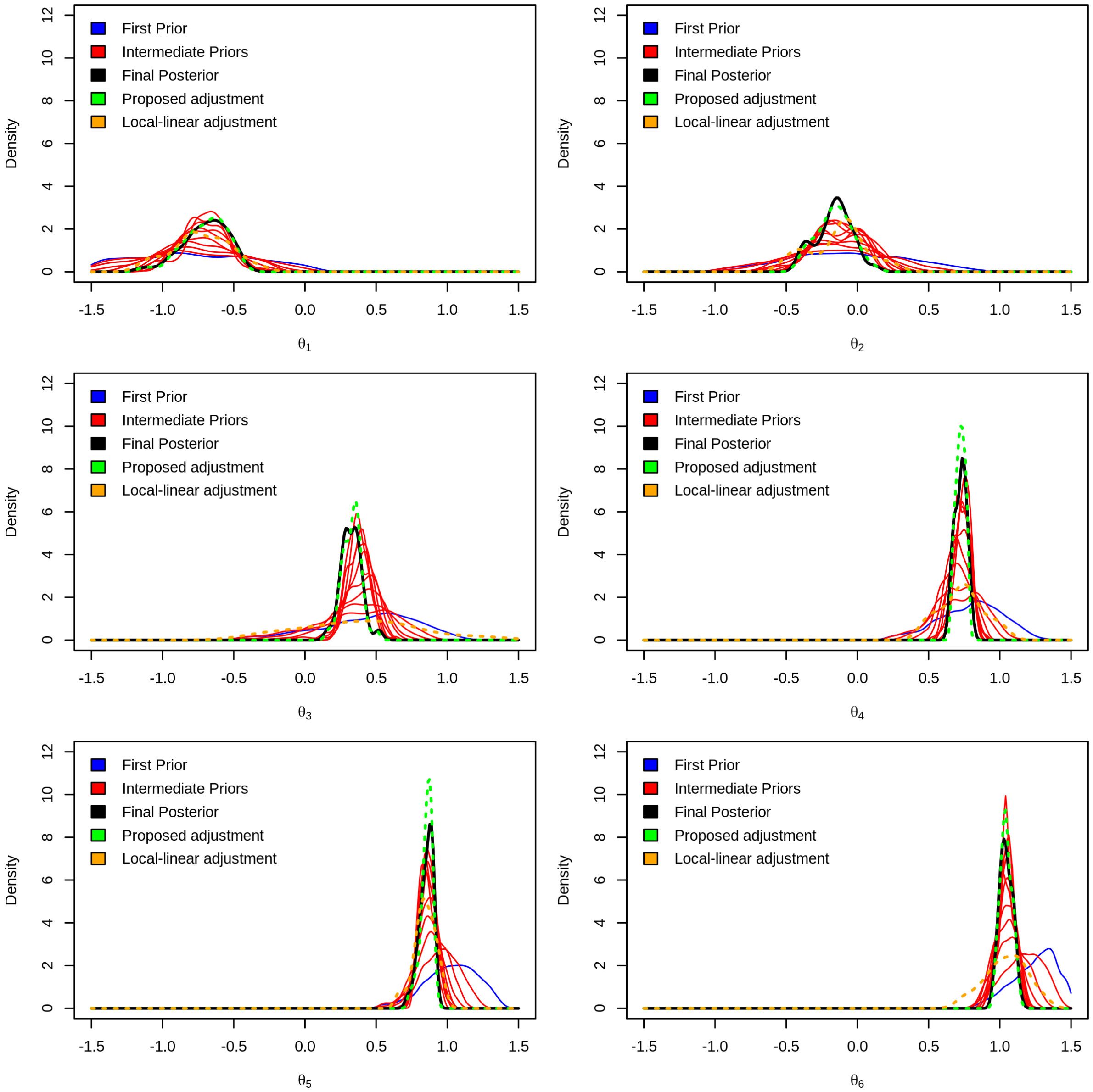

Figure 5.6: Goodness-of-fit density plots of the (unadjusted) approximate posterior distribution (in black) for the 6 parameters of the toy model against the sequentially improving prior distributions and the adjusted posterior from the two regression adjustments at $N = 3000$ (on logarithmic scale).

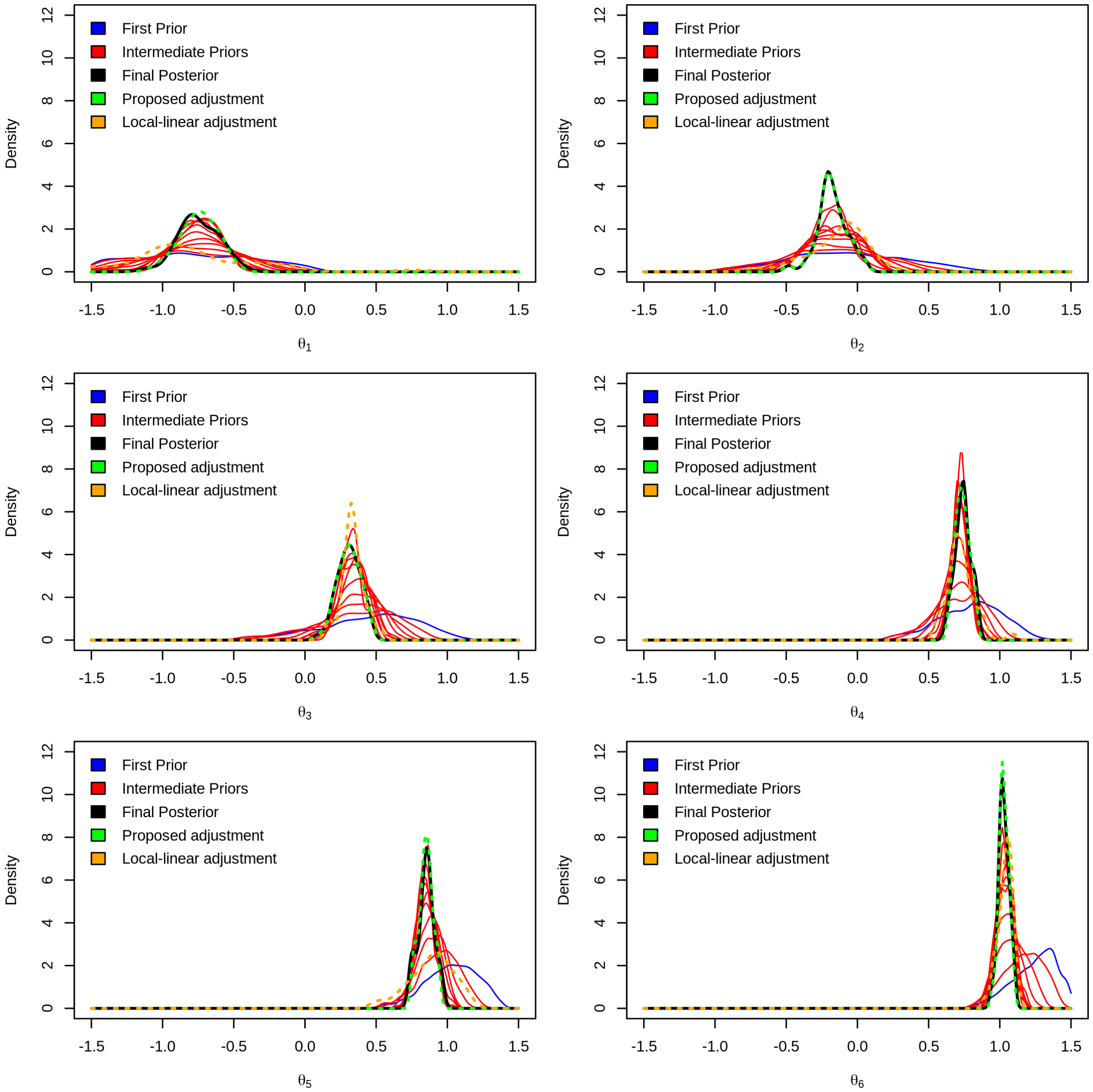

Figure 5.7: Goodness-of-fit density plots of the (unadjusted) approximate posterior distribution (in black) for the 6 parameters of the toy model against the sequentially improving prior distributions and the adjusted posterior from the two regression adjustments at $N = 4000$ (on logarithmic scale).

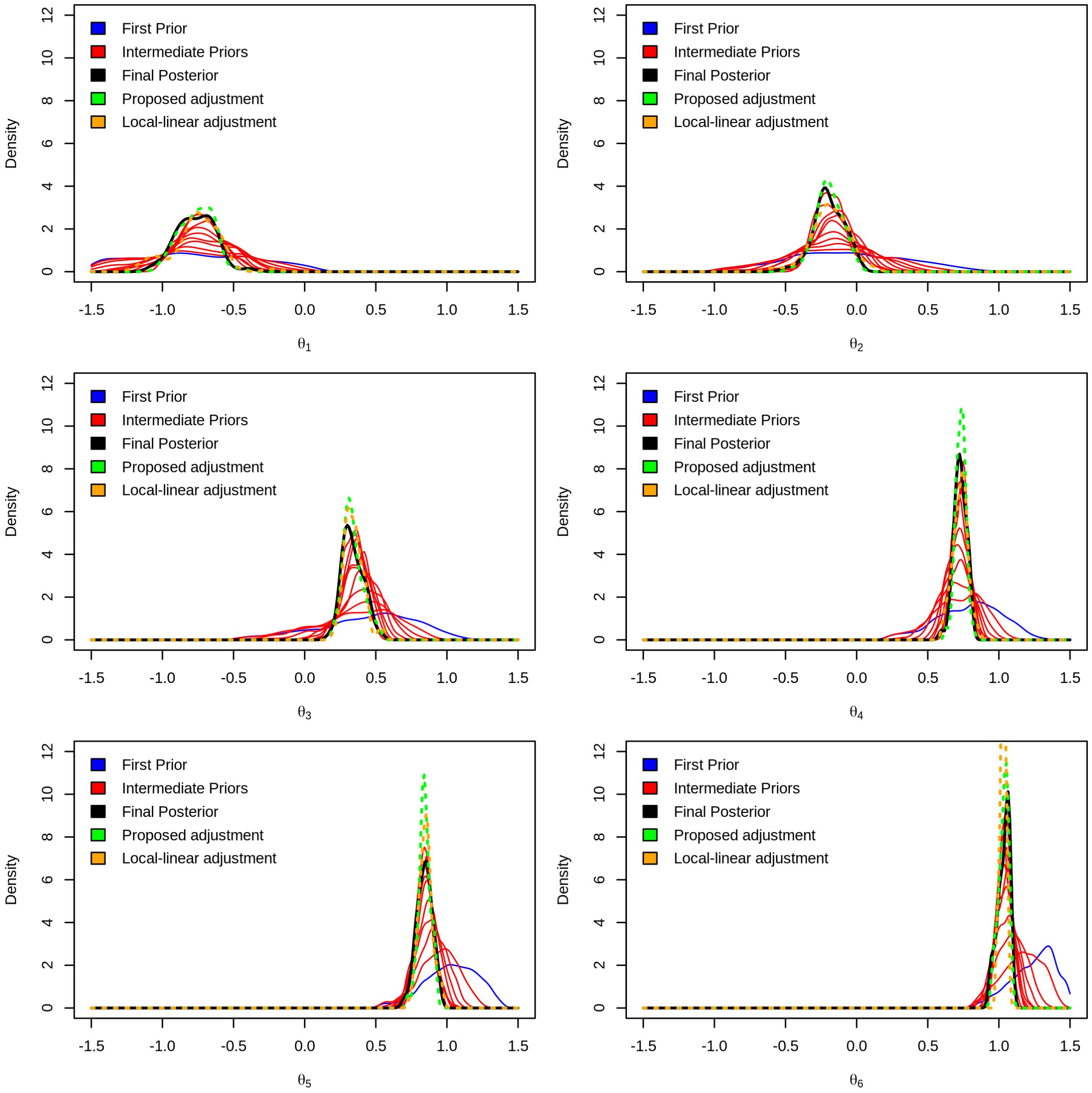

Figure 5.8: Goodness-of-fit density plots of the (unadjusted) approximate posterior distribution (in black) for the 6 parameters of the toy model against the sequentially improving prior distributions and the adjusted posterior from the two regression adjustments at $N = 5000$ (on logarithmic scale).

Chapter 6

Novel stochastic simulation model

6.1 Introduction

The multi-state Markov model (outlined in section 2.3.3) was not able to include spatial information and other relevant information about parasite fecundity, age group (young or old parasite), parasite mortality, parasite mobility and host immune response while exploring host survival and parasite infrapopulation dynamics. Thus, a more sophisticated stochastic model can incorporate such relevant data and help provide answers to other unknown biological questions and additional insights about the infrapopulation and mixed-population dynamics of the gyrodactylid-fish system.

Here, we compare the infection dynamics of the three *Gyrodactylus* parasite strains (*Gt3*, *Gt* and *Gb*) across the three host populations (OS, LA and UA fish), by developing a multidimensional continuous-time Markov Chain (CTMC) model via a hybrid τ-leaping simulation. The leap size selection (equation 4.26) based on the B-D-C process (the auxiliary stochastic model) also provided additional means of accelerating the multidimensional simulation model (see Chapter 4). The model simulates (conditioned on relevant information such as fish sex, fish size, fish type and parasite strain) the movement of parasites for two age groups over the external surfaces (four major body regions) of a fish over a 17-day infection period with population carrying capacity (dependant on host size and area of body regions). Based on findings from the spatial and temporal parasite dynamics of the *Gyrodactylus* species (see Chapter 2), the eight body regions of fish (tail fin, lower body, upper body, anal fin, dorsal fin, pelvic fins, pectoral fins and head) shown in Figure 6.1 were re-categorised into four major body locations: tail, lower region (comprising of the lower body, anal fin, pelvic fins and dorsal fin), upper region (made up of the upper body and pectoral fins) and the head (Figure 6.2) in the multidimensional stochastic model.

It was observed from the empirical data (described fully in section 2.2.1) that there were lower mean parasite intensities at the pectoral, pelvic, dorsal and anal fins compared to the tail, lower body, upper body, and head regions due to either fish being maintained in isolation or difference in the surface area of these body regions. Individual host isolation suggested that there was no opportunity for host-to-host transmission to occur via the fins and, thus, the need for the body re-categorisation into the four major regions as represented in the transition diagram or movement model for a single parasite (Figure 6.2). The model was parameterised by the birth, death and movement rates of young and older parasites in the presence or absence of the host's immune response. Host death was assumed to occur at a rate proportional to the total number of parasites on the fish. Parasite body preference which depends on the parasite strain (microhabitat preference), is included in the stochastic model. The preference for parasites to move back and forth on the host and the effective carrying capacity (total parasites that can occupy each body location) of a fish are additional model parameters that are estimated. The underlying specific assumptions of the complex stochastic model (see section 6.2.2) were motivated by the findings from the multi-state Markov model in exploring the infection progression of the *Gyrodactylus* species, in terms of host survival and parasite virulence of the parasite strains across the three host populations (see Chapter 2). The CTMC simulation model was fitted using a modified weighted-iterative ABC (developed in Chapter 5).

Continuous-time Markov chain is often used to model biological systems or processes where there are a low population count and a high degree of uncertainty associated with transitions across different states of the process [21]. For this study, the infection dynamics of the laboratory-bred *G. turnbulli* (*Gt3*), a wild *G. turnbulli* strain (*Gt*) and wild *G. bullatarudis* (*Gb*) across three different fish stocks (OS, LA and UA fish) are being compared. The gyrodactylid parasite population are usually low among the host population

[307]; however, the infection dynamics over time varies across different parasite strains and different fish populations (Chapter 2). Hence, developing a CTMC simulation model for the gyrodactylid-fish system can capture the stochasticity of this system in the best possible way and incorporate relevant complexities into the model; with the aid of existing empirical data. The multidimensional stochastic model is formally defined in section 6.2 with some numerical experiments performed in section 6.2.3; whereas section 6.3 focuses on the ABC fitting of the multidimensional stochastic model, including other hypotheses testing results.

6.2 Construction of the CTMC simulation model

This section presents the framework of the CTMC stochastic model for the gyrodactylid-fish system (section 6.2.1), the hybrid τ-leaping algorithm for the multidimensional CTMC simulation model with its underlying assumptions (section 6.2.2) as well as the pseudo-codes of the hybrid τ-leaping algorithms (section 6.2.2.1) for simulating the spatial and temporal infection dynamics for a fish (conditioned on relevant information such as fish sex, fish size, fish type and parasite strain). Nine parasite-fish groups (described in Chapter 2) namely: *Gt3*-OS, *Gt3*-LA, *Gt3*-UA, *Gt*-OS, *Gt*-LA, *Gt*-UA, *Gb*-OS, *Gb*-LA and *Gb*-UA, are being compared. A linear function (least square regression equation) is used to project the number parasites after fish mortality until the end of the observation period (to aid in ABC fitting) as presented under section 5.3.3. The effect of different error bound values on simulation accuracy and speed for the CTMC simulation model are further explored; and a reasonable error threshold is chosen based on the trade-off between accuracy-speed trade-off (section 6.2.3).

6.2.1 Model framework

Suppose individual gyrodactylid parasites on infected fish can transition between four discrete states or major body locations: tail (state 1), lower region (state 2), upper region (state 3) and head (state 4) as represented by the transition diagram (Figure 6.2). Let $\{A^{(i)}_{j,k}(t); t \geq 0\}$ be $j \times k$ matrix denoting the number of gyrodactylid parasites at body location j $(j = 1,2,3,4)$ of fish i for parasite age group k $(k = 1,2)$ at any time t; where $k = 1$ represent young parasites (daughter yet to reproduce) and $k = 2$ denote old parasite (mother). Let $\{X_j^{(i)}(t); t \geq 0\}$ be the total number of young and old gyrodactylid parasites at any time t at the jth body location of fish i from any parasite-fish group, such that $X_j^{(i)}(t) = \sum_{k=1}^{2} A_{j,k}^{(i)}(t)$. For each fish $i = 1,2,\cdots,n_l$, where n_l is the total number of fish in the lth parasite-fish group (for $l = 1,2,\cdots,9$), we have observations $X_j^{(i)} = \left\{X_{j0}^{(i)}, X_{j1}^{(i)}, \cdots, X_{j9}^{(i)}\right\}$ at times $t_0 = 0$, $t_1 = 1$, $t_2 = 3$, $\cdots$, $t_9 = 17$. For simplicity, let assume $X_j^{(i)}(t) = X_{ju}^{(i)} = \sum_{k=1}^{2} A_{j,k}^{(i)}(t)$ for $t \in [t_{u-1}, t_u)$ (where $u = 1,2,\cdots,9$ are observed time indexes). Let $z_i = \{z_{i1}, z_{i2}, z_{i3}, z_{i4}\}$ be the realized values of the covariates: fish sex, fish size, fish stock and parasite strain, respectively, for fish i. Let also assume that $B_j \to \{0,1\}$ is an (unobserved) indicator function representing the immune state of the jth body region of a host over time; such that a value of 0 indicates the absence of immune response, while a value of 1 implies the presence of immune response. We suppose that $\{A_{j,k}^{(i)}(t); t \geq 0\}$ is a multidimensional time-homogeneous Markov chain with state space $S = \{0,1,2,\cdots\}$ defined by the number of parasites per age group k (young or old parasite) at each jth body region at time t $\cup$ immune states $\cup$ mortality state of fish i, and satisfies the following scheme at any time t:

Event	**Transition**	**Rate**
Parasite birth at region j	$A_{j,k}^{(i)} \to A_{j,k}^{(i)} + 1$	$A_{j,k}^{(i)} \times \left[1 - \frac{A_{j,k}^{(i)}}{\xi(f_j, z_{i2}, \kappa)}\right] \times b_k(B_j, z_{i4})$
Parasite death at region j	$A_{j,k}^{(i)} \to A_{j,k}^{(i)} - 1$	$A_{j,k}^{(i)} \times \left[1 - \frac{A_{j,k}^{(i)}}{\xi(f_j, z_{i2}, \kappa)}\right] \times d_k(B_j, z_{i4})$
Forward movement from region j to $j+1$	$A_{j,k}^{(i)} \to A_{j,k}^{(i)} - 1$, $A_{j+1,k}^{(i)} \to A_{j+1,k}^{(i)} + 1$	$A_{j,k}^{(i)} \times m_k(B_j) \times \epsilon(z_{i4})$
Backward movement from region j to $j-1$	$A_{j-1,k}^{(i)} \to A_{j-1,k}^{(i)} + 1$, $A_{j,k}^{(i)} \to A_{j,k}^{(i)} - 1$	$A_{j,k}^{(i)} \times m_k(B_j) \times (1 - \epsilon(z_{i4}))$
Immune response at region j	$\sum_{k=1}^{2} A_{j,k}^{(i)} \to 0$	$\left[\sum_{k=1}^{2} A_{j,k}^{(i)}\right] \times r(z_{i1}, z_{i3})$
Fish mortality	$\sum_{j=1}^{4}\sum_{k=1}^{2} A_{j,k}^{(i)} \to 0$	$\left[\sum_{j=1}^{4}\sum_{k=1}^{2} A_{j,k}^{(i)}\right] \times s(z_{i1}, z_{i2})$

where $b_k(B_j, z_{i4})$ is the birth rate for parasites age k (which depends on the immune state

B_j at body region j and parasite strain z_{i4}), $d(B_j, z_{i4})$ is the death rate for parasites age k (which depends on B_j and z_{i4}), $m_k(B_j)$ is the movement rate for parasites age k (which depends on B_j), $\epsilon(z_{i4})$ is the movement rate adjustment (which depends on the parasite strain z_{i4}), $r(z_{i1}, z_{i3})$ is the immune response rate by a single parasite (which depends the fish sex z_{i1} and fish type z_{i3}), $s(z_{i1}, z_{i2})$ is the fish mortality rate caused by a single parasite (which depends on the fish sex z_{i1} and fish size z_{i2}), $\xi(f_j, z_{i2}, \kappa)$ is the population carrying capacity (which depends on the area of body region f_j, fish size z_{i2} and the effective carrying capacity per unit area of each body region κ). The main model parameters of underlying the stochastic simulation to be estimated are described in Table 6.1.

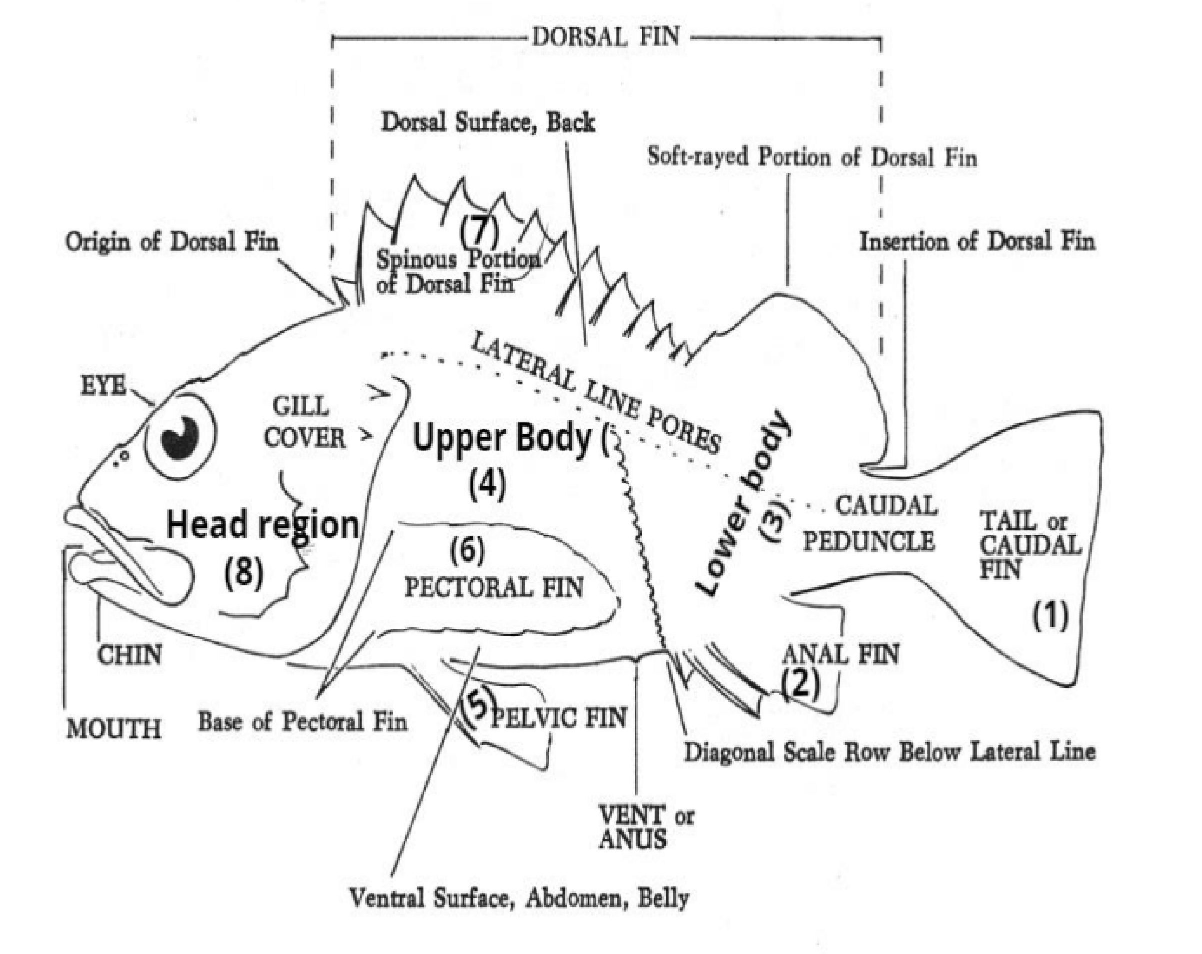

Figure 6.1: Conceptual framework showing the eight body locations of fish.

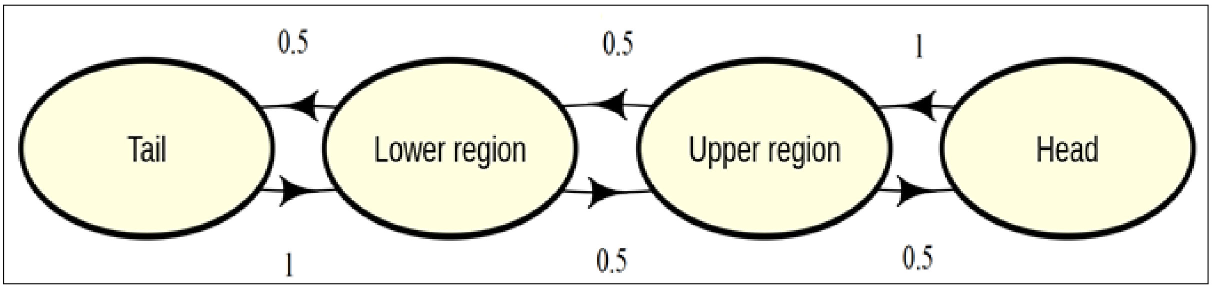

Figure 6.2: Transition diagram across the four major body regions of fish used as states for the CTMC model for a single parasite.

6.2.2 Hybrid τ-leaping algorithm for the multidimensional CTMC simulation model

The CTMC stochastic simulation model is developed using a hybrid τ-leaping algorithm whose leap size, τ_{leap}, (given by equation 6.1) is a modified version of the optimal leap size of the B-D-C process (the auxiliary model) given by equation 4.26; such that

$$\tau_{\text{leap}} = \min\left\{ \frac{\epsilon(\bar{b}+\bar{d})}{|(\bar{b}-\bar{d})|\max(\bar{b},\bar{d})}, \frac{\epsilon^2(\bar{b}+\bar{d})^2\left[\sum_{j=1}^{4}\sum_{k=1}^{2} A_{j,k}^{(i)}\right]}{(\bar{b}+\bar{d})\max(\bar{b}^2,\bar{d}^2)} \right\}, \tag{6.1}$$

where $\bar{b}$ is the average birth rate of young and old parasites, $\bar{d}$ is the average death rate of parasites in the presence or absence of host immune response, and ϵ is the error bound of the τ-leaping algorithm (which is pre-determined based on the trade-off between simulation accuracy and speed). The leap condition is determined by $\frac{1}{10a_0\left(A_{j,k}^{(i)}\right)}$ where $a_0\left(A_{j,k}^{(i)}\right)$ is the total event rate (which depends on state $A_{j,k}^{(i)}$) for fish i as specified in Algorithm 3 for simulating the B-D-C process. Thus, the hybrid τ-leaping is set up such that if the leap size τ_{leap} (equation 6.1) $> \frac{1}{10a_0\left(A_{j,k}^{(i)}\right)}$, the τ-leaping algorithm is implemented for a single fish (as given by Algorithm 6), whereas we forego τ-leaping and use the exact stochastic simulation algorithm (i.e., exact SSA given by Algorithm 5) when the leap condition is not met. The hybrid τ-leaping simulation at an error bound of 0 ($\epsilon = 0$) result in exact SSA only since at $\epsilon = 0$, the leap size $\tau_{\text{leap}} = 0$ for any state value and birth-death parameter values > 0. Thus, the leap condition is not satisfied for

τ-leaping at $\epsilon = 0$. The probability of that a single parasite will move between the four major body regions of fish within the simulation model (J) is assumed to be constant over time (as shown in Figure 6.2), and it is given as

$$J = \begin{array}{c} \\ \text{Tail} \\ \text{Lower region} \\ \text{Upper region} \\ \text{Head} \end{array} \begin{array}{c} \begin{array}{cccc} \text{Tail} & \text{Lower region} & \text{Upper region} & \text{Head} \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 \end{pmatrix} \end{array}.$$

The specific underlying assumptions of the CTMC simulation model are as follows:

1. The birth rate of young parasites are greater than the old parasites' birth rate.

2. The death rate of young and old parasites are assumed to be equal but higher in the presence of host immune response.

3. The birth rate per age depends on the parasite strain; whereas, the death rate with or without host immune response depends on the parasite strain.

4. Host mortality occur at a rate proportional to the total number of parasites on the body of the fish, fish sex and fish size.

5. The rate of movement of each parasite depends its age, strain and immune response.

6. Localised host immune response at each body region occurs at a rate proportional to the effective population carrying capacity per unit area, fish sex and fish stock. The localised immune response can also occur at any time within the observed infection period.

7. The fish size is measured by its standard length, and the unit area of the host's body regions depend its size and sex.

8. The population carrying capacity depends on the unit area of the host's body regions, fish size and the effective carrying capacity (maximum number of parasites per unit area of body regions).

9. The transition or event rates are time-homogeneous and dependent on the current state of the process (independent of past states) within any infinitesimal amount of time or time step of the τ-leaping simulation.

Table 6.1: Main model parameters of the multidimensional CTMC stochastic simulation

Parameters	Description
Base simulation parameters	
b_{11}	birth rate for young *Gt3* parasites
b_{12}	birth rate for old *Gt3* parasites
b_{21}	birth rate for young *Gt* parasites
b_{22}	birth rate for old *Gt* parasites
b_{31}	birth rate for young *Gb* parasites
b_{32}	birth rate for old *Gb* parasites
d_{11}	death rate for *Gt3* parasites without host immune response
d_{12}	death rate for *Gt3* parasites with host immune response
d_{21}	death rate for *Gt* parasites without host immune response
d_{22}	death rate for *Gt* parasites with host immune response
d_{31}	death rate for *Gb* parasites without host immune response
d_{32}	death rate for *Gb* parasites with host immune response
m	movement rate for a single parasite
r	immune response rate caused by a single parasite
s	host mortality rate caused by a single parasite
κ	effective carrying capacity per each body region
Additional simulation parameters	
ϵ_1	movement rate adjustment for *Gt3* parasites
ϵ_2	movement rate adjustment for *Gt* parasites
ϵ_3	movement rate adjustment for *Gb* parasites
r_1	immune response rate adjustment for LA fish (ref: UA fish)
r_2	immune response rate adjustment for OS fish (ref: UA fish)
r_3	immune response rate adjustment for male fish (ref: female)
s_1	host mortality rate adjustment for male fish (ref: female)

Remark. Other additional notations in the simulation model are f_j for $j = 1,2,3,4$ representing the unit area of the four major body regions, B_j are immune states at each body region (no response: $B_j = 1$; response: $B_j = 2$), and $\mathbf{x}$ representing the survival status of fish. In the multidimensional simulation model, the fish is initially infected with at least 2 parasites at the tail (in a similar fashion as the observed empirical data described in section 2.2.1), and the total number of parasites at each body region is recorded over time. For each simulation realisation, the stochastic model is set-up to simulate the entire fish population corresponding to each parasite-fish groups as observed in the empirical experimental data (conditioned on descriptive information such as fish sex and fish size).

6.2.2.1 Pseudo-codes of exact simulation and τ-leaping for the multidimensional model

The novel pseudo-codes of exact stochastic simulation (SSA) and the hybrid τ-leaping algorithm for the multidimensional CTMC model for simulating infection dynamics of a single fish are given by Algorithms 5 and 6, respectively. For the R codes used in simulating the infection dynamics of a single fish, and a group of fish corresponding to the observed empirical data, see Appendix G.

```
Algorithm 5: Exact SSA of the simulation model (pseudo-code)
Input: $A_0$, $B_0$, $J$, $\mathbf{x}$, $b_k$, $d_k$, $m_k$, $r$, $s$, $s_1$, $\kappa$, $\epsilon_1$, $\epsilon_2$, $\epsilon_3$, $f$, $t$, $t_{\text{final}}$, fish sex ($z_{i1}$),
       fish size ($z_{i2}$), fish stock ($z_{i3}$), parasite strain ($z_{i4}$) and host survival
       status ($\mathbf{x}$).
Output: Number of parasites at each body region over time, $X_j(t) = \sum_{k=1}^{2} A_{j,k}^{(i)}(t)$
        for fish $i$; host survival status (alive: $\mathbf{x} = 1$; dead: $\mathbf{x} = 2$)
1 while $t < t_{final}$ and $\mathbf{x} = 1$ do
2     Set initial time $t = t_0$; state $A_{1,1} = A_0 = 2$ & zero elsewhere; immune state
        $B_j = B_0 = 1$; host survival status $\mathbf{x} = 1$.
3     Calculate the event rates $a_\delta\left(A_{j,k}^{(i)}\right)$ for events $\delta = 1, 2, \cdots, 6$ such that:
        Birth $= A_{j,k}^{(i)} \times \left[1 - (A_{j,k}^{(i)}/(f_j \times z_{i2} \times \kappa))\right] \times b_k(B_j, z_{i4})$,
        Death $= A_{j,k}^{(i)} \times \left[1 - (A_{j,k}^{(i)}/(f_j \times z_{i2} \times \kappa))\right] \times d_k(B_j, z_{i4})$,
        Forward movement $= A_{j,k}^{(i)} \times m_k(B_j) \times \epsilon(z_{i4})$,
        Backward movement $= A_{j,k}^{(i)} \times m_k(B_j) \times (1 - \epsilon(z_{i4}))$,
        Immune response $= \sum_{k=1}^{2} A_{j,k}^{(i)} \times r(z_{i1}, z_{i3})$,
        Fish mortality $= \sum_j \sum_k A_{j,k}^{(i)} \times s(z_{i1}, z_{i2})$.
4     Compute the total rate, $a_0 = \sum_{\delta=1}^{6} a_\delta$, for events $\delta = 1, 2, \cdots, 6$ (from step 3).
5     Determine the event to occur at the host's body regions using a random
        number $\mathbf{u}$ from $Uniform(0, a_0)$ at a probability equal to $\frac{a_\delta}{a_0}$, and update
        state $A_{j,k}^{(i)}$ for fish $i$ according to the scheme defined in section 6.2.1.
6     Generate time increment $\tau_{\text{SSA}}$ from $Exponential(a_0)$, and update the time
        such that $t = t + \tau_{\text{SSA}}$.
7     Record $\left(X_j = \sum_{k=1}^{2} A_{j,k}^{(i)}, \mathbf{x}\right)$ at the desired discrete times for fish $i$ and
        $j = 1, 2, 3, 4$.
8 end
```

Algorithm 6: Hybrid τ-leaping algorithm for simulation model (pseudo-code)

```
   Input: $A_0$, $B_0$, $J$, $\mathbf{x}$, $b_k$, $d_k$, $m_k$, $r$, $s$, $s_1$, $\kappa$, $\epsilon_1$, $\epsilon_2$, $\epsilon_3$, $f$, $t$, $t_{\text{final}}$, fish sex ($z_{i1}$),
          fish size ($z_{i2}$), fish stock ($z_{i3}$), parasite strain ($z_{i4}$) and host survival
          status ($\mathbf{x}$).
   Output: Number of parasites at each body region over time, $X_j(t) = \sum_{k=1}^{2} A_{j,k}^{(i)}(t)$
           for fish $i$; host survival status (alive: $\mathbf{x} = 1$; dead: $\mathbf{x} = 2$)
1  while $t < t_{final}$ and $\mathbf{x} = 1$ do
2      Set initial time $t = t_0$; state $A_{1,1} = A_0 = 2$ & zero elsewhere; immune state
         $B_j = B_0 = 1$; host survival status $\mathbf{x} = 1$.
3      Calculate the event rates $a_\delta\left(A_{j,k}^{(i)}\right)$ for events $\delta = 1, 2, \cdots, 6$ such that:
         Birth $= A_{j,k}^{(i)} \times \left[1 - (A_{j,k}^{(i)}/(f_j \times z_{i2} \times \kappa))\right] \times b_k(B_j, z_{i4})$,
         Death $= A_{j,k}^{(i)} \times \left[1 - (A_{j,k}^{(i)}/(f_j \times z_{i2} \times \kappa))\right] \times d_k(B_j, z_{i4})$,
         Forward movement $= A_{j,k}^{(i)} \times m_k(B_j) \times \epsilon(z_{i4})$,
         Backward movement $= A_{j,k}^{(i)} \times m_k(B_j) \times (1 - \epsilon(z_{i4}))$,
         Immune response $= \sum_{k=1}^{2} A_{j,k}^{(i)} \times r(z_{i1}, z_{i3})$,
         Fish mortality $= \sum_j \sum_k A_{j,k}^{(i)} \times s(z_{i1}, z_{i2})$.
4      Compute the total rate, $a_0 = \sum_{\delta=1}^{6} a_\delta$, for events $\delta = 1, 2, \cdots, 6$ (from step 3).
5      Compute the leap size $\tau_{\text{leap}}$ given by equation 6.1.
6      if $\tau_{leap} > \frac{1}{10a_0}$ then
7          set $t = t + \tau_{\text{leap}}$ and choose a random number $\mathbf{u}$ from $Uniform(0, a_0)$
8          if $\mathbf{u} < a_6$ then
9              Set $\mathbf{x} = 2$ and break (host mortality event occurs)
10         end
11
12         if $a_6 < \mathbf{u} < a_5 + a_6$ then
13             Set $B_j = 2$ (immune response event occurs)
14         end
15
16         if $a_5 + a_6 < \mathbf{u} < a_1 + a_2 + a_3 + a_5 + a_6$ then
17             update $A_{j,k} = A_{j,k} + \sum_\delta v_\delta P(a_\delta, \tau_{\text{leap}})$ where $P \sim Poisson(a_\delta \tau_{\text{leap}})$ and
                 $v_\delta$ is state-change vector (birth, death and forward movement events
                 occur)
18         else
19             update $A_{j,k} = A_{j,k} + \sum_\delta v_\delta P(a_\delta, \tau_{\text{leap}})$ (birth, death and backward
                 movement events occur)
20         end
21     else
22         Execute exact SSA (Algorithm 5)
23     end
24
25     Record $\left(X_j = \sum_{k=1}^{2} A_{j,k}^{(i)}, \mathbf{x}\right)$ at the desired discrete times for fish $i$ and
         $j = 1, 2, 3, 4$.
26 end
```

6.2.3 Determining an error bound for the Hybrid τ-leaping simulation model

A reasonable choice of the error bound ϵ ($0 < \epsilon \ll 1$) for the hybrid τ-leaping simulation model was further investigated by exploring the trade-off between simulation accuracy and computational speed at some fixed parameter values (Table 6.2) based on 100 different simulation realisations or repetitions; where each simulation realisation corresponded to the nine observed parasite-fish groups (given fish sex, fish size, fish stock and parasite strain). The simulation accuracy was quantified by the mean square error (given by equation 6.2) based on the mean (simulated) parasite numbers over time (day 1 to 17) from the exact SSA (Algorithm 5) and the hybrid τ-leaping algorithm (Algorithm 6) at 10 different error bound values ($\epsilon = 0.002, 0.004, 0.006, 0.008, 0.01, 0.02, 0.04, 0.06, 0.08$ and 0.1); such that

$$\text{MSE}\left(\bar{X}^{(g)}_{\text{leap}}(t), \bar{X}^{(g)}_{\text{SSA}}(t)\right) = \frac{1}{100}\sum_{r=1}^{100}\left(\bar{X}^{(g)}_{\text{leap},r}(t) - \bar{X}^{(g)}_{\text{SSA},r}(t)\right)^2, \tag{6.2}$$

where $\bar{X}^{(g)}_{\text{leap},r}(t)$ and $\bar{X}^{(g)}_{\text{SSA},r}(t)$ are the mean parasite numbers over time t from hybrid $\tau - leaping$ and exact SSA, respectively; across each of the nine parasite-fish groups (g) and simulation realisation (r). The respective confidence intervals of the mean over time between the two simulation methods were also compared at $0 < \epsilon < 0.1$ for each parasite-fish group over time. The simulation speed was quantified by the computational time (computer's CPU time measured in seconds).

It was discovered that the simulation accuracy reduces in the hybrid τ-leaping algorithm as the error bound (ϵ) increases from $\epsilon = 0.002$ to $\epsilon = 0.1$ (see Figure 6.3). From Figures 6.4-6.13, it can be observed that the mean parasite numbers from the hybrid τ-leaping simulations were relatively consistent with the exact SSA at error bounds, $0.002 \leq \epsilon \leq 0.01$, across the nine parasite-fish groups (see Figures 6.4–6.8); including their respective confidence intervals. At $\epsilon \geq 0.02$, the τ-leaping algorithm started to perform badly across the parasite-fish groups as the error bound increased towards $\epsilon = 0.1$ (see Figures 6.9–

6.13). Figure 6.14 shows that at $\epsilon = 0.008$ or $\epsilon = 0.01$, the τ-leaping algorithm was relatively fast but not very significant from the computational time of the exact SSA. This may be due to either the smaller number of simulation repetitions (100 repetitions) or the number of parasites from the simulations being relatively small over time at the pre-specified parameter values (which were randomly chosen). Thus, the leaping condition was not met most of the time for just these simple explorations. However, it has already been shown in Chapter 4 that once the leap condition is met, τ-leaping is much faster compared to the exact SSA (otherwise, the latter is used the proposed hybrid $\tau-$leaping algorithm given by Algorithm 6). Based on the simulation accuracy and computational speed, $\epsilon = 0.01$ can be a reasonable choice of the error bound for subsequent simulations from the multidimensional stochastic model.

Table 6.2: Fixed parameter values for choosing an error bound

Parameters	Fixed values
Base simulation parameters	
b_{11}	0.668
b_{12}	0.018
b_{21}	0.668
b_{22}	0.018
b_{31}	0.668
b_{32}	0.018
d_{11}	0.008
d_{12}	0.071
d_{21}	0.008
d_{22}	0.071
d_{31}	0.008
d_{32}	0.071
m	0.083
r	0.001
s	0.009
κ	182
Additional simulation parameters	
ϵ_1	0.545
ϵ_2	0.333
ϵ_3	0.001
r_1	0.351
r_2	0.196
r_3	0.994
s_1	0.041

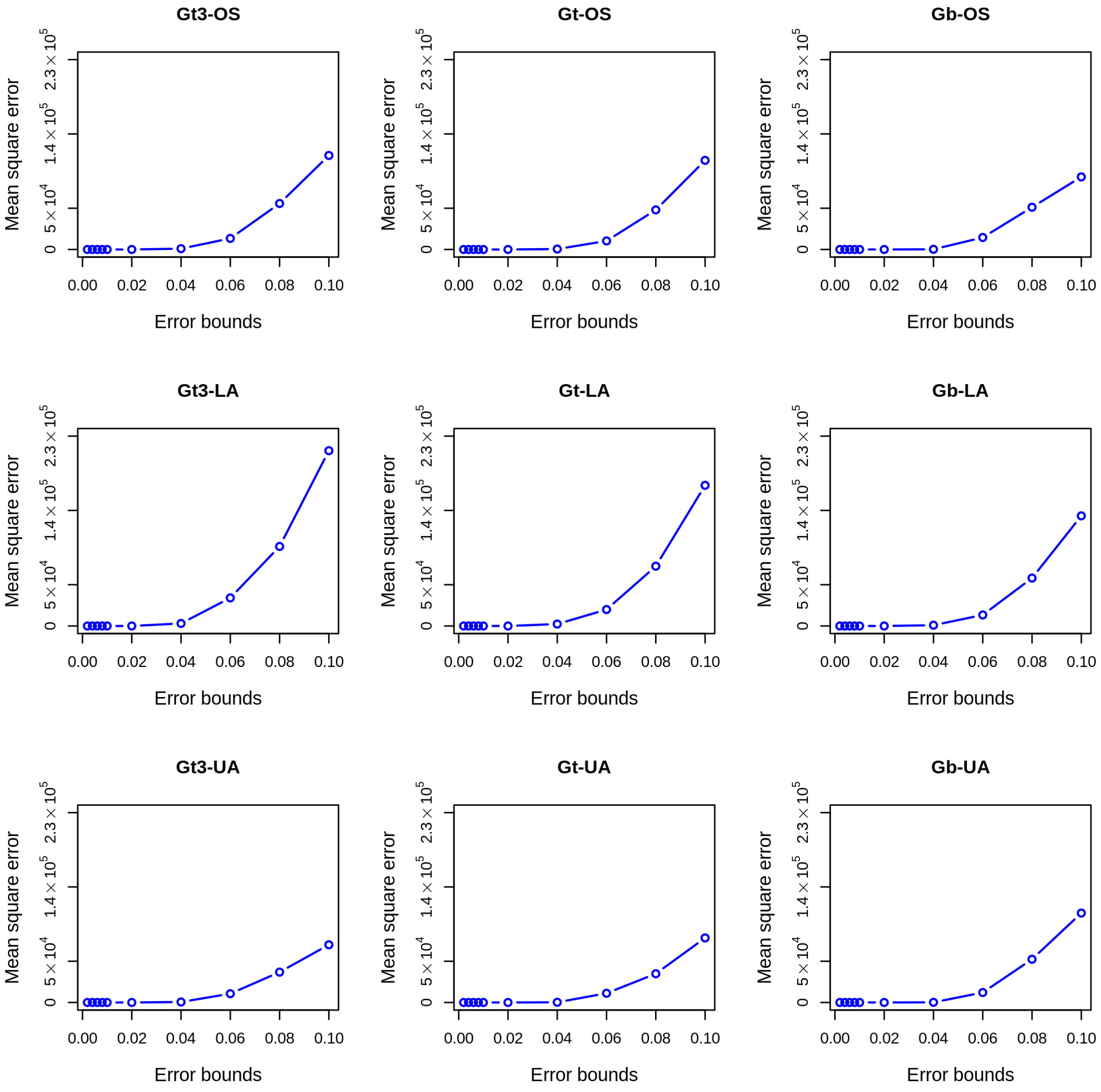

Figure 6.3: Mean square error from the Hybrid τ-leaping algorithm at different error bounds.

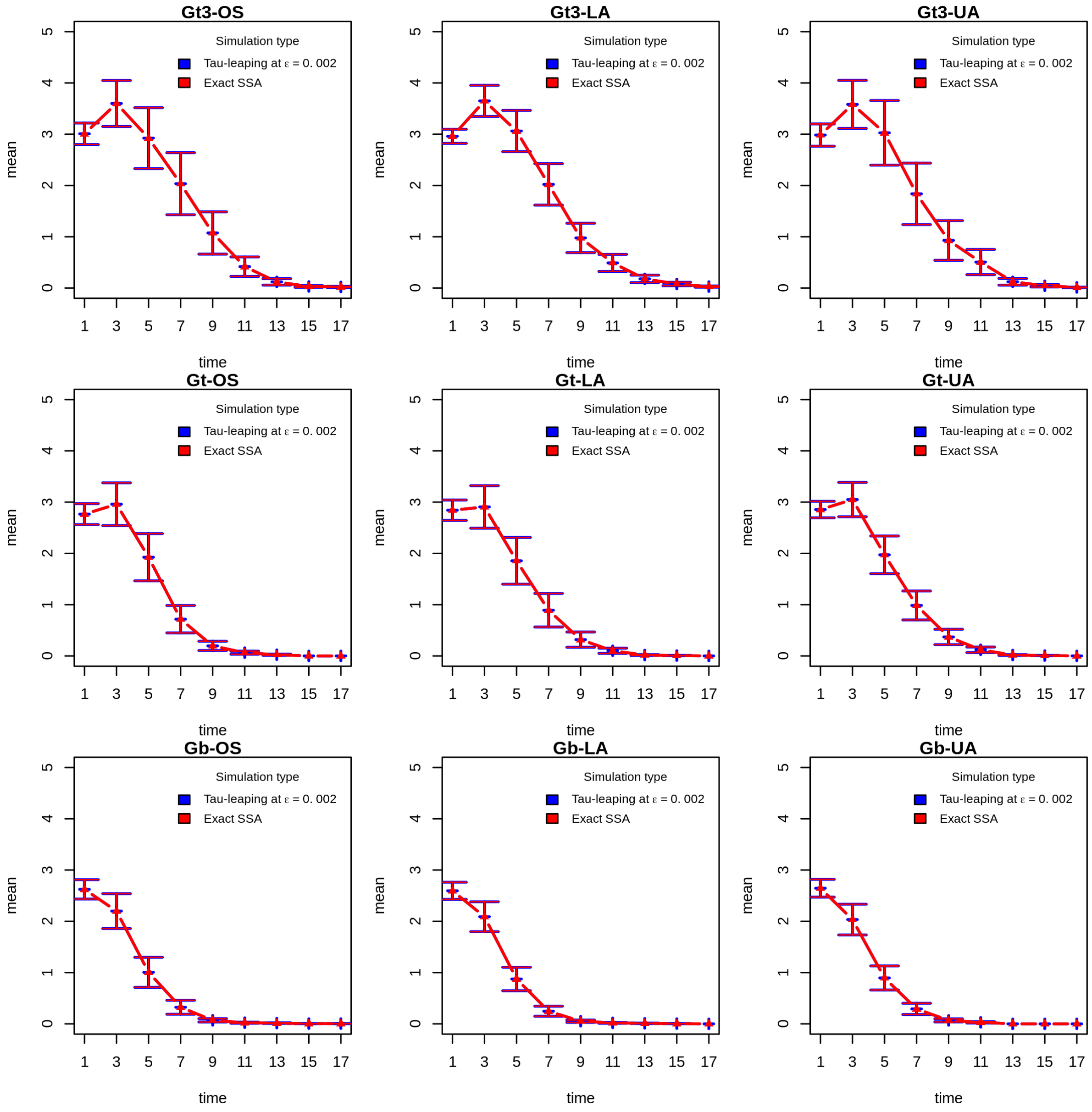

Figure 6.4: Mean comparison between exact SSA and Hybrid τ-leaping simulations at $\epsilon = 0.002$.

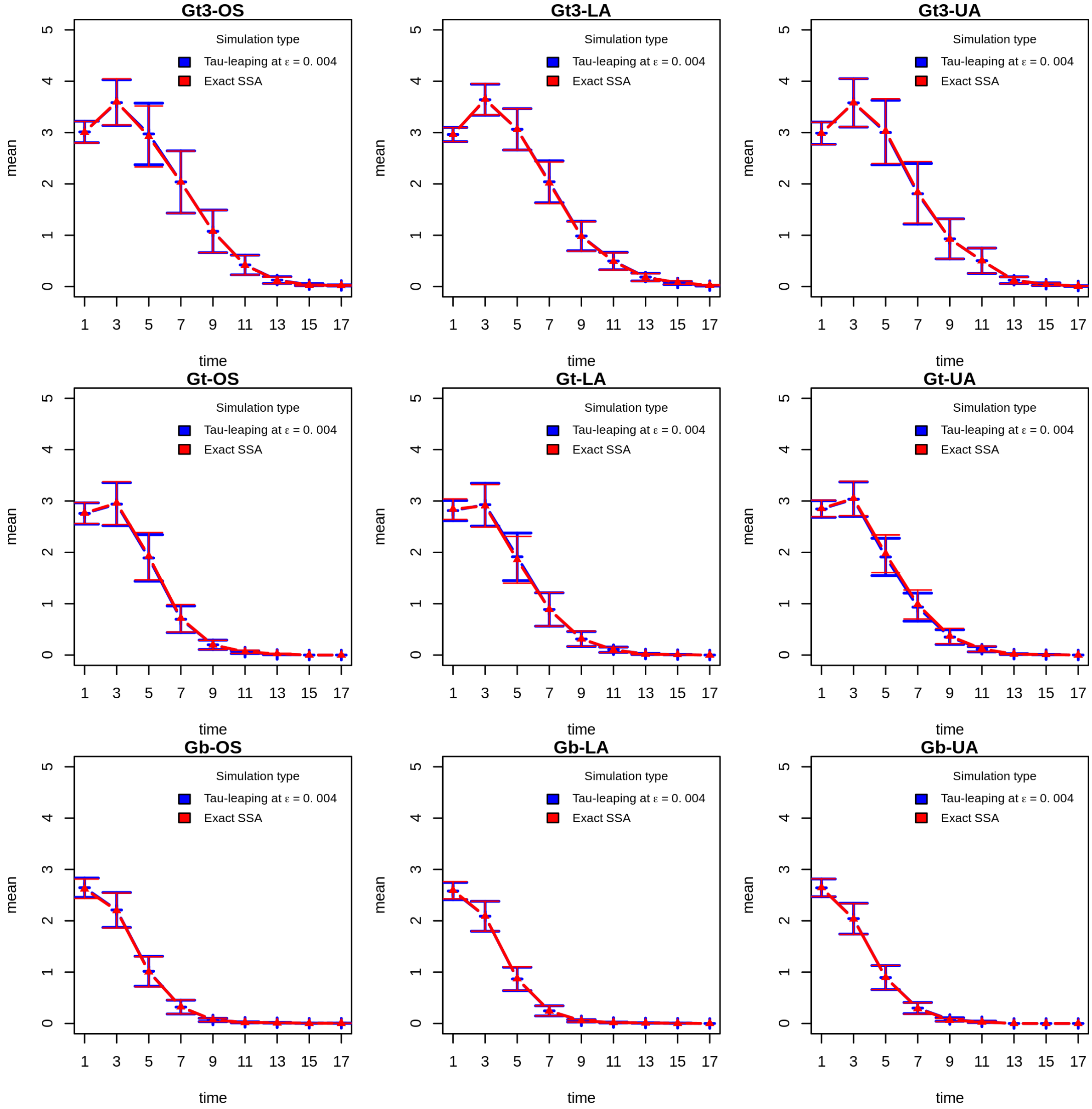

Figure 6.5: Mean comparison between exact SSA and Hybrid τ-leaping simulations at $\epsilon = 0.004$.

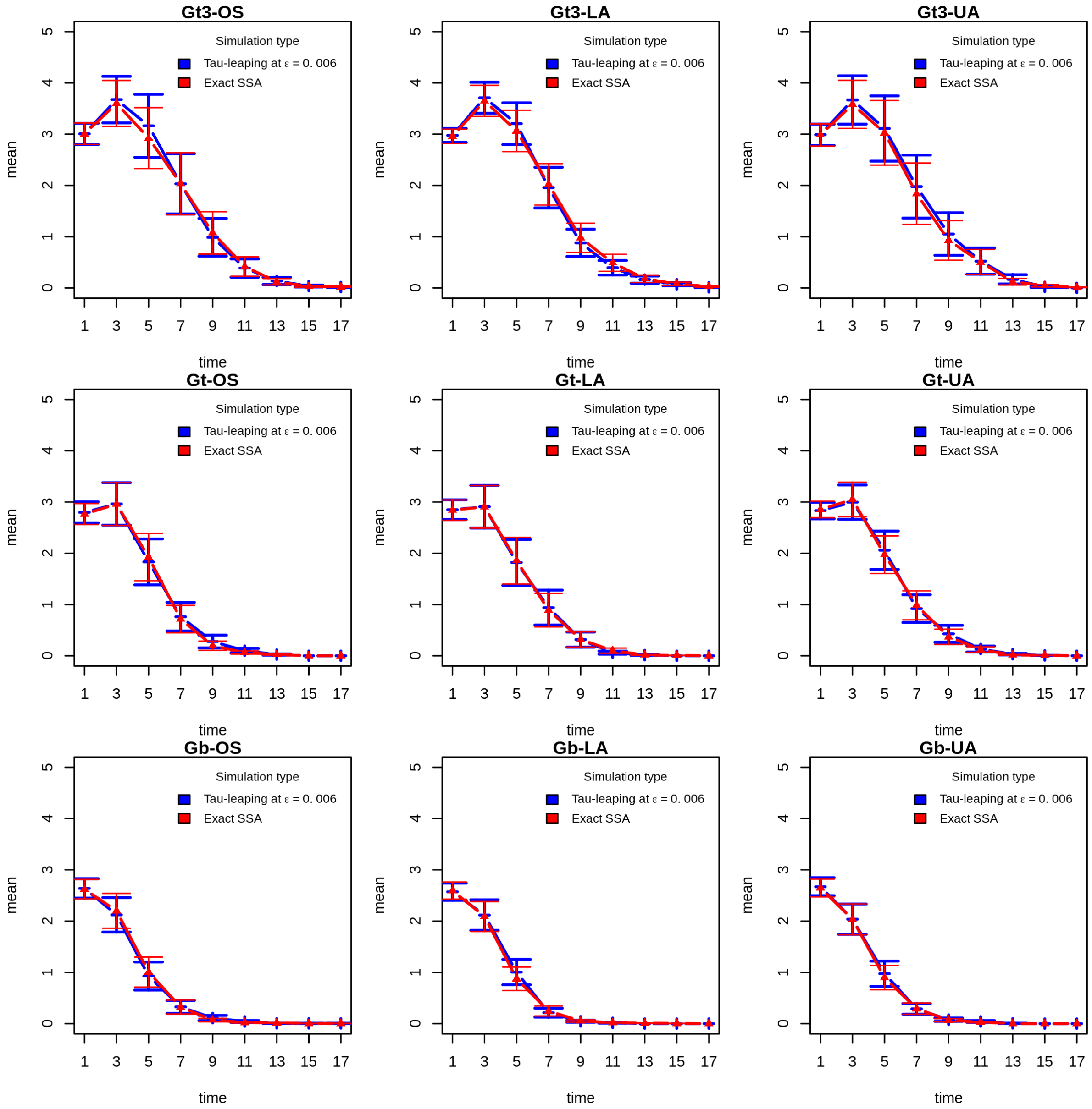

Figure 6.6: Mean comparison between exact SSA and Hybrid τ-leaping simulations at $\epsilon = 0.006$.

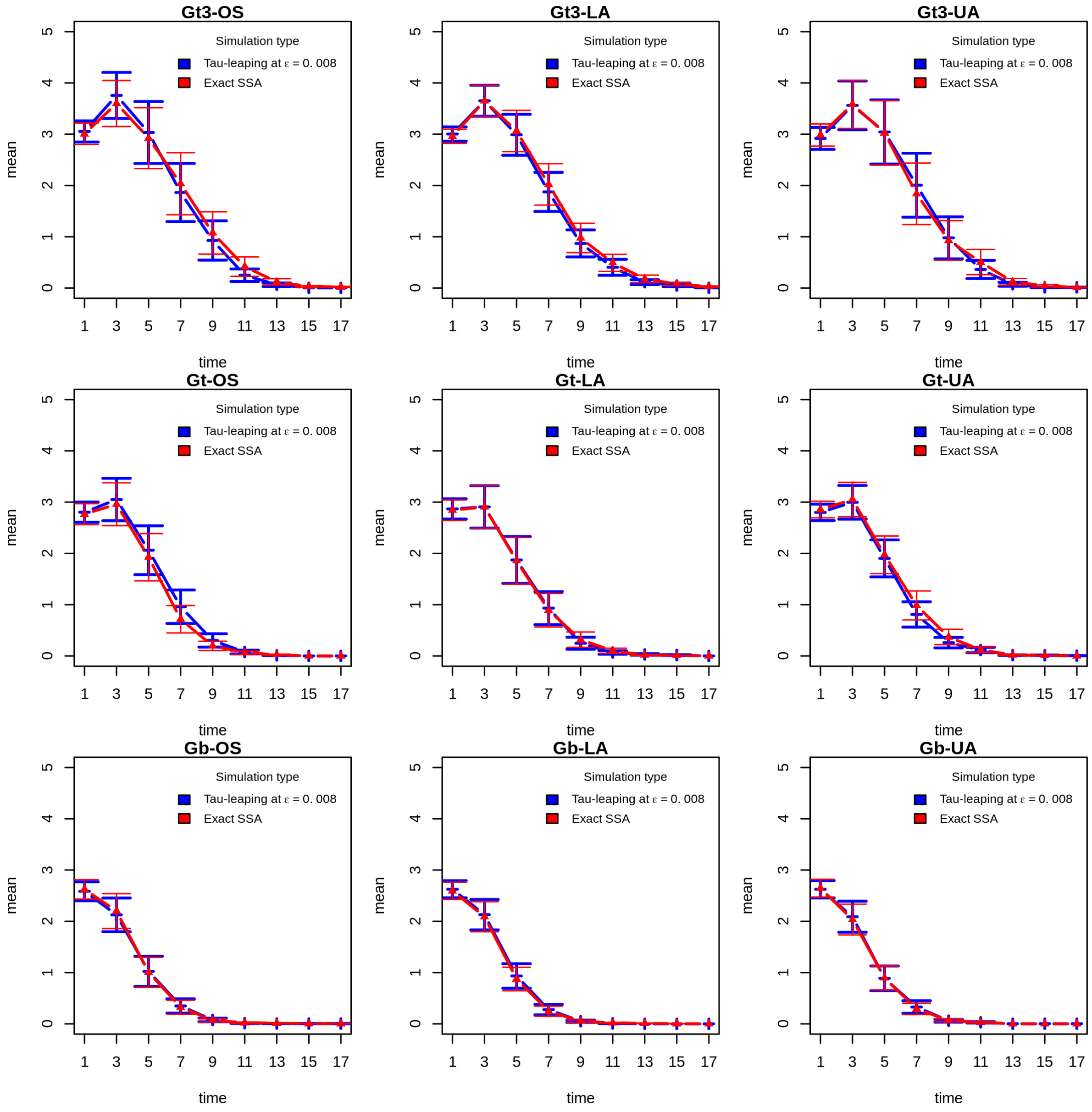

Figure 6.7: Mean comparison between exact SSA and Hybrid τ-leaping simulations at $\epsilon = 0.008$.

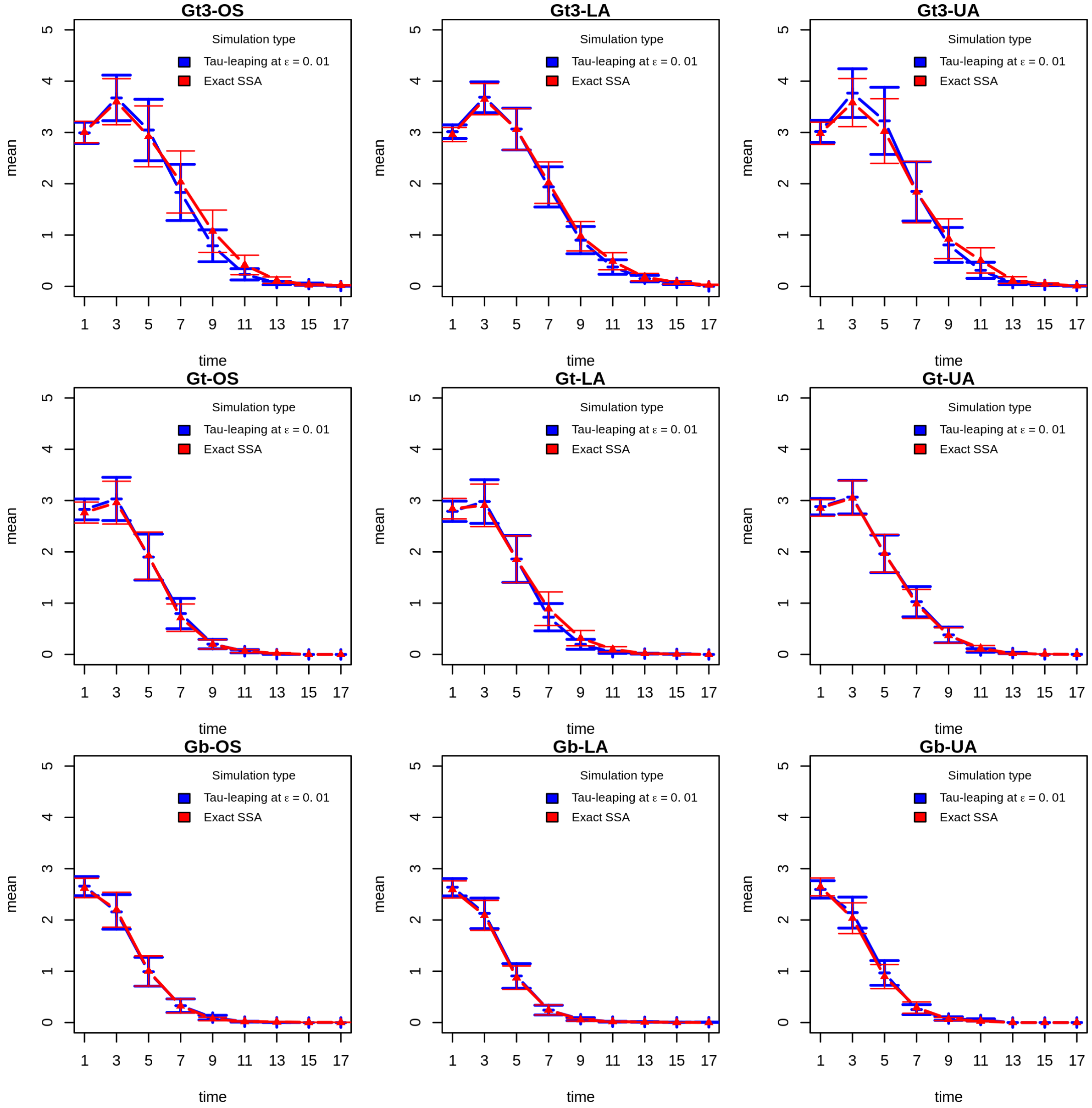

Figure 6.8: Mean comparison between exact SSA and Hybrid τ-leaping simulations at $\epsilon = 0.01$.

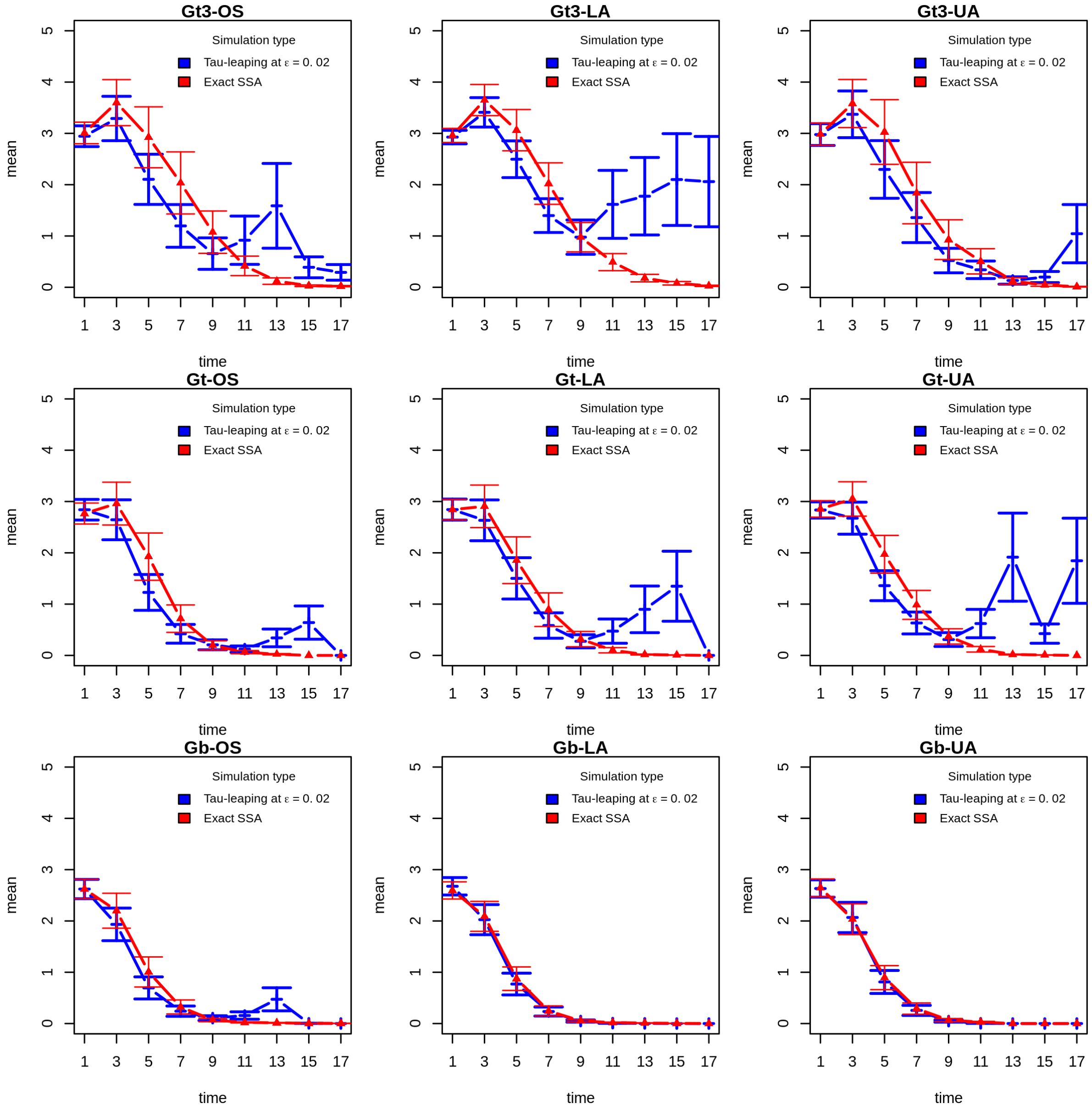

Figure 6.9: Mean comparison between exact SSA and Hybrid τ-leaping simulations at $\epsilon = 0.02$.

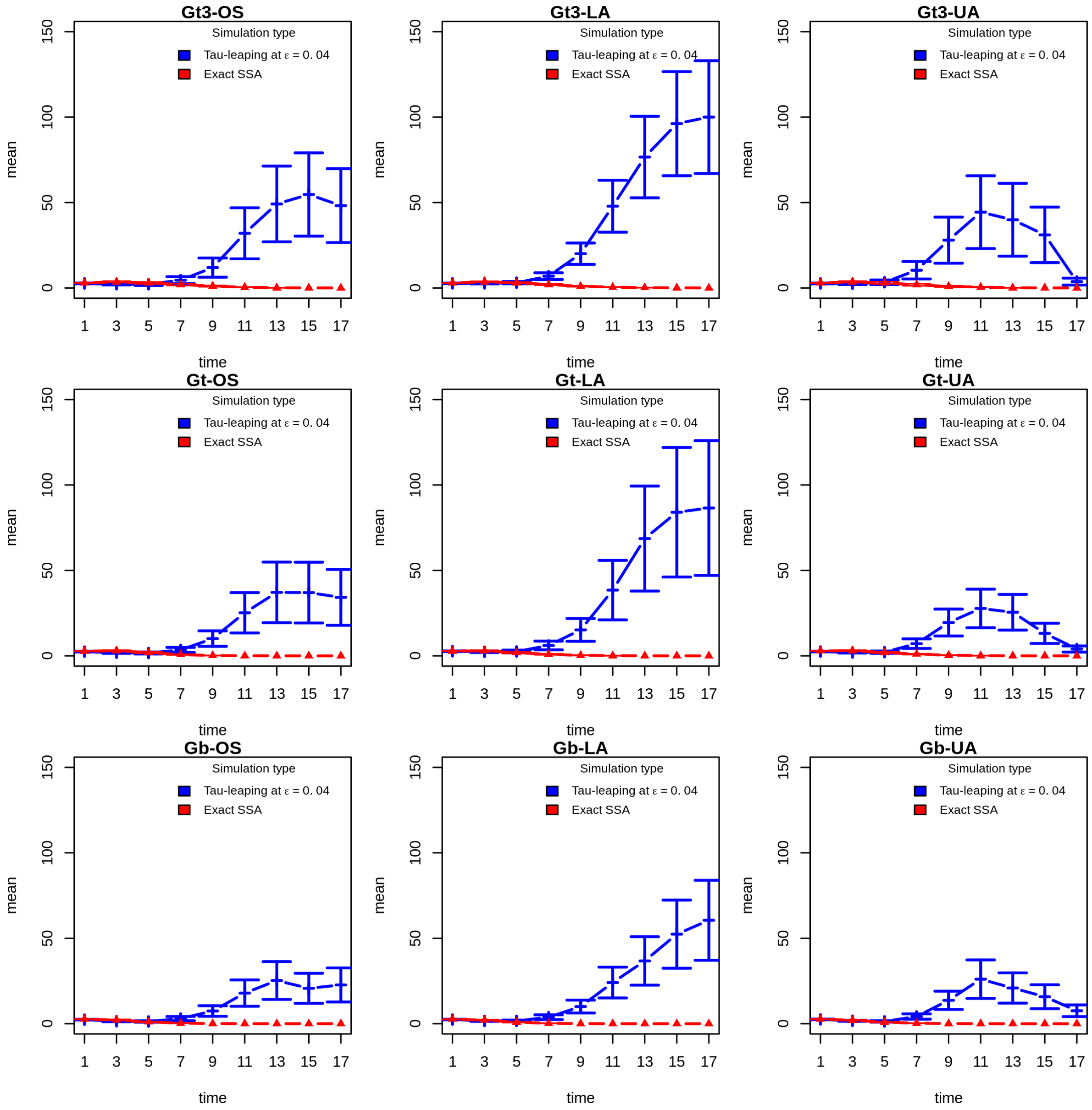

Figure 6.10: Mean comparison between exact SSA and Hybrid τ-leaping simulations at $\epsilon = 0.04$.

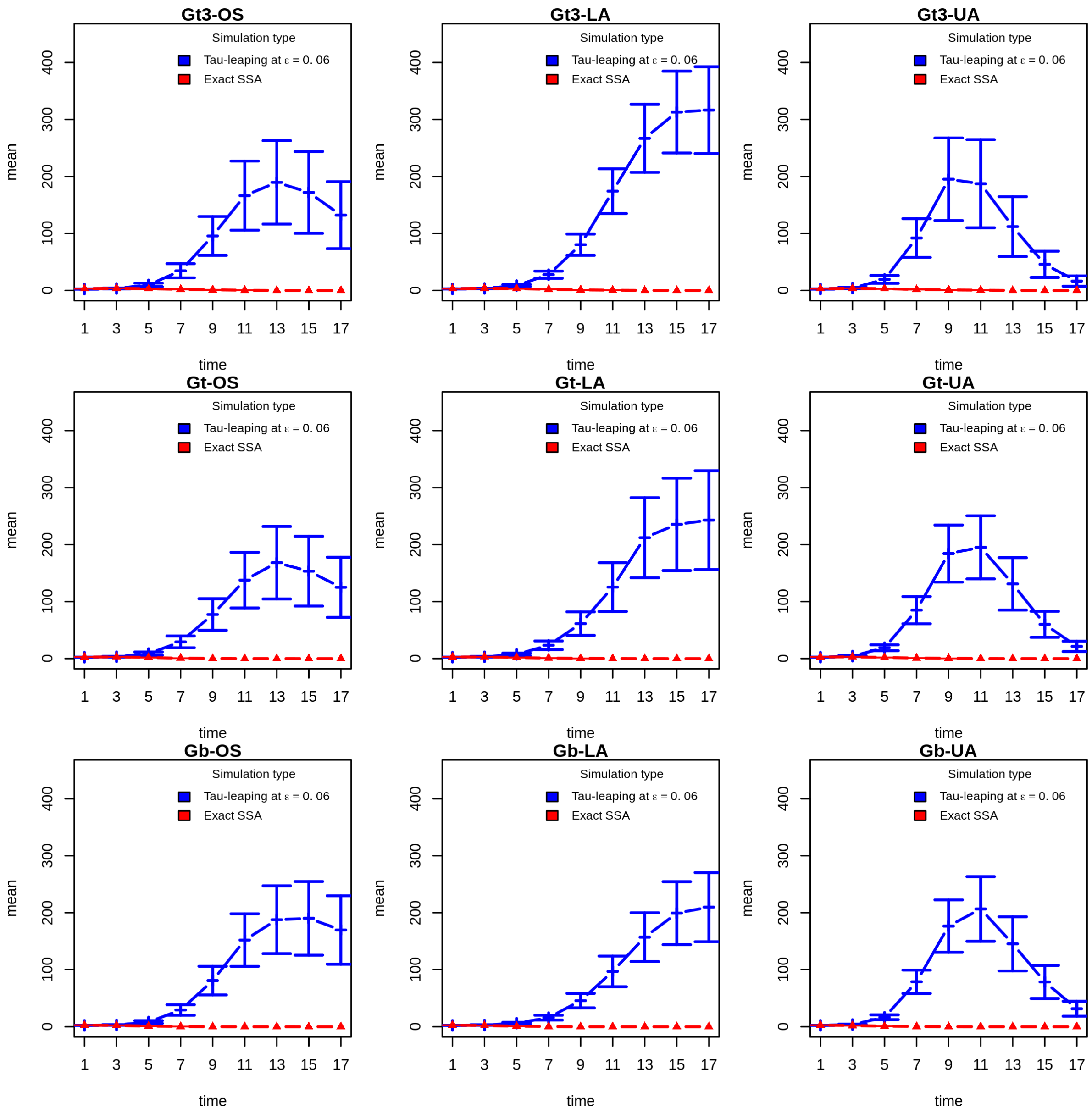

Figure 6.11: Mean comparison between exact SSA and Hybrid τ-leaping simulations at $\epsilon = 0.06$.

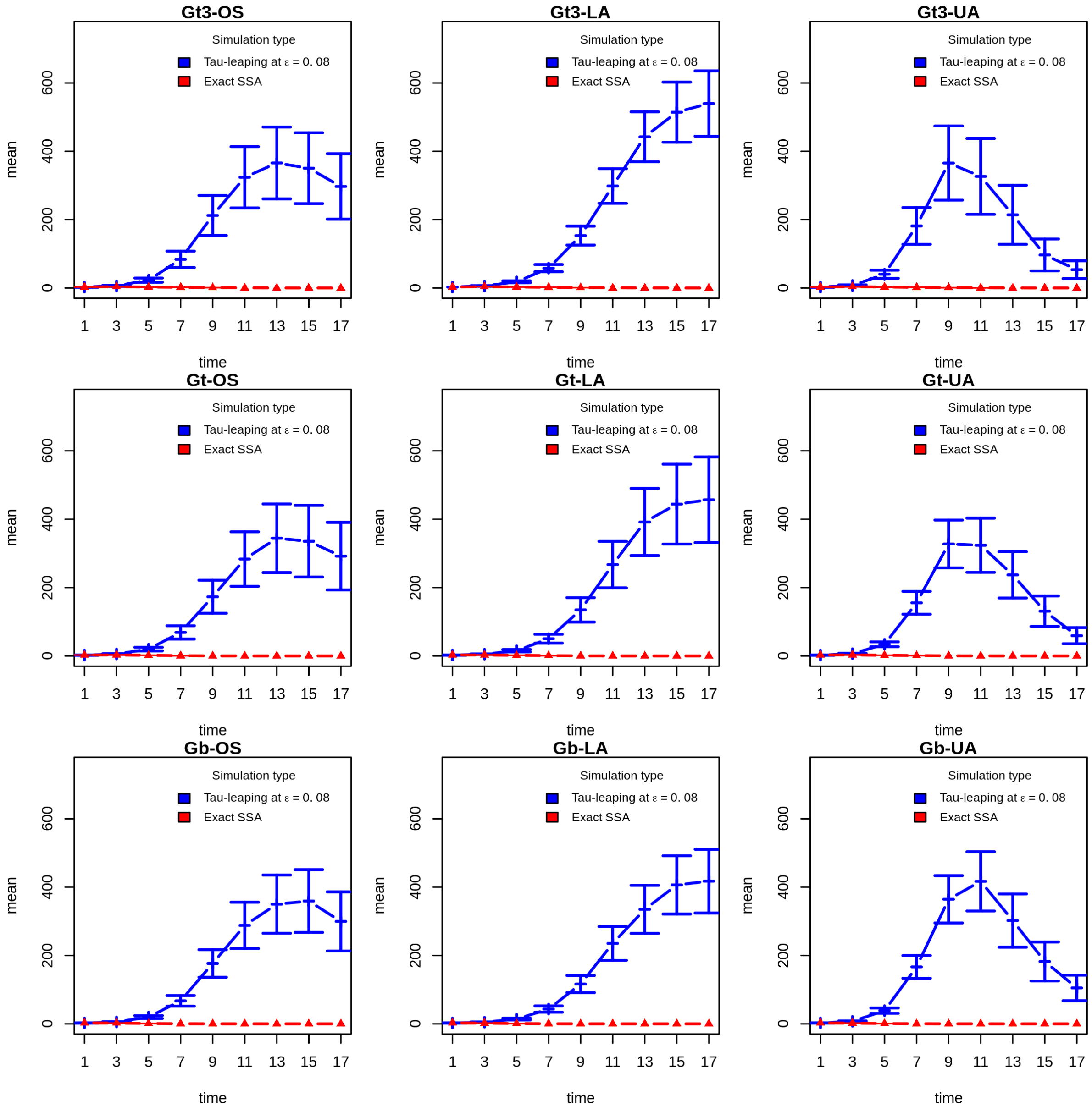

Figure 6.12: Mean comparison between exact SSA and Hybrid τ-leaping simulations at $\epsilon = 0.08$.

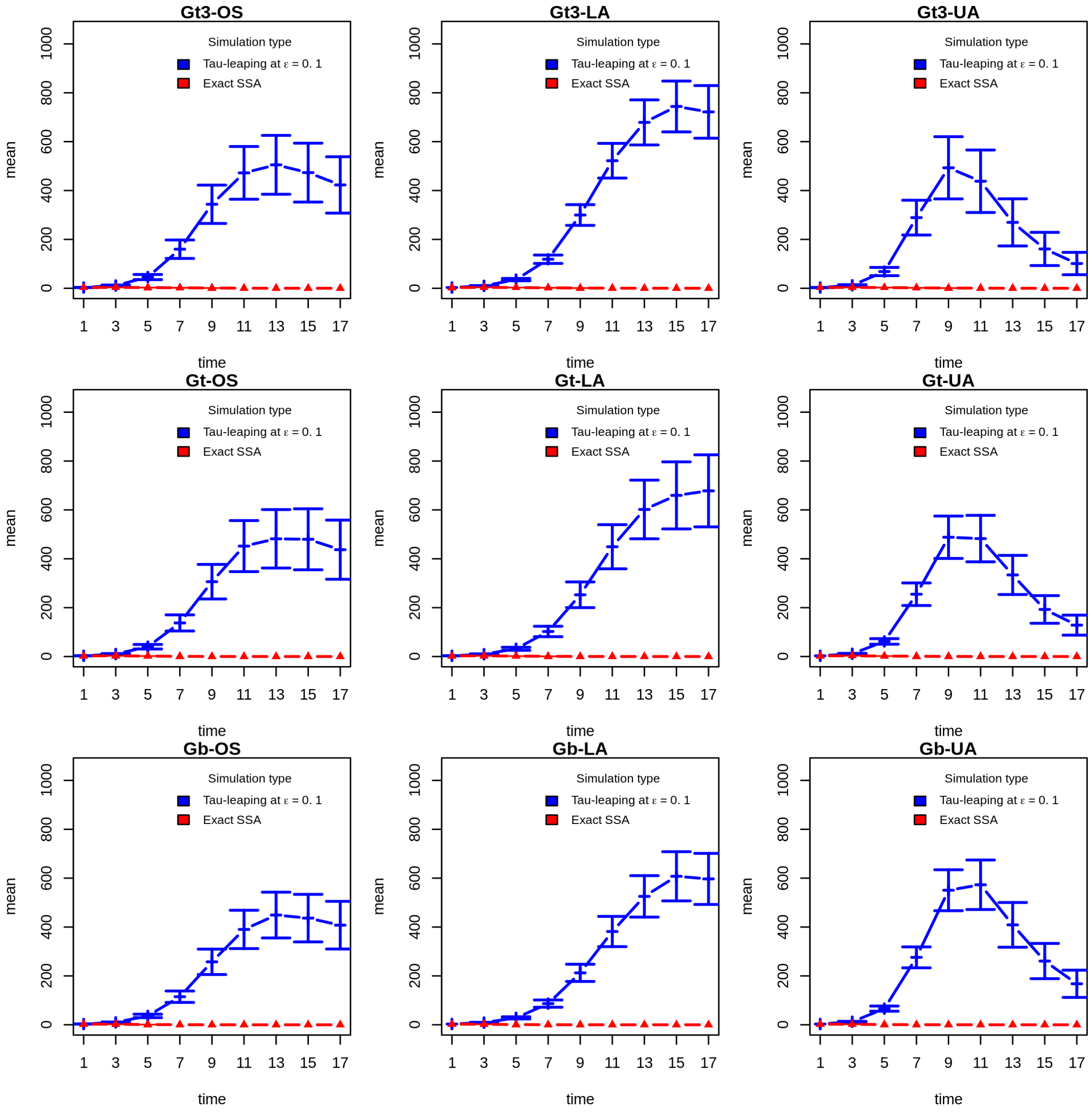

Figure 6.13: Mean comparison between exact SSA and Hybrid τ-leaping simulations at $\epsilon = 0.1$.

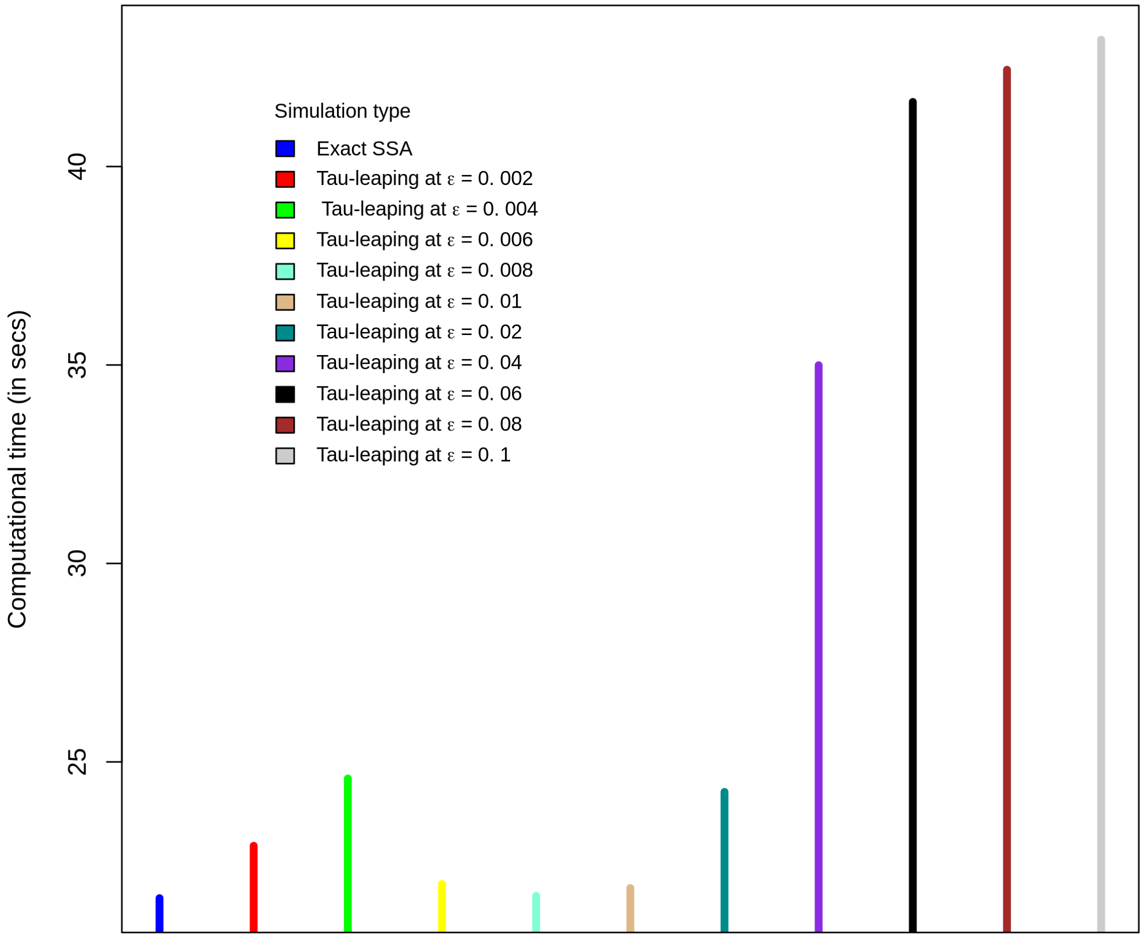

Figure 6.14: Comparison between computational time between exact SSA and Hybrid τ-leaping simulation at different error bounds $0 < \epsilon \leq 0.1$.

6.3 ABC fitting of the novel stochastic model

6.3.1 Introduction

This section presents the results of the ABC fitting of the multidimensional stochastic model with 23 parameters (outlined in section 6.2.1) using the weighted-iterative ABC at pre-specified tolerance thresholds and ABC-SMC total time steps as well as the modified regression adjustment (with $L2$ regularisation) for estimating the posterior means (motivated by findings from the numerical experiments presented in section 5.3.5). The results of ABC fitting of the complex stochastic simulation model as well as the ABC post-processing analysis using the modified local-linear regression with $L2$ regularisation are presented in section 6.3.2. Finally, section 6.3.3 presents the findings on hypothesis testing in relation to research questions 5-9 based on the adjusted posterior samples of the underlying model parameters.

6.3.2 ABC fitting of the novel multidimensional stochastic model

The complex stochastic model (with multi-parameters described in Table 6.1) was then successfully fitted using the proposed weighted-iterative ABC with sequential Monte Carlo and adaptive importance sampling outlined by Algorithm 4 at $N = 500$ (see Figure 6.15); where the overall ABC computational time was 46485.99 seconds. The modified local-linear regression model with $L2$ regularisation (described in section 5.3.4.2) was further used to obtain the adjusted posterior mean estimates of the model parameters (using equation 5.33). The ABC posterior distributions at were also adjusted based on equation 5.34, and the corresponding 95% credible intervals of the adjusted mean estimates were obtained for each parameter of the complex stochastic simulation model. Due to high multicollinearity between some of the regression predictors (in the neighbourhood of the observed summary statistics) as shown by Figure 6.16, Beaumont et al. [27] local-linear regression was impossible to implement.

Consequently, the proposed regression adjustment was able to deal with the high multicollinearity by shrinking the regression coefficients, resulting in predictors with minor contributions to the posterior samples (the outcome variable) having coefficients close to zero (but not equal to zero) in order to minimise their respective standard errors. Table 6.3 summarises the unadjusted and adjusted posterior mean estimates of the underlying model parameters with their respective 95% credible intervals. It can be observed from Table 6.3 that the width of the estimated 95% credible intervals based on the adjusted posterior distribution are relatively smaller compared to that of the unadjusted ABC posterior distribution. The goodness-of-fit density plots of the unadjusted and adjusted posterior distributions of the 23 parameters against the sequentially improving priors are shown by Figures 6.17–6.20. It can be concluded based on the fitted model that the effective (infrapopulation) carrying capacity at each of the four main body regions of the host (i.e., tail, lower region, upper region and head) is between 93 and 117 with an average number of 104 parasites per region (see Table 6.3).

Remark. The adjusted posterior distributions from the proposed regression adjustment method with $L2$ regularisation are further used to test several research hypotheses which aim to provide specific answers to the research questions numbered 5-9 (see section 6.3.3).

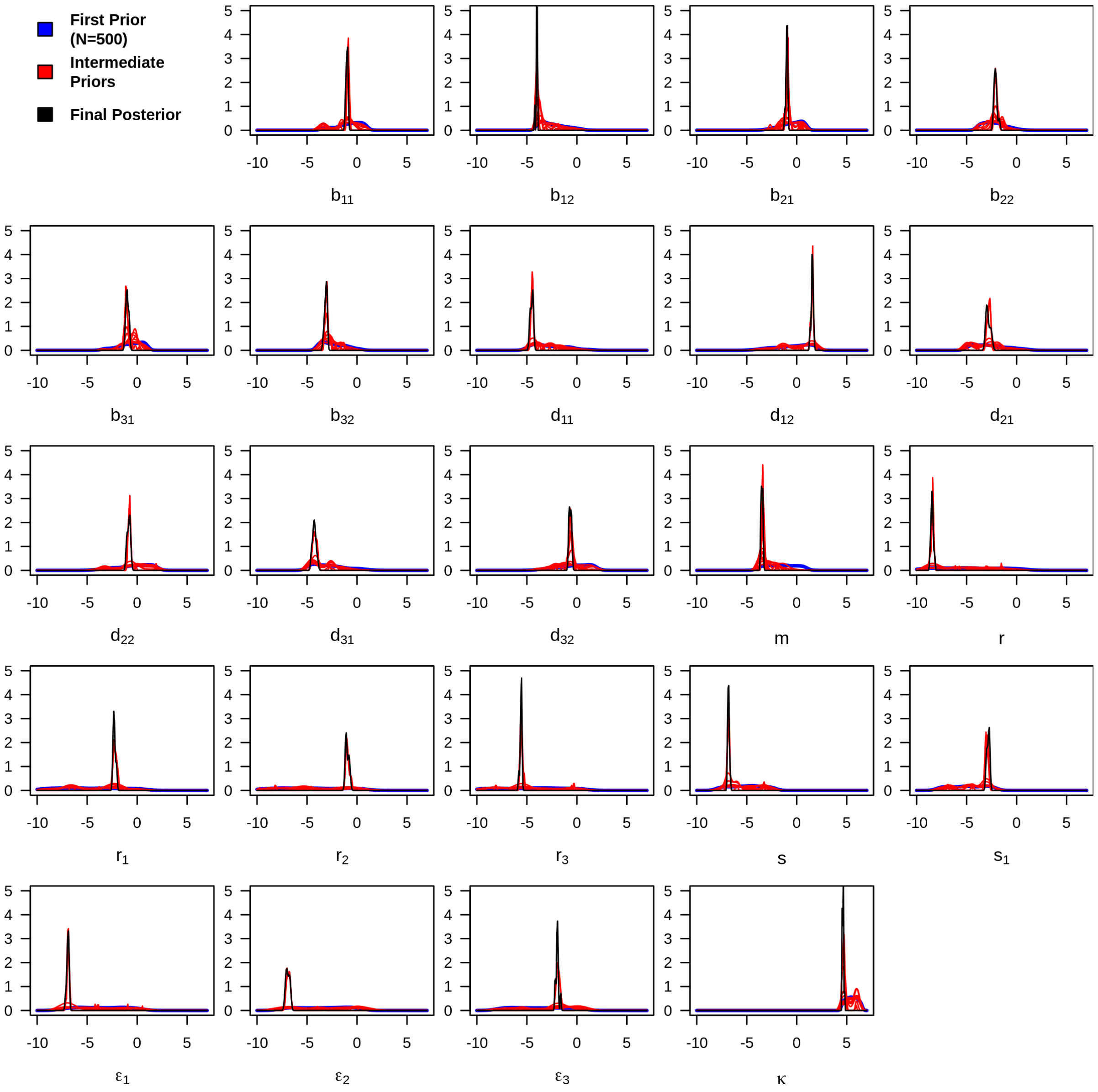

Figure 6.15: Goodness-of-fit density plots of the (marginal) approximate posterior distributions (in black) for the 23 parameters of the complex stochastic model against the sequentially improving prior distributions at $N = 500$ (on logarithmic scale).

Table 6.3: Unadjusted and adjusted posterior mean estimates of the 23 parameters of the multidimensional stochastic model with their respective 95% credible intervals (C.I.).

Parameters	Unadjusted mean	95% C.I.	Adjusted mean	95% C.I.
Base simulation parameters				
b_{11}	0.3664	0.3133—0.4093	0.3651	0.3120—0.4071
b_{12}	0.0181	0.0153—0.0191	0.0531	0.0407—0.0554
b_{21}	0.3682	0.3124—0.3987	0.3632	0.3121—0.3983
b_{22}	0.1236	0.1010—0.1659	0.1249	0.1011—0.1661
b_{31}	0.3778	0.3097—0.4515	0.5913	0.5275—0.7080
b_{32}	0.0444	0.0353—0.0502	0.0447	0.0353—0.0502
d_{11}	0.0110	0.0089—0.0126	0.0108	0.0089—0.0127
d_{12}	4.7294	3.8521—5.4808	4.2187	3.4749—4.8169
d_{21}	0.0594	0.0465—0.0818	0.0798	0.0667—0.1133
d_{22}	0.4346	0.3474—0.5247	0.4450	0.3466—0.5260
d_{31}	0.0141	0.0107—0.0186	0.0143	0.0107—0.0187
d_{32}	0.5179	0.4467—0.6026	0.5212	0.4471—0.6031
m	0.0307	0.0276—0.0338	0.0231	0.0211—0.0242
r	2.15×10^{-4}	1.79×10^{-4}—2.63×10^{-4}	2.16×10^{-4}	1.77×10^{-4}—2.60×10^{-4}
s	1.09×10^{-3}	9.95×10^{-4}—1.26×10^{-3}	1.13×10^{-3}	1.04×10^{-3}—1.26×10^{-3}
κ	103.597	93.503—116.387	104.546	93.439—116.291
Additional simulation parameters				
ϵ_1	9.85×10^{-4}	7.99×10^{-4}—1.12×10^{-3}	1.22×10^{-3}	1.01×10^{-3}—1.40×10^{-3}
ϵ_2	1.00×10^{-3}	7.69×10^{-4}—1.25×10^{-3}	9.62×10^{-4}	7.69×10^{-4}—1.26×10^{-3}
ϵ_3	0.1427	0.1136—0.1857	0.2778	0.2417—0.3569
r_1	0.1034	0.0088—1.2791	0.1029	0.0088—0.1279
r_2	0.3824	0.3240—0.4859	0.4009	0.3505—0.5216
r_3	3.74×10^{-3}	3.09×10^{-3}—4.11×10^{-3}	3.69×10^{-3}	3.09×10^{-3}—4.12×10^{-3}
s_1	0.0588	0.0480—0.0685	0.0591	0.0481—0.0685

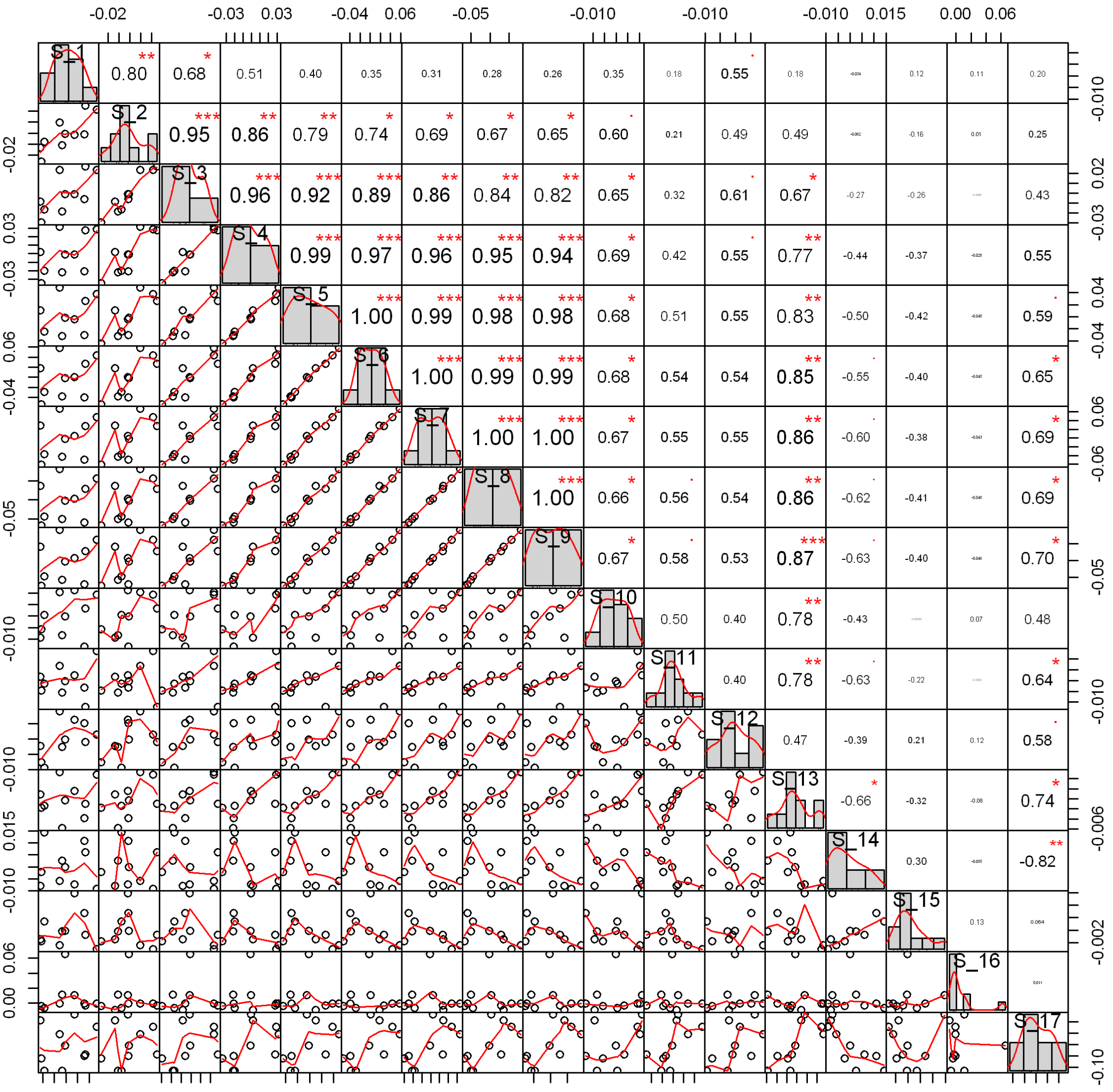

Figure 6.16: Correlation matrix plot indicating high multicollinearity between some of the 17 regression predictors (denoted by S_i, $1 \leq i \leq 17$ in the neighbourhood of the observed summary statistics) in the modified regression-adjusted ABC (with $L2$ regularisation).

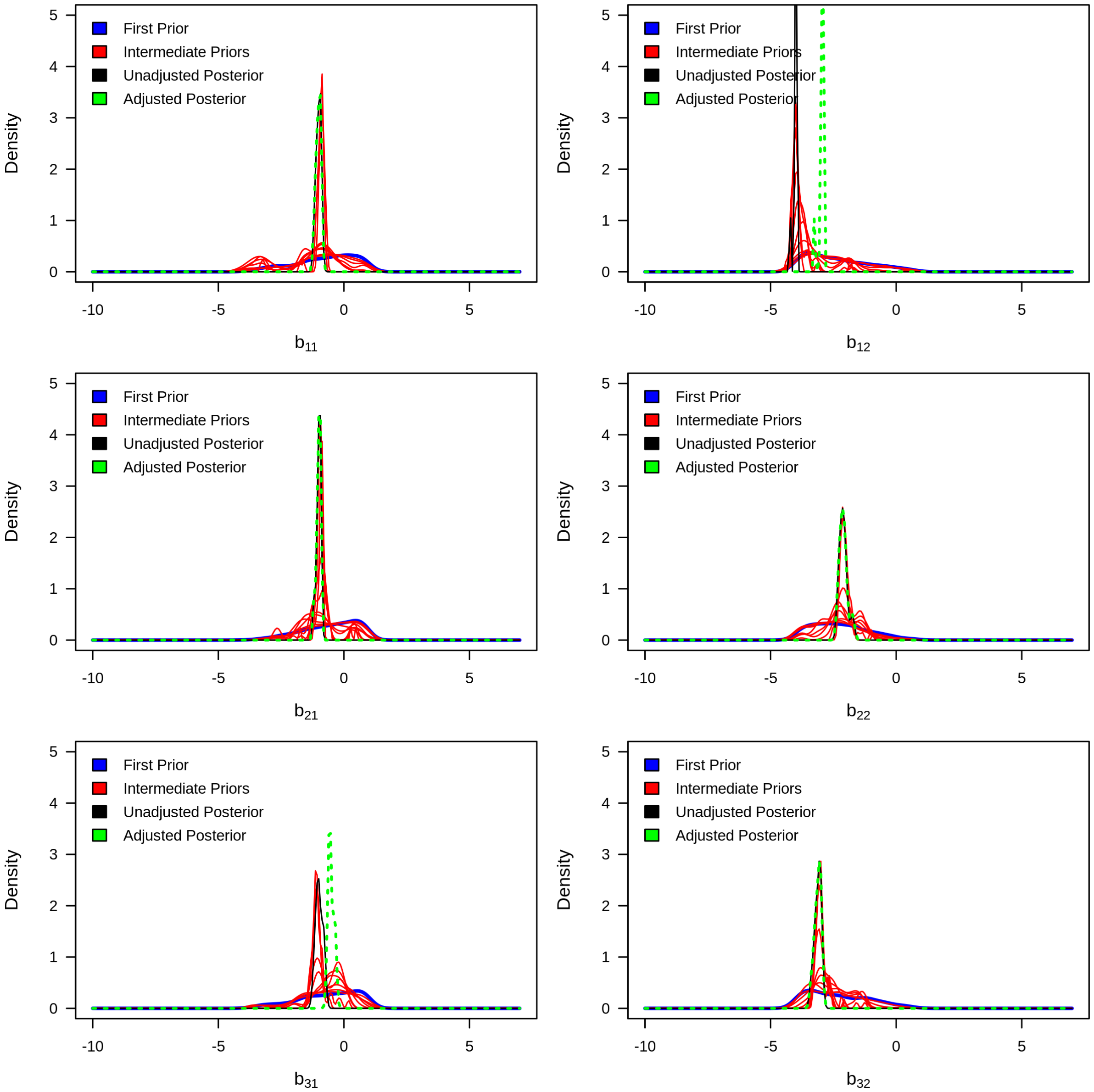

Figure 6.17: Goodness-of-fit density plots of the unadjusted (in black) and adjusted (in green) posterior distributions of model parameters: b_{11}, b_{12}, b_{21}, b_{22}, b_{31}, and b_{32} against the sequentially improving prior distributions (on logarithmic scale).

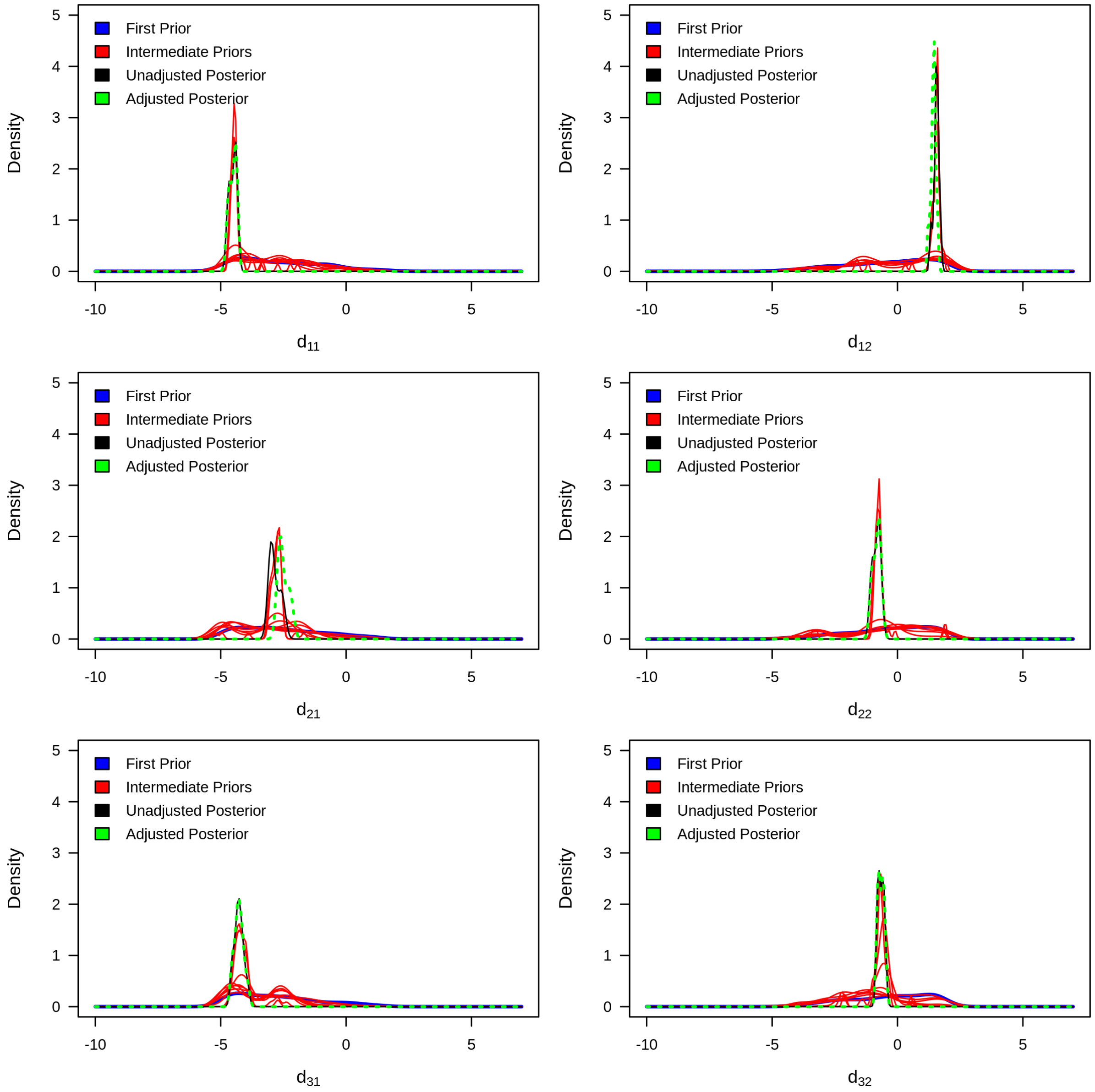

Figure 6.18: Goodness-of-fit density plots of the unadjusted (in black) and adjusted (in green) posterior distributions of model parameters: d_{11}, d_{12}, d_{21}, d_{22}, d_{31}, and d_{32} against the sequentially improving prior distributions (on logarithmic scale).

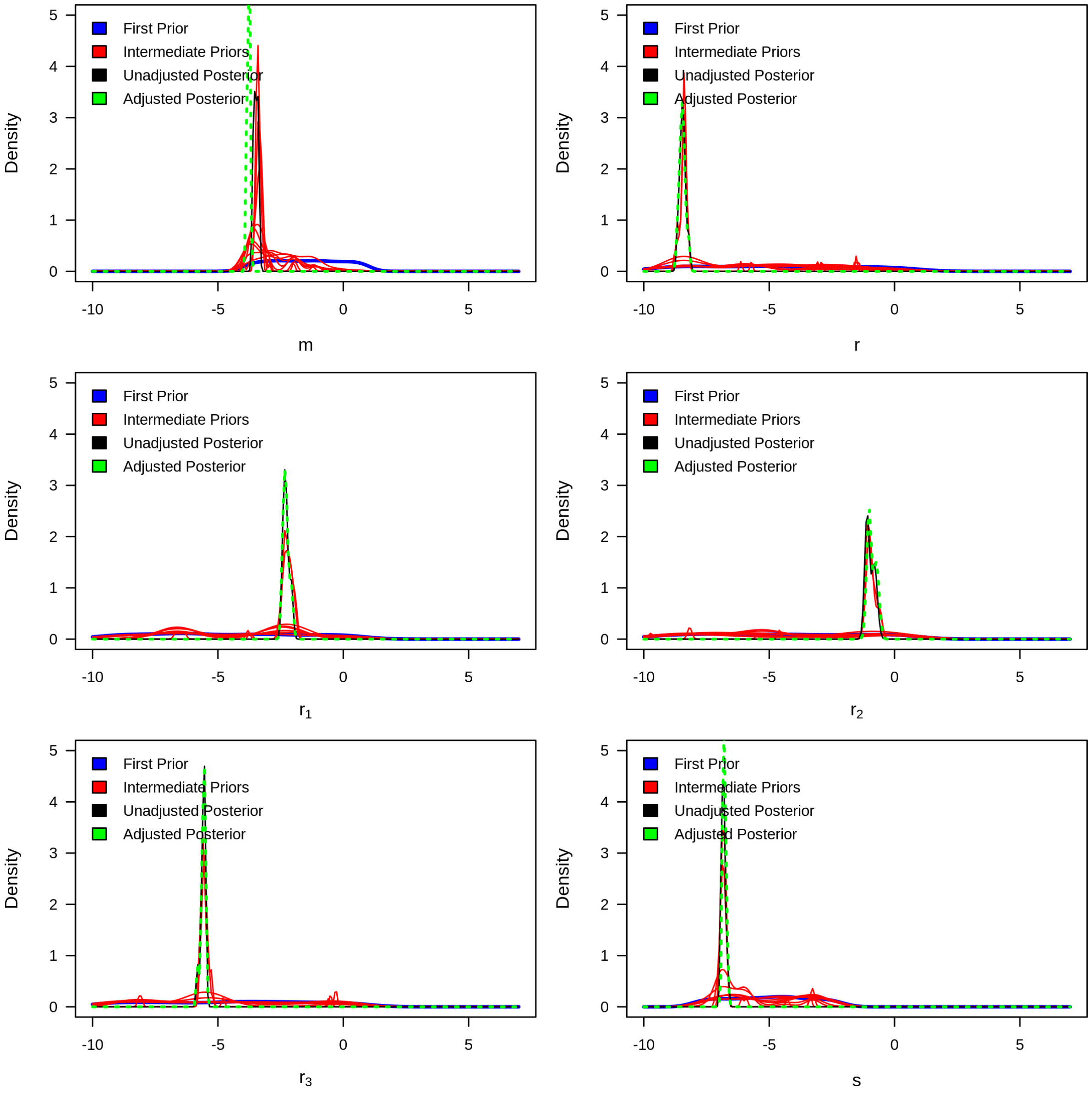

Figure 6.19: Goodness-of-fit density plots of the unadjusted (in black) and adjusted (in green) posterior distributions of model parameters: m, r, r_1, r_2, r_3, and s against the sequentially improving prior distributions (on logarithmic scale).

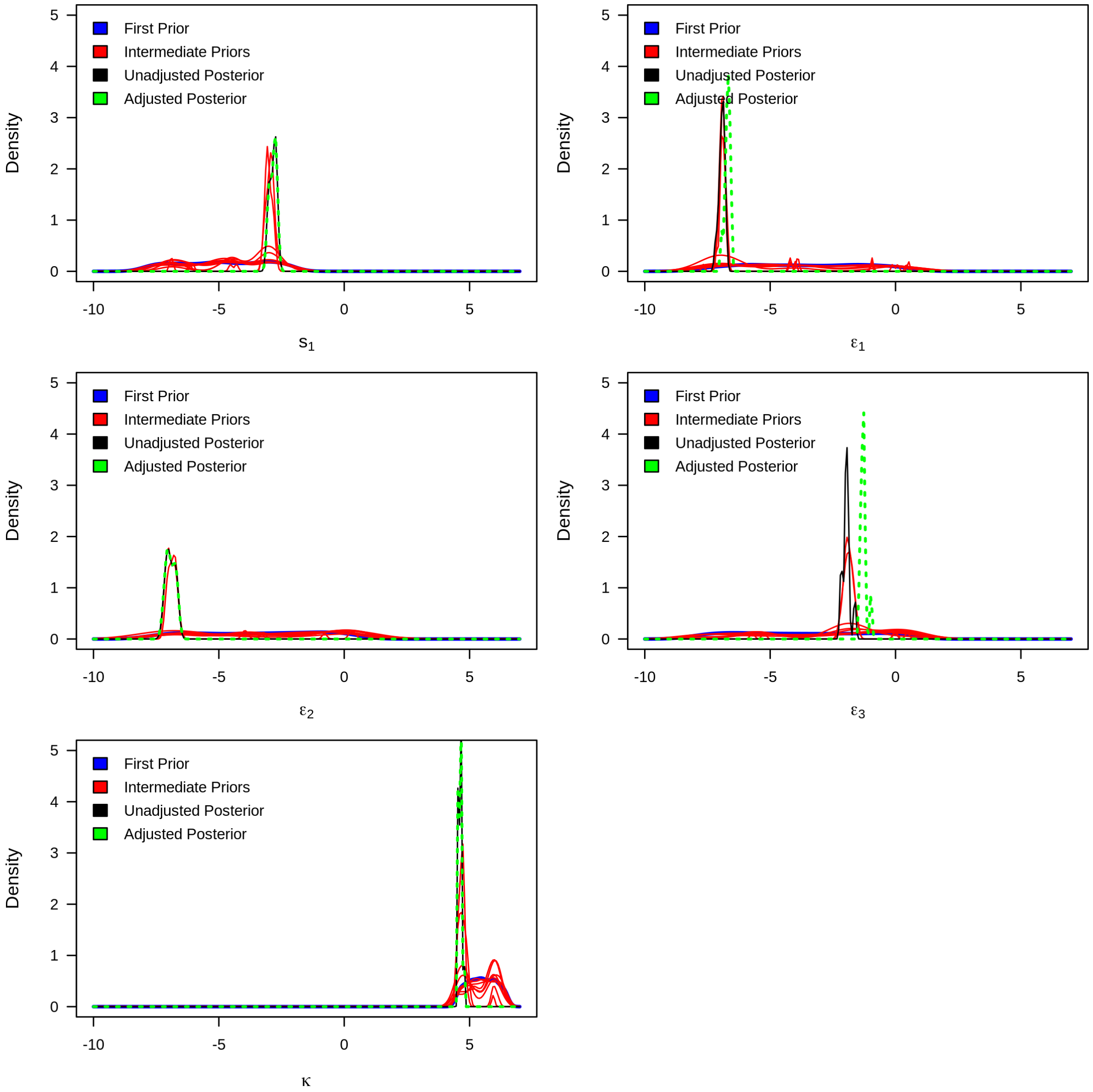

Figure 6.20: Goodness-of-fit density plots of the unadjusted (in black) and adjusted (in green) posterior distributions of model parameters: s_1, ϵ_1, ϵ_2, ϵ_3, and κ against the sequentially improving prior distributions (on logarithmic scale).

6.3.3 Bayesian hypothesis testing based on adjusted posterior samples

6.3.3.1 Introduction

Classical null hypothesis significance testing (NHST) often uses a dichotomous decision rule to conclude on a parameter value of interest (i.e., the null value) based on either *Pvalue* of a test statistic or an estimated confidence interval of the underlying parameter. For the latter method (which is preferred over the highly criticised *Pvalue*-dependent NHST decision [186]), we reject the parameter value under the null hypothesis if it falls outside a confidence interval. Nonetheless, confidence intervals cannot correctly capture the uncertainty about parameters and usually suffer from coverage probability issues [317]. Other studies attempt to apply similar logic to Bayesian posterior distributions and reject a parameter value if it falls outside a credible posterior interval [182]. According to Kruschke and Liddell [182], this standard decision rule causes two statistical issues. First, it can only reject and never accept a parameter value. Second, even if a null value is true, the decision process will eventually reject it, given large posterior samples of the underlying parameter. Other studies have proposed a more accurate decision rule, analogous to frequentist equivalence testing [263, 316]. This new Bayesian approach requires the integration of a region of practical equivalence (ROPE) around the null value and an estimated $100(1-\alpha)\%$ highest density interval (HDI) [182]. Consequently, it has been recommended that if an HDI is used to evaluate null values as part of a decision rule, the decision should also rely on a ROPE around the null value [180, 213]. In other words, a null value should not be rejected simply because it falls outside an HDI, as observed in previous studies [181]. Therefore, it has been recommended to reject the null only when HDI strictly falls outside the ROPE (meaning the parameter's most credible values are not practically equivalent to the null value). We then accept the null if the HDI lies entirely within the ROPE, and we remain indecisive if there is an overlap [182, 269]. For a wide range of Bayesian decisions using ROPE (including more technical reports), see work by Schwaferts and Augustin [269].

In the current study, we simultaneously used the ROPE and HDI (which is dubbed in the literature as ROPE+HDI) to test relevant hypotheses concerning differences between some underlying parameters of our novel stochastic simulation model with the help of the adjusted posterior samples and the bayestestR package in R [204]. McElreath [213] and Kruschke [180] have recommended an 89% HDI to be an ideal choice compared to the usual 95% HDI for Bayesian hypothesis testing with ROPE. According to Kruschke [180] the 95% HDI might not be the most appropriate for Bayesian posterior distributions due to potentially lacking stability if not enough posterior samples are drawn (as observed in the current study). Hence, an appropriate ROPE and an 89% HDI are considered for testing sets of hypotheses, respectively. Results from the Bayesian hypothesis test will aid in providing answers to research questions 5-8. Now, let suppose a null hypothesis $H_0 : \theta_1 = \theta_2$ (or $d = \theta_1 - \theta_2 = 0$), where $\theta_g \in R$ denotes model parameters corresponding to some independent groups $g = 1, 2$ (possibly identically distributed). The alternative hypothesis is defined as $H_1 : \theta_1 \neq \theta_2$ (or $d = \theta_1 - \theta_2 \neq 0$). Let $\mathcal{A}_{\mathrm{I}} = \{[a,b] \mid a,b \in \Theta, a < b\}$ represent the action space w.r.t the HDI of the posterior distribution of $d = \theta_1 - \theta_2$, and let $\mathcal{A}_R = [-0.5\sigma_d, 0.5\sigma_d]$ denote the ROPE range [recommended by 226], where σ_d is the standard deviation of the posterior samples of d. Let also suppose $\gamma = P(\mathcal{A}_{\mathrm{I}} \subseteq \mathcal{A}_R \mid d)$ denote the ROPE coverage probability (or the probability that elements of $\mathcal{A}_{\mathrm{I}}$ fall within $\mathcal{A}_R$ given the posterior samples of d). Following Kruschke and Liddell [182], we also reject or accept H_0 according to the following HDI+ROPE decision rule:

$$\text{ROPE equivalence decision} = \begin{cases} \text{reject} \quad H_0, & \gamma = 0 \\ \text{indecisive}, & 0 < \gamma < 1 \\ \text{accept} \quad H_0, & \gamma = 1. \end{cases}$$

The null hypothesis and the ROPE+HDI test described above can be modified to compare differences between model parameters corresponding to more than two groups similarly (as done to the subsequent sections). For the main accompanying R codes on the Bayesian hypothesis testing based on the adjusted posterior samples, see Appendix G.9.

6.3.3.2 Assessing differences between the birth rate model parameters

We first tested three major hypotheses in relation to the birth rate parameters of the fitted stochastic model based on ROPE+HDI Bayesian tests (Table 6.4). The null hypotheses tested were as follows:

H_{01}: $b_{i1} - b_{j1} = 0$, for $i \neq j$ and $1 \leq i, j, \leq 3$.

H_{02}: $b_{i2} - b_{j2} = 0$, for $i \neq j$ and $1 \leq i, j, \leq 3$.

H_{03}: $b_{i1} - b_{j2} = 0$, for $i = j$ and $1 \leq i, j, \leq 3$.

For the first null hypotheses (H_{01}), there were no enough evidence to either accept or reject H_{01}. Thus, it can be inferred that the birth rates of young gyrodactylids (i.e., daughters yet to reproduce) may or may not be significantly different across the three parasite strains. However, second (H_{02}) and third (H_{03}) null hypotheses where rejected, respectively. This confirms that the birth rates of old parasites (i.e., mothers) differ significantly across the three parasite strains with the old *Gt* parasites having the highest birth rate (with the birth rate of old *Gb*>*Gt3*); whereas the birth rate of young parasites are significantly greater than the mothers irrespective of their strain (Table 6.4).

Because gyrodactylids are protogynous (with female reproductive organs developing before the male organs) and the first daughter (i.e., the younger parasite) most likely born asexually, this long-lived strategy guarantees that a high proportion of the older population has a functional male system, whereas asexually derived younger parasites make up a lesser proportion of the entire population [128]. This could explain the indecision regarding H_{01}.

Table 6.4: Results from the test of statistical differences between the birth rate parameters.

Parameter	89% HDI	ROPE range	ROPE coverage (%)	Decision
First hypotheses				
$b_{11} - b_{21}$	-0.0426 — 0.0376	-0.0142 — 0.0142	62.5	indecisive
$b_{11} - b_{31}$	-0.0802 — 0.0729	-0.0308 — 0.0308	12.5	indecisive
$b_{21} - b_{31}$	-0.0740 — 0.0760	-0.0276 — 0.0276	37.5	indecisive
Second hypotheses				
$b_{12} - b_{22}$	-0.1372 — -0.0833	-0.0107 — 0.0107	0	rejected
$b_{12} - b_{32}$	-0.0333 — -0.0182	-0.0029 — 0.0029	0	rejected
$b_{22} - b_{32}$	0.0591 — 0.1091	-0.0101 — 0.0101	0	rejected
Third hypotheses				
$b_{12} - b_{12}$	0.3024 — 0.3903	-0.0173 — 0.0173	0	rejected
$b_{21} - b_{22}$	0.1825 — 0.2955	-0.0220 — 0.0220	0	rejected
$b_{31} - b_{32}$	0.2694 — 0.4125	-0.0275 — 0.0275	0	rejected

6.3.3.3 Assessing differences between the death rate model parameters

Also, we tested three major hypotheses concerning the death rate model parameters (Table 6.5). The null hypotheses tested were as follows:

H_{04}: $d_{i1} - d_{j1} = 0$, for $i \neq j$ and $1 \leq i, j, \leq 3$.

H_{05}: $d_{i2} - d_{j2} = 0$, for $i \neq j$ and $1 \leq i, j, \leq 3$.

H_{06}: $d_{i1} - d_{j2} = 0$, for $i = j$ and $1 \leq i, j, \leq 3$.

With the exception of the fifth null hypotheses (H_{04}) which was inconclusive for one of its tests, Table 6.5 showed that the other null hypotheses (H_{05} and H_{06}) were rejected. These findings implies that, in the absence of host immune response, the death rate of the wild *G. turnbulli* is significantly higher than the other two parasite strains. However, in the presence of host response (due to potentially rapid infrapopulation growth, high parasite virulence or intense competition for resources), laboratory-bred *G. turnbulli* had the highest rate of death (with that of that of $Gb{>}Gt$). It can also be inferred across all the parasite strains that the immune response induced death rates was far greater than the case of no response.

In addition to the low birth rate of old *Gt3* parasite strain comparatively (as discovered in section 6.3.3.2), the high death of *Gt3* parasites and low death rate of the two wild parasite strains after adaptive immune response, can also explain the low parasite mean intensity and relatively low parasite load found in the *Gt3* parasite population over time as discovered in the spatial-temporal analysis (refer to Chapter 2). Although temperature also controls population dynamics of gyrodactylids, it has been shown in other studies of *Gyrodactylus* that adaptive host immunity (which develops in most fish populations) is responsible for the extinction of gyrodactylid populations on a fish host [see 265].

Table 6.5: Results from the test of statistical differences between death rate parameters.

Parameter	89% HDI	ROPE range	ROPE coverage (%)	Decision
Fourth hypotheses				
$d_{11} - d_{21}$	-0.0674 — -0.0362	-0.0062 — 0.0062	0	rejected
$d_{11} - d_{31}$	-0.0072 — 0.0013	-0.0017 — 0.0017	37.5	indecisive
$d_{21} - d_{31}$	0.0338 — 0.0677	-0.0072 — 0.0072	0	rejected
Fifth hypotheses				
$d_{12} - d_{22}$	3.5664 — 4.9886	-0.2647 — 0.2647	0	rejected
$d_{12} - d_{32}$	3.4814 — 4.9070	-0.2597— 0.2597	0	rejected
$d_{22} - d_{32}$	-0.1465— 0.0087	-0.0312— 0.0312	0	rejected
Sixth hypotheses				
$d_{12} - d_{12}$	-5.3631— -3.9705	-0.2538—0.2538	0	rejected
$d_{21} - d_{22}$	-0.4622— -0.2694	-0.0373 — 0.0373	0	rejected
$d_{31} - d_{32}$	-0.5764 — -0.4370	-0.0286 — 0.0286	0	rejected

6.3.3.4 Assessing differences between the movement rate adjustment parameters

We further tested differences between movement rate adjustment parameters across the three parasite strains (Table 6.6). The strain-specific movement rate adjustment parameters are expected to account for the unique caudal-rostral preferences of the gyrodactylid strains in the simulation model (as confirmed in Chapter 2). Here, the null hypotheses were defined as follows:

H_{07}: $\epsilon_i - \epsilon_j = 0$, for $i \neq j$ and $1 \leq i, j, \leq 3$.

Table 6.6 reveals an inconclusive decision concerning the statistical difference between *Gt3* and *Gt* movement rate adjustment parameters (possibly because they belong to the same species); however, the *Gb* movement rate adjustment parameter was significantly higher than the two strains of *G. turnbulli*. It can be inferred that the stochastic model was able to distinguish between the unique microhabitat preferences of the two distinct *Gyrodactylus* species previously justified in Section 2.3.2 (after initial infection at the caudal region of the host). Thus, the high movement rate of *Gb* parasite strain may be a possible reason why *Gb* parasites could rapidly move towards the rostral regions of their fish host (starting from the caudal region) over time, as discovered in the spatial-temporal analysis concerning the parasites' microhabitat preference (and it tends to prefer the head region of their host as the infection progresses). Nonetheless, the low movement rate associated with the two *G. turnbulli* strains after initial infection at the host's caudal region may imply that they are relatively less mobile (possibly due to tail preference) and, thus, tend to stay at the host's tail region for a more extended period.

Table 6.6: Results from the test of statistical differences between the movement rate adjustment parameters.

Parameter	89% HDI	ROPE range	ROPE coverage (%)	Decision
Seventh hypotheses				
$\epsilon_1 - \epsilon_2$	-0.00023 — 0.00024	-0.00009 — 0.00009	37.5	indecisive
$\epsilon_1 - \epsilon_3$	-0.1738 — -0.1140	-0.0115 — 0.0115	0	rejected
$\epsilon_2 - \epsilon_3$	-0.1737 — -0.1139	-0.0114 — 0.0114	0	rejected

6.3.3.5 Assessing differences between the immune response rate adjustment parameters as well as the sex-specific host mortality parameter

Finally, we tested two different hypotheses in relation to the immune response rate adjustment parameters and the sex-specific host mortality parameter, respectively. The null hypotheses of these tests are defined as:

H_{08}: $r_i - r_j = 0$, for $i \neq j$ and $1 \leq i, j, \leq 3$.

H_{09}: $s_1 = 0$.

Table 6.7 summarises the results on H_{08} and H_{09}. It was found that the immune response rate adjustment parameters were significantly different, and the model parameter s_1 (i.e., host mortality rate adjustment for male fish relative female fish) was significant from zero. These resulted helped to make inference on whether the adaptive host immune response is sex and host-dependent, and whether the mortality rate of male fish is not higher than female fish (based on the stochastic simulation and evidence drawn from the empirical data during ABC fitting). It can be inferred from Table 6.7 that male fish are more likely to die than female fish (which was kept as a reference category in the simulation model).

It can also be deduced that the immune response rate of OS fish was significantly greater than that of both Trinidadian stocks (with an immune response rate of LA fish > UA fish) and the male stock (relative female fish). The high immune response rate of OS stock against the gyrodactylid infection over time may confirm the findings from the previously fitted multi-state Markov model (see section 2.3.3.2), where the Ornamental fish was found to be less likely to die from gyrodactylid infection compared to the Trinidadian stocks (with LA fish less likely to die than UA fish) . It also suggests why OS fish had a higher mean parasite intensity as the infection progressed, leading to a potentially higher rate of adaptive immune response (relative to the other host populations). On the other hand, the high immune response of the male fish may also explain why the multi-state Markov model predicted the time for male fish to remain infected to be relatively lower than the infected female fish across all parasite strains, fish stocks, and host sizes.

Table 6.7: Results from the test of statistical differences between the immune response rate adjustment parameters and sex-specific host mortality parameter.

Parameter	89% HDI	ROPE range	ROPE coverage (%)	Decision
Eighth hypotheses				
$r_1 - r_2$	-0.3521 — -0.2301	-0.0267 — 0.0267	0	rejected
$r_1 - r_3$	0.0854 — 0.1221	-0.0068 — 0.0068	0	rejected
$r_2 - r_3$	0.3219 — 0.4711	-0.0323 — 0.0323	0	rejected
Ninth hypothesis				
s_1	0.0483 — 0.0676	-0.0038 — 0.0038	0	rejected

Chapter 7

Conclusions

This chapter first summarises how we addressed the research questions of the study (section 7.1). An overview of the entire study, highlighting the main contributions (which are of both mathematical and biological relevance), is provided in section 7.2. Finally, we outline directions for future works on the gyrodactylid-fish system in section 7.3.

7.1 Summary of answers to the study's research questions

The study focused on the infection dynamics of a gyrodactylid-fish system by using novel mathematical and stochastic simulation models to add to our understanding of the system. For the first time in the current study, we have provided answers to the nine major research questions (outlined in section 1.6):

- The first research question wanted to determine whether the caudal and rostral preferences of the gyrodactylid strains were consistent over time and across different fish stocks with the help of existing empirical data. Here, we adapted a rank-based multivariate Kruskal-Wallis test coupled with its *post-hoc* tests and informative graphical summaries (in Chapter 2) to investigate the spatial and temporal parasite distribution of three different gyrodactylid strains across three host populations (OS, LA and UA stocks). Two out of the three parasite strains were *Gyrodactylus turnbulli*, a laboratory-bred strain (*Gt3*) and a wild *turnbulli* strain (*Gt*); whereas the third strain was *G. bullatarudis*, also a wild type. We revealed that *Gt3* and *Gb* strains preferred the caudal and rostral regions respectively across different fish stocks; however, *Gt* strain changed microhabitat preference over time, indicating microhabitat preference of gyrodactylids is host and time-dependent.

- The second research question asked whether fish sex, fish size, fish stock, and parasite strain affect gyrodactylid infection progression (host recovery and host mortality over time). By adopting a time-inhomogeneous multi-state Markov model (MSM), we improved previous estimates of survival probabilities (under Chapter 2). We further showed that: i) parasite-related mortalities are host, sex, and time-dependent, and ii) fish size is the key determinant of host recovery.

- The third research question wanted to know the average infection time of infected fish conditioned on the significant predictors. Here, we derived exact mathematical expressions (in the time-inhomogeneous case) to estimate other relevant epidemiological quantities such as the mean time of host to remain infected and probability of infected host to either recover or die conditioned on the significant predictors. For the first time, we showed that the average time of host infection before recovery or death was between 6 and 14 days.

- The fourth research question sought to determine the parasite virulence (quantified by both host mortality and recovery) of the gyrodactylid strains time-varying and dependent on the covariates (fish sex, fish size, fish stock and parasite strain). We provided answers to this in Chapter 2 by quantifying parasite virulence of three different strains as a function of host mortality and recovery across different fish stocks and sexes based on fitted MSM. We found that a longer period of host infection leads to a higher chance of host recovery and a small chance of host mortality. Male fish from the three host populations consistently had a higher rate of host mortality than female fish stocks over time. Parasite virulence was thus significantly time-dependent and generally increased towards the end of the infection period.

- The subsequent research questions (5 to 9) wanted to i) determine if the birth rates (for young and old parasites) and death rates (with or without immune response) of *Gyrodactylus* parasites were significantly different across the three parasite strains, ii) determine whether an adaptive immune response from gyrodactylid infection progression, sex and host-dependent, iii) determine whether the mortality rate of male fish from gyrodactylid infection significantly higher than female fish, iv) ascertain whether the microhabitat preferences of *Gyrodactylus turnbulli* and *G. bullatarudis* parasite species driven by their rate of movement on their fish host, and finally v) determine the effective population carrying capacity of *Gyrodactylus* parasites at the major body regions of their fish host.

To address all research questions, we developed a more sophisticated (individual-based) stochastic simulation model. We parameterised the model to include spatial information and other relevant information about parasite fecundity, age group (young or old parasite), parasite mortality, parasite mobility, and host immune response. This also motivated the need to propose a more robust ABC methodology (defined in Chapter 5) to fit this complex stochastic model (with over 23 parameters) in Chapter 6. We also adopted an auxiliary stochastic model, which simplifies the more sophisticated simulation model (in Chapter 4) to aid in refining our modified sequential ABC samplers (by providing good theoretical justification of the auxiliary model, including its parameter estimation techniques, amongst others). After ABC fitting of the complex simulation model and further correcting the resulting posterior (based on another modified regression adjustment methodology), it was found that the effective carrying capacity at the host's major body regions was between 93 and 117 parasites (with an average value of 104). Given the adjusted posterior samples of model parameters, we used a developing Bayesian hypothesis test (whose decision rule relies on a region of practical equivalence and the highest density interval) to test underlying hypotheses. Based on a statistical test of differences between appropriate model parameters, we provided specific answers to the remaining research questions (i-iv) by testing nine sets of hypotheses in section 6.3.3.

7.2 Main biological and mathematical contributions

This interdisciplinary research work has made novel contributions to the gyrodactylid-fish system and the mathematical or ecological modelling community. First, we have offered new epidemiological insights into the gyrodactylid-fish system by analysing empirical data in Chapter 2. By adopting a multi-state Markov model for the first time in a parasitological study, we have justified the need to realistically model the host's entire infection history (before death eventually occurs) and estimate other relevant quantities concerning parasite virulence and host survival. Three open biological questions concerning parasite microhabitat preference, host survival, and parasite virulence were mathematically answered. We identified host-parasite strain-specific microhabitat preferences, discovered determinants of host survival, and quantified host-specific parasite virulence as a function of both host mortality and recovery.

Improving upon the multi-state Markov model (developed in Chapter 2) and the existing agent-based model for the gyrodactylid-fish system, we developed and calibrated our novel individual-based stochastic simulation model in Chapter 6 (with documented R scripts made publicly available on GitHub). This novel simulation model is robust enough to help simulate the infection dynamics of three different parasite strains over the external surfaces of three different host populations within a standard 17-day infection period. Furthermore, based on specified demographic information (such as parasite strain, fish type, fish sex, and fish size) and specified model parameters, the developed stochastic simulation model can provide information regarding parasite numbers at the major body locations (tail, lower region, upper region, and head) of fish over time for a given host population. In addition, the fish survival status and the exact time to fish mortality are other essential outputs of the simulation model. Hence, this proposed individual-based stochastic model can facilitate experimental data collection and help investigate specific biological questions and the system's complexity that may be difficult to control and implement experimentally.

Considering the linear birth-death process with catastrophe extinction (B-D-C process) as an auxiliary model to the novel stochastic simulation model in Chapter 4, we analytically derived, for the first time with numerical justifications and *in silico* experiments, the exact transition function and theoretical moments of the B-D-C process in the setting of discretely observed processes. We further established three efficient approaches (based on maximum likelihood estimation, generalised method of moments and the embedded Galton-Watson estimation method) to estimate the B-D-C model parameters. Also, we demonstrated an approach to simulate the B-D-C process with discrete-state space via τ-leaping, for the first time in a hybrid manner, by separately setting up the catastrophe event (where the entire population extinct within an infinitesimal time interval) different from the birth and death events using standard Monte Carlo technique. The motivation behind the hybrid setup was that, in simulating continuous-time Markov processes, the state variable could not change by more than one within an infinitesimal time interval. Findings from the B-D-C parameter estimation and its hybrid τ-leaping simulation provided additional insights on accelerating the novel stochastic simulation model (based on its proposed leap size) and aided in the computation of some components of the multidimensional summary statistics during ABC fitting.

The current study also proposed a modified sequential Monte ABC algorithm (dubbed weighted-iterative ABC) which: i) adaptively integrates importance weights for importance sampling and summary statistics weights (based on accepted simulations by computing the harmonic mean between previous and current summary statistics weights due to mismatch in their respective proposal densities via the iterative ABC procedure with importance sampling) to improve ABC posterior approximations, ii) adopt a weighted distance metric for estimating discrepancy between high-dimensional simulated and observed summary statistics (in the case where the summary statistics have a bi-dimensional space), iii) employ a computationally efficient multivariate normal perturbation kernel (with bandwidth matrix optimally determined), and iv) then separately adjust the resulting ABC posterior using a modified heterogeneous local-linear regression with $L2$ regularisation, robust enough to deal with high multicollinearity and supercollinearity in the neighbourhood of the observed summaries. In the instance of high multicollinearity during ABC posterior adjustment, the standard local-linear regression based on weighted least-squares is impossible to implement due to matrix singularity issues. However, the proposed posterior correction in the current study is implementable in the presence of high multicollinearity by shrinking regression coefficients of predictors with less importance. Additionally, the proposed ABC post-processing method can effectively deal with supercollinearity (where the predictors outnumber the posterior samples) as opposed to the standard local-linear regression developed by Beaumont et al. [27] which may result in poor estimates (as revealed in this study). The ABC methodologies proposed in this study can aid in the parameter estimation of either complex or simple likelihood-free models sequentially across a whole population.

7.3 Future research directions

The proposed mathematical and stochastic simulation models can be extended and adapted for different host-parasite systems and other ecological systems. Furthermore, the modified population-based ABC posterior estimation methodologies can be employed to calibrate other multi-parameter models with many correlating or independent high-dimensional summary statistics. Specifically, the following are future works concerning the novel stochastic simulation model, the modified sequential Monte Carlo ABC with adaptive importance sampling and the gyrodactylid-fish system:

- Within the complex stochastic simulation model (developed in Chapter 6), we assumed that the rate of localised host immune response (which occurs temporally as a function of the number of parasites at any of the host body locations) also depends on fish sex (with two levels) and type of fish (with three levels). Thus, the current study only considered the additive impact of the covariates (fish sex and fish type) on the immune response rate without considering interaction effects.

In contrast, future studies should also consider the multiplicative or interaction effects of these covariates on the rate of localised immune response and compare the modified model with the current version with additive immune response rates.

- In the modified sequential Monte Carlo ABC algorithm with importance sampling, we pre-specified or fixed the set of decreasing tolerance thresholds and the final ABC stopping time (with ten ABC time steps and tolerances). Future studies can propose an adaptive or automated stopping rule so that the ABC algorithm terminates after posterior convergence to improve its computational speed by minimising the number of sequential ABC iterations. In addition, the set of decreasing tolerance thresholds can be determined adaptively based on either some quantiles of the distances of accepted proposal samples from the previous time step or some quantiles of the effective sample size values. The impact of different optimal perturbation kernels, such as component-wise perturbation kernels, other multivariate perturbation kernels (such as multivariate uniform kernels) and local perturbation kernels with or without nearest neighbours can also be investigated further for the modified ABC algorithm to identify other good choices of the perturbation kernel.

- Moreover, the novel stochastic simulation model's simulation time axis (or preferably the observed time points) can be extended further to make predictions beyond the standard 17-day infection period. Other researchers are yet to study this exciting and crucial biological modelling problem concerning the gyrodactylid-fish system across an entire host population. The extended model can help discover what happens to infected fish beyond the standard 17-days by assessing how infections are maintained in the long term among different host populations. The extended model should also be fitted using the proposed ABC methodologies with the help of observed experimental data.

- Based on our proposed simulation model, future studies can further conduct biological experiments that are challenging to explore experimentally because of similarities between gyrodactylids and other unfavourable conditions. This can be done by modifying our stochastic simulation model to investigate mixed gyrodactylid parasite populations or co-infections on a single or population of fish based on what is already known about these gyrodactylid species (i.e., *G. turnbulli* and *G. bullatarudis* parasites). In addition, relevant ecological questions can be explored regarding how the two (very different) *Gyrodactylus* species interact or compete and which one temporally wins at an individual host and population levels.

- Finally, future studies can use a social network model coupled with the proposed stochastic simulation model to describe the infection dynamics of a fish population and their interactions. The social network model should capture the parasite load for each fish over time, but must not necessarily give the exact spatial locations of parasites on an individual host. The model must then be calibrated using approximate Bayesian computation (ABC).

Bibliography

1. Abbott, R. D. (1985). Logistic regression in survival analysis. *American journal of epidemiology*, 121(3):465–471.

2. Aeschbacher, S., Beaumont, M. A., and Futschik, A. (2012). A novel approach for choosing summary statistics in approximate Bayesian computation. *Genetics*, 192(3):1027–1047.

3. Allen, L. J. (2015). Stochastic population and epidemic models. *Mathematical biosciences lecture series, stochastics in biological systems.*

4. Allen, L. J. (2017). A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. *Infectious Disease Modelling*, 2(2):128–142.

5. Andersen, P. K. and Borgan, Ø. (1984). Counting Process Models for Life History Data: A Review. *Preprint series. Statistical Research Report http://urn. nb. no/URN: NBN: no-23420.*

6. Anderson, R. M. and May, R. M. (1978). Regulation and stability of host-parasite population interactions: I. Regulatory processes. *The journal of animal ecology*, pages 219–247.

7. Anderson, R. M. and May, R. M. (1981). The population dynamics of microparasites and their invertebrate hosts. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 291(1054):451–524.

8. Anderson, R. M. and May, R. M. (1992). *Infectious diseases of humans: dynamics and control.* Oxford university press.

9. Andersson, H. and Britton, T. (2012). *Stochastic epidemic models and their statistical analysis*, volume 151. Springer Science & Business Media.

10. Arkin, R. G. and Montgomery, D. C. (1980). Augmented robust estimators. *Technometrics*, 22(3):333–341.

11. Aryal, N. R. and Jones, O. D. (2020). Fitting the Bartlett–Lewis rainfall model using Approximate Bayesian Computation. *Mathematics and Computers in Simulation*, 175:153–163.

12. Asmussen, S. and Glynn, P. W. (2007). *Stochastic simulation: algorithms and analysis*, volume 57. Springer Science & Business Media.

13. Athreya, K. B. (1970). Branching process. *The Annals of Mathematical Statistics*, 41(1):195–202.

14. Bailey, N. T. et al. (1975). *The mathematical theory of infectious diseases and its applications*. Number 2nd ediition. Charles Griffin & Company Ltd 5a Crendon Street, High Wycombe, Bucks HP13 6LE.

15. Bakke, T. A., Cable, J., and Harris, P. D. (2007). The Biology of Gyrodactylid Monogeneans: The “Russian-Doll Killers". *Advances in Parasitology*, 64(0318).

16. Bakke, T. A., Harris, P. D., and Cable, J. (2002). Host specificity dynamics: observations on gyrodactylid monogeneans. *International Journal for Parasitology*, 32(3):281–308.

17. Bakke, T. A., Nilsen, K. B., and Shinn, A. P. (2004). Chaetotaxy applied to Norwegian *Gyrodactylus salaris* Malmberg, 1957 (Monogenea) clades and related species from salmonids. *Folia Parasitologica*, 51(2-3):253–261.

18. Ball, F. (1985). Deterministic and stochastic epidemics with several kinds of susceptibles. *Advances in Applied Probability*, 17(1):1–22.

19. Ball, F., Britton, T., and Sirl, D. (2013). A network with tunable clustering, degree correlation and degree distribution, and an epidemic thereon. *Journal of Mathematical Biology*, 66(4):979–1019.

20. Banisch, S. (2015). *Markov chain aggregation for agent-based models*. Springer.

21. Banks, H. T., Broido, A., Canter, B., Gayvert, K., Hu, S., Joyner, M., and Link, K. (2012). Simulation algorithms for continuous time Markov chain models. *Studies in Applied Electromagnetics and Mechanics*, 37:3–18.

22. Barber, S., Voss, J., and Webster, M. (2015). The rate of convergence for approximate Bayesian computation. *Electronic Journal of Statistics*, 9(1):80–105.

23. Barnes, C. P., Filippi, S., Stumpf, M. P., and Thorne, T. (2012). Considerate approaches to constructing summary statistics for ABC model selection. *Statistics and Computing*, 22(6):1181–1197.

24. Barthelmé, S., Chopin, N., and Cottet, V. (2018). Divide and conquer in ABC: Expectation-propagation algorithms for likelihood-free inference. In *Handbook of Approximate Bayesian Computation*, pages 415–434. Chapman and Hall/CRC.

25. Bartlett, M. (1949). Some evolutionary stochastic processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(2):211–229.

26. Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., and Robert, C. P. (2009). Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990.

27. Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

28. Becker, N. (1979). The uses of epidemic models. *Biometrics*, pages 295–305.

29. Begon, M., Bennett, M., Bowers, R. G., French, N. P., Hazel, S., and Turner, J. (2002). A clarification of transmission terms in host-microparasite models: numbers, densities and areas. *Epidemiology & Infection*, 129(1):147–153.

30. Berec, L. (2002). Techniques of spatially explicit individual-based models: Construction, simulation, and mean-field analysis. *Ecological Modelling*, 150(1-2):55–81.

31. Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019). Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(2):235–269.

32. Berry, S. D., Ngo, L., Samelson, E. J., and Kiel, D. P. (2010). Competing Risk of Death: An Important Consideration in Studies of Older Adults. *Journal of the American Geriatrics Society*, 58(4):783–787.

33. Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98.

34. Biau, G., Cérou, F., and Guyader, A. (2015). New insights into approximate Bayesian computation. In *Annales de l'IHP Probabilités et statistiques*, volume 51, pages 376–403.

35. Biggins, J. (1995). The growth and spread of the general branching random walk. *The Annals of Applied Probability*, pages 1008–1024.

36. Black, A. J. and McKane, A. J. (2012). Stochastic formulation of ecological models and their applications. *Trends in ecology & evolution*, 27(6):337–345.

37. Blum, M. G. (2010a). Approximate Bayesian computation: a nonparametric perspective. *Journal of the American Statistical Association*, 105(491):1178–1187.

38. Blum, M. G. (2010b). Choosing the summary statistics and the acceptance rate in approximate Bayesian computation. In *Proceedings of COMPSTAT'2010*, pages 47–56. Springer.

39. Blum, M. G. (2018). Regression approaches for ABC. In *Handbook of Approximate Bayesian Computation*, pages 71–85. Chapman and Hall/CRC.

40. Blum, M. G. and François, O. (2010). Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing*, 20(1):63–73.

41. Boeger, W. A., Kritsky, D. C., Pie, M. R., and Engers, K. B. (2005). Mode of transmission, host switching, and escape from the Red Queen by viviparous gyrodactylids (Monogenoidea). *Journal of Parasitology*, 91(5):1000–1008.

42. Bohner, M., Streipert, S., and Torres, D. F. (2019). Exact solution to a dynamic SIR model. *Nonlinear Analysis: Hybrid Systems*, 32:228–238.

43. Bolker, B. M. (2000). Moment methods for ecological processes in continuous space. *The geometry of ecological interactions*, pages 388–411.

44. Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the national academy of sciences*, 99(suppl 3):7280–7287.

45. Bonassi, F. V. and West, M. (2015). Sequential Monte Carlo with adaptive weights for approximate Bayesian computation. *Bayesian Analysis*, 10(1):171–187.

46. Bonsall, M. B. (2009). Population models. *Ecology-Volume II*, page 235.

47. Bortot, P., Coles, S. G., and Sisson, S. A. (2007). Inference for stereological extremes. *Journal of the American Statistical Association*, 102(477):84–92.

48. Brauer, F. (2008). Compartmental models in epidemiology. In *Mathematical epidemiology*, pages 19–79. Springer.

49. Britton, T. (2020). Epidemic models on social networks—With inference. *Statistica Neerlandica*, 74(3):222–241.

50. Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC press.

51. Burr, T. and Skurikhin, A. (2013). Selecting summary statistics in approximate bayesian computation for calibrating stochastic models. *BioMed research international*, 2013.

52. Cable, J., Archard, G. A., Mohammed, R. S., McMullan, M., Stephenson, J. F., Hansen, H., and van Oosterhout, C. (2013). Can parasites use predators to spread between primary hosts? *Parasitology*, 140(9):1138–1143.

53. Cable, J. and Harris, P. (2002). Gyrodactylid developmental biology: historical review, current status and future trends. *International Journal for Parasitology*, 32(3):255–280.

54. Cable, J., Harris, P., and Bakke, T. A. (2000). Population growth of *Gyrodactylus salaris* (Monogenea) on Norwegian and Baltic Atlantic salmon (Salmo salar) stocks. *Parasitology*, 121(6):621–629.

55. Cable, J., Scott, E. C. G., Tinsley, R., and Harris, P. (2002). Behavior favoring transmission in the viviparous monogenean *Gyrodactylus turnbulli*. *Journal of Parasitology*, 88(1):183–184.

56. Cable, J. and van Oosterhout, C. (2007a). The impact of parasites on the life history evolution of guppies (Poecilia reticulata): The effects of host size on parasite virulence. *International Journal for Parasitology*, 37(13):1449–1458.

57. Cable, J. and van Oosterhout, C. (2007b). The role of innate and acquired resistance in two natural populations of guppies (*Poecilia reticulata*) infected with the ectoparasite *Gyrodactylus turnbulli*. *Biological Journal of the Linnean Society*, 90(4):647–655.

58. Caflisch, R. E. (1998). Monte carlo and quasi-monte carlo methods. *Acta numerica*, 7:1–49.

59. Carrington, P. J., Scott, J., and Wasserman, S. (2005). *Models and methods in social network analysis*, volume 28. Cambridge university press.

60. Caswell, H. (2000). *Matrix population models*, volume 1. Sinauer Sunderland, MA, USA.

61. Chakladar, S., Liao, R., Landau, W., Gamalo, M., and Wang, Y. (2022). Discrete Time Multistate Model With Regime Switching for Modeling COVID-19 Disease Progression and Clinical Outcomes. *Statistics in Biopharmaceutical Research*, 14(1):52–66.

62. Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2012). *Monte Carlo methods in Bayesian computation*. Springer Science & Business Media.

63. Chiang, C. L. (1961). On the probability of death from specific causes in the presence of competing risks. In *Proceedings of the fourth Berkeley symposium on mathematical*

statistics and probability, volume 4, pages 169–180. University of California Press, Berkeley, CA.

64. Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, 89(3):539–552.

65. Chowell, G., Sattenspiel, L., Bansal, S., and Viboud, C. (2016). Mathematical models to characterize early epidemic growth: A review. *Physics of life reviews*, 18:66–97.

66. Cisewski-Kehe, J., Weller, G., and Schafer, C. (2019). A preferential attachment model for the stellar initial mass function. *Electronic Journal of Statistics*, 13(1):1580–1607.

67. Clinical, I., Advisor, V., and Information, B. (2018). Monogenea Learn more about Monogenea Flukes (Monogenean Parasites).

68. Collette, B. B. (2020). The Future of Bluefin Tunas: Ecology, Fisheries Management, and Conservation. *Reviews in Fisheries Science & Aquaculture*, 28(1):136–137.

69. Corander, J., Fraser, C., Gutmann, M. U., Arnold, B., Hanage, W. P., Bentley, S. D., Lipsitch, M., and Croucher, N. J. (2017). Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. *Nature ecology & evolution*, 1(12):1950–1960.

70. Cox, D. R. (2006). *Principles of statistical inference.* Cambridge university press.

71. Cox, D. R. and Miller, H. D. (2017). *The theory of stochastic processes.* Routledge.

72. Croft, D. P., Edenbrow, M., Darden, S. K., Ramnarine, I. W., van Oosterhout, C., and Cable, J. (2011). Effect of gyrodactylid ectoparasites on host behaviour and social network structure in guppies *Poecilia reticulata*. *Behavioral Ecology and Sociobiology*, 65(12):2219–2227.

73. Csilléry, K., François, O., and Blum, M. G. (2012). abc: an R package for approximate Bayesian computation (ABC). *Methods in ecology and evolution*, 3(3):475–479.

74. Daley, D. J. and Gani, J. (1999). Epidemic Modeling: An Introduction. *Cambridge Studies in Mathematical Biology*, 15.

75. Davison, A. C., Hautphenne, S., and Kraus, A. (2021). Parameter estimation for discretely observed linear birth-and-death processes. *Biometrics*, 77(1):186–196.

76. De Roos, A. M., Mccauley, E., and Wilson, W. G. (1991). Mobility versus density-limited predator-prey dynamics on different spatial scales. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 246(1316):117–122.

77. DeAngelis, D. L. (2018). *Individual-based models and approaches in ecology: populations, communities and ecosystems.* CRC Press.

78. DeAngelis, D. L. and Grimm, V. (2014). Individual-based models in ecology after four decades. *F1000prime reports*, 6.

79. Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436.

80. Del Moral, P., Doucet, A., and Jasra, A. (2012). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and computing*, 22(5):1009–1020.

81. Denholm, S. J. (2013). Mathematical models for investigating the long-term impact of *Gyrodactylus salaris* infections on Atlantic salmon populations.

82. des Clers, S. (1993). Modelling the impact of disease-induced mortality on the population size of wild salmonids. *Fisheries Research*, 17(1-2):237–248.

83. Di Crescenzo, A., Giorno, V., Nobile, A. G., and Ricciardi, L. M. (2008). A note on birth–death processes with catastrophes. *Statistics & probability letters*, 78(14):2248–2257.

84. Didelot, X., Everitt, R. G., Johansen, A. M., and Lawson, D. J. (2011). Likelihood-free estimation of model evidence. *Bayesian analysis*, 6(1):49–76.

85. Diekmann, O. and Heesterbeek, J. A. P. (2000). *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*, volume 5. John Wiley & Sons.

86. Diekmann, O., Heesterbeek, J. A. P., and Metz, J. A. (1990). On the definition and the computation of the basic reproduction ratio r_0 in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology*, 28(4):365–382.

87. Divya, B. and Kavitha, K. (2020). A Review on Mathematical Modelling in Biology and Medicine. *Advances in Mathematics: Scientific Journal*, 9:5869–5879.

88. Dobson, A. P. and Hudson, P. J. (1992). Regulation and stability of a free-living host-parasite system: Trichostrongylus tenuis in red grouse. II. Population models. *Journal of Animal Ecology*, pages 487–498.

89. Douc, R. and Cappé, O. (2005). Comparison of resampling schemes for particle filtering. In *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005.*, pages 64–69. IEEE.

90. Durrett, R. and Levin, S. (1994a). The importance of being discrete (and spatial). *Theoretical population biology*, 46(3):363–394.

91. Durrett, R. and Levin, S. A. (1994b). Stochastic spatial models: a user's guide to ecological applications. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 343(1305):329–350.

92. Eames, K. T. and Keeling, M. J. (2002). Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases. *Proceedings of the national academy of sciences*, 99(20):13330–13335.

93. Efron, B. (1988). Logistic regression, survival analysis, and the Kaplan-Meier curve. *Journal of the American statistical Association*, 83(402):414–425.

94. El-Naggar, M., El-Naggar, A., and Kearn, G. (2004). Swimming in *Gyrodactylus*

rysavyi [Monogenea, Gyrodactylidae] from the Nile catfish, *Clarias gariepinus*. *Acta Parasitologica*, 49(2).

95. Elandt-Johnson, R. C. (1976). Conditional failure time distributions under competing risk theory with dependent failure times and proportional hazard rates. *Scandinavian Actuarial Journal*, 1976(1):37–51.

96. Ezanno, P., Vergu, E., Langlais, M., and Gilot-Fromont, E. (2012). Modelling the dynamics of host-parasite interactions: basic principles. In *New Frontiers of Molecular Epidemiology of Infectious Diseases*, pages 79–101. Springer.

97. Fan, Y., Nott, D. J., and Sisson, S. A. (2013). Approximate Bayesian computation via regression density estimation. *Stat*, 2(1):34–48.

98. Farley, J. (1992). Parasites and the germ theory of disease. *Hospital Practice*, 27(9):175–196.

99. Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474.

100. Feller, W. (1940). On the integro-differential equations of purely discontinuous Markoff processes. *Transactions of the American Mathematical Society*, 48:488–515.

101. Ferrari, M. J., Perkins, S. E., Pomeroy, L. W., and Bjørnstad, O. N. (2011). Pathogens, social networks, and the paradox of transmission scaling. *Interdisciplinary perspectives on infectious diseases*, 2011.

102. Filippi, S., Barnes, C. P., Cornebise, J., and Stumpf, M. P. (2013). On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. *Statistical applications in genetics and molecular biology*, 12(1):87–107.

103. Forsberg White, L. and Pagano, M. (2008). A likelihood-based method for real-time estimation of the serial interval and reproductive number of an epidemic. *Statistics in medicine*, 27(16):2999–3016.

104. Frazier, D. T., Martin, G. M., Robert, C. P., and Rousseau, J. (2018). Asymptotic properties of approximate Bayesian computation. *Biometrika*, 105(3):593–607.

105. Frazier, D. T., Robert, C. P., and Rousseau, J. (2020). Model misspecification in approximate Bayesian computation: consequences and diagnostics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):421–444.

106. Fromm, B. and Museum, N. H. (2014). Occurrence and phylogenetic implications of microRNAs in the fish parasite *Gyrodactylus salaris* (Platyhelminthes : Neodermata : Monogenea) and related species.

107. Gaba, S., Cabaret, J., Ginot, V., and Silvestre, A. (2006). The early drug selection of nematodes to anthelmintics: stochastic transmission and population in refuge. *Parasitology*, 133(3):345–356.

108. Gallagher, S. (2017). Comparing compartment and agent-based models. In *Joint Statistical Meeting, Baltimore*.

109. Gani, J. (1980). Mathematical models of epidemics. *The Mathematical Intelligencer*, 3(1):41–43.

110. Gillespie, D. T. (2001). Approximate accelerated stochastic simulation of chemically reacting systems. *Journal of Chemical Physics*, 115(4):1716–1733.

111. Gillespie, D. T. and Petzold, L. R. (2003). Improved lead-size selection for accelerated stochastic simulation. *Journal of Chemical Physics*, 119(16):8229–8234.

112. Gordon, N. J., Salmond, D. J., and Smith, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proceedings F-radar and signal processing*, volume 140, pages 107–113. IET.

113. Gourbière, S., Morand, S., and Waxman, D. (2015). Fundamental factors determining the nature of parasite aggregation in hosts. *PloS one*, 10(2):e0116893.

114. Gradmann, C. (2006). Robert Koch and the white death: from tuberculosis to tuberculin. *Microbes and infection*, 8(1):294–301.

115. Grazian, C. and Fan, Y. (2020). A review of approximate Bayesian computation methods via density estimation: Inference for simulator-models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12(4):e1486.

116. Greenwood, M. (1931). On the statistical measure of infectiousness. *Epidemiology & Infection*, 31(3):336–351.

117. Grenfell, B., Wilson, K., Isham, V., Boyd, H., and Dietz, K. (1995). Modelling patterns of parasite aggregation in natural populations: trichostrongylid nematode–ruminant interactions as a case study. *Parasitology*, 111(S1):S135–S151.

118. Griffiths, S. W. and Magurran, A. E. (1998). Sex and schooling behaviour in the Trinidadian guppy. *Animal Behaviour*, 56(3):689–693.

119. Grimm, V. and Railsback, S. F. (2013). *Individual-based modeling and ecology.* Princeton university press.

120. Guidoum, A. C. (2020). Kernel Estimator and Bandwidth Selection for Density and its Derivatives: The kedd Package. *arXiv preprint arXiv:2012.06102.*

121. Haccou, P., Jagers, P., and Vatutin, V. A. (2005). *Branching processes: variation, growth, and extinction of populations.* Number 5. Cambridge university press.

122. Hagen, O., Hartmann, K., Steel, M., and Stadler, T. (2015). Age-dependent speciation can explain the shape of empirical phylogenies. *Systematic biology*, 64(3):432–440.

123. Hamer, W. H. (1906). *Epidemic disease in England: the evidence of variability and of persistency of type.* Bedford Press.

124. Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054.

125. Haris, T. E. (2002). *The theory of Branching Processes.* Dover, Mineola, NY.

126. Harris, P. (1983). The morphology and life-cycle of the oviparous *Oögyrodactylus farlowellae* gen. et sp. nov. (Monogenea, Gyrodactylidea). *Parasitology*, 87(3):405–420.

127. Harris, P. (1989). Interactions between population growth and sexual reproduction in the viviparous monogenean *Gyrodactylus turnbulli* Harris, 1986 from the guppy, *Poecilia reticulata* Peters. *Parasitology*, 98(2):245–251.

128. Harris, P., Jansen, P., and Bakke, T. (1994). The population age structure and reproductive biology of *Gyrodactylus salaris* Malmberg (Monogenea). *Parasitology*, 108(2):167–173.

129. Harris, P. D. (1988). Changes in the site specificity of *Gyrodactylus turnbulli* Harris, 1986 (Monogenea) during infections of individual guppies (*Poecilia reticulata* Peters, 1859). *Canadian Journal of Zoology*, 66(12):2854–2857.

130. Harris, P. D. (1993). Interactions between reproduction and population biology in gyrodactylid monogeneans-A review. *Bulletin Français de la Pêche et de la Pisciculture*, 1:47–65.

131. Harris, P. D. and Lyles, A. M. (1992a). Infections of *Gyrodactylus bullatarudis* and *Gyrodactylus turnbulli* on Guppies (*Poecilia reticulata*) in Trinidad. *The Journal of Parasitology*, 78(5):912.

132. Harris, P. D. and Lyles, A. M. (1992b). Infections of *Gyrodactylus bullatarudis* and *Gyrodactylus turnbull*i on guppies (*Poecilia reticulata*) in Trinidad. *The Journal of parasitology*, pages 912–914.

133. Harris, P. D., Shinn, A., Cable, J., and Bakke, T. A. (2004). Nominal species of the genus *Gyrodactylus* von Nordmann 1832 (Monogenea: Gyrodactylidae), with a list of principal host species. *Systematic Parasitology*, 59(1):1–27.

134. Harris, P. D., Shinn, A. P., Cable, J., Bakke, T. A., and Bron, J. (2008). GyroDb: gyrodactylid monogeneans on the web. *Trends in Parasitology*, 24(3):109–111.

135. Harris, T. E. (1950). Some mathematical models for branching processes. Technical report, RAND CORP SANTA MONICA CA.

136. Harrison, J. U. and Baker, R. E. (2020). An automatic adaptive method to combine summary statistics in approximate Bayesian computation. *PloS one*, 15(8):e0236954.

137. Harvell, C. D., Mitchell, C. E., Ward, J. R., Altizer, S., Dobson, A. P., Ostfeld, R. S., and Samuel, M. D. (2002). Climate warming and disease risks for terrestrial and marine biota. *Science*, 296(5576):2158–2162.

138. Harvey, J. A., Malcicka, M., and Ellers, J. (2015). Integrating more biological and ecological realism into studies of multitrophic interactions. *Ecological Entomology*, 40(4):349–352.

139. Hastie, T. and Qian, J. (2014). Glmnet vignette. *Retrieved June*, 9(2016):1–30.

140. Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

141. Hatcher, M. J., Dick, J. T., and Dunn, A. M. (2006). How parasites affect interactions between competitors and predators. *Ecology Letters*, 9(11):1253–1271.

142. Hautphenne, S., Krings, G., Delvenne, J.-C., and Blondel, V. D. (2015). Sensitivity analysis of a branching process evolving on a network with application in epidemiology. *Journal of Complex Networks*, 3(4):606–641.

143. He, F., Mazumdar, S., Tang, G., Bhatia, T., Anderson, S. J., Dew, M. A., Krafty, R., Nimgaonkar, V., Deshpande, S., Hall, M., et al. (2017). Non-parametric MANOVA approaches for non-normal multivariate outcomes with missing values. *Communications in Statistics-Theory and Methods*, 46(14):7188–7200.

144. Heesterbeek, H. (2005). The law of mass-action in epidemiology: a historical perspective. *Ecological paradigms lost: routes of theory change*, pages 81–104.

145. Heesterbeek, J. A. P. (2002). A brief history of R_0 and a recipe for its calculation. *Acta biotheoretica*, 50(3):189–204.

146. Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM review*, 42(4):599–653.

147. Hoem, J. M., Keiding, N., Kulokari, H., Natvig, B., Barndorff-Nielsen, O., and Hilden, J. (1976). The statistical theory of demographic rates: A review of current developments [with discussion and reply]. *Scandinavian Journal of Statistics*, pages 169–185.

148. Høgåsen, H. and Brun, E. (2003). Risk of inter-river transmission of Gyrodactylus salaris by migrating Atlantic salmon smolts, estimated by Monte Carlo simulation. *Diseases of Aquatic Organisms*, 57(3):247–254.

149. Holland, P. W. (1973). Weighted ridge regression: Combining ridge and robust regression methods. *NBER Working Paper*, (w0011).

150. Holt, R. D. and Polis, G. A. (1997). A theoretical framework for intraguild predation. *The American Naturalist*, 149(4):745–764.

151. Hougaard, P. (1999). Multi-state models: a review. *Lifetime data analysis*, 5(3):239–264.

152. Hunter, D. R., Krivitsky, P. N., and Schweinberger, M. (2012). Computational statistical methods for social network models. *Journal of Computational and Graphical Statistics*, 21(4):856–882.

153. Huston, M., DeAngelis, D., and Post, W. (1988). New computer models unify ecological theory: computer simulations show that many ecological patterns can be explained by interactions among individual organisms. *BioScience*, 38(10):682–691.

154. Huyse, T., Audenaert, V., and Volckaert, F. A. (2003). Speciation and host-parasite relationships in the parasite genus *Gyrodactylus* (Monogenea, Platyhelminthes) infecting gobies of the genus *Pomatoschistus* (Gobiidae, Teleostei). *International journal for parasitology*, 33(14):1679–1689.

155. Inés, M., González, M., Gutiérrez, C., Martínez, R., Minuesa, C., Molina, M., Mota, M., and Ramos, A. (2016). *Branching Processes and Their Applications*, volume 219. Springer.

156. Ishida, E. E., Vitenti, S. D., Penna-Lima, M., Cisewski, J., de Souza, R. S., Trindade, A. M., Cameron, E., Busti, V. C., collaboration, C., et al. (2015). Cosmoabc: likelihood-free inference via population Monte Carlo approximate Bayesian computation. *Astronomy and Computing*, 13:1–11.

157. Izquierdo, L. R., Izquierdo, S. S., Galan, J. M., and Santos, J. I. (2009). Techniques to understand computer simulations: Markov chain analysis. *Journal of Artificial Societies and Social Simulation*, 12(1):6.

158. Jackson, C. H. (2011). Multi-state models for panel data: the msm package for R. *Journal of statistical software*, 38(8):1–29.

159. Jackson, C. H., Sharples, L. D., Thompson, S. G., Duffy, S. W., and Couto, E. (2003). Multistate Markov models for disease progression with classification error. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(2):193–209.

160. Jacob, C. (2010). Branching processes: their role in epidemiology. *International journal of environmental research and public health*, 7(3):1186–1204.

161. Jansen, P. A., Matthews, L., and Toft, N. (2007). Geographic risk factors for inter-river dispersal of Gyrodactylus salaris in fjord systems in Norway. *Diseases of aquatic organisms*, 74(2):139–149.

162. Janson, S. (2004). Functional limit theorems for multitype branching processes and generalized Pólya urns. *Stochastic Processes and their Applications*, 110(2):177–245.

163. Jiang, B. (2018). Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy. In *International conference on artificial intelligence and statistics*, pages 1711–1721. PMLR.

164. Jost, J., Kell, M., and Rodrigues, C. S. (2015). Representation of Markov chains by random maps: existence and regularity conditions. *Calculus of Variations and Partial Differential Equations*, 54(3):2637–2655.

165. Joyce, P. and Marjoram, P. (2008). Approximately sufficient statistics and Bayesian computation. *Statistical applications in genetics and molecular biology*, 7(1).

166. Jung, H. and Marjoram, P. (2011). Choice of summary statistic weights in approximate Bayesian computation. *Statistical applications in genetics and molecular biology*, 10(1).

167. Karlin, S. and Tavaré, S. (1982). Linear birth and death processes with killing. *Journal of Applied Probability*, 19(3):477–487.

168. Keeling, M. and Danon, L. (2009). Mathematical modelling of infectious diseases. *British medical bulletin*, 92(1).

169. Keeling, M. J. and Eames, K. T. (2005). Networks and epidemic models. *Journal of the royal society interface*, 2(4):295–307.

170. Keeling, M. J. and Rohani, P. (2011). *Modeling infectious diseases in humans and animals*. Princeton university press.

171. Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721.

172. Kermack, W. O. and McKendrick, A. G. (1991). Contributions to the mathematical theory of epidemics–I. 1927. *Bulletin of mathematical biology*, 53(1-2):33–55.

173. Kesten, H., Ney, P., and Spitzer, F. (1966). The Galton-Watson process with mean one and finite variance. *Theory of Probability & Its Applications*, 11(4):513–540.

174. Keymer, A. E. and Anderson, R. (1979). The dynamics of infection of Tribolium confusum by Hymenolepis diminuta: the influence of infective-stage density and spatial distribution. *Parasitology*, 79(2):195–207.

175. Khazeiynasab, S. R. and Qi, J. (2021). Generator parameter calibration by adaptive approximate bayesian computation with sequential monte carlo sampler. *IEEE Transactions on Smart Grid*, 12(5):4327–4338.

176. King, T. and Cable, J. (2007). Experimental infections of the monogenean *Gyrodactylus turnbulli* indicate that it is not a strict specialist. *International Journal for Parasitology*, 37(6):663–672.

177. Kobayashi, G. and Kozumi, H. (2015). Generalized multiple-point Metropolis algorithms for approximate Bayesian computation. *Journal of Statistical Computation and Simulation*, 85(4):675–692.

178. Kopp, M. and Gabriel, W. (2006). The dynamic effects of an inducible defense in the Nicholson–Bailey model. *Theoretical Population Biology*, 70(1):43–55.

179. Kritsky, D. C., Vianna, R. T., and Boeger, W. A. (2007). Neotropical Monogenoidea. 50. Oviparous gyrodactylids from loricariid and pimelodid catfishes in Brazil, with the proposal of Phanerothecioides ng, Onychogyrodactylus ng and Aglaiogyrodactylus ng (Polyonchoinea: Gyrodactylidea). *Systematic parasitology*, 66(1):1–34.

180. Kruschke, J. (2014). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.

181. Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, 6(3):299–312.

182. Kruschke, J. K. and Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic bulletin & review*, 25(1):178–206.

183. Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological methods & research*, 33(2):188–229.

184. Kurtz, T. G. (1971). Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *Journal of Applied Probability*, 8(2):344–356.

185. Lakhani, S. (1993). Early clinical pathologists: Robert Koch (1843-1910). *Journal of clinical pathology*, 46(7):596.

186. Lee, D. K. (2016). Alternatives to P value: confidence interval and effect size. *Korean journal of anesthesiology*, 69(6):555.

187. Leung, T. L. and Bates, A. E. (2013). More rapid and severe disease outbreaks for aquaculture at the tropics: Implications for food security. *Journal of Applied Ecology*, 50(1):215–222.

188. Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times*, volume 107. American Mathematical Soc.

189. Li, J., Nott, D. J., Fan, Y., and Sisson, S. A. (2017a). Extending approximate Bayesian computation methods to high dimensions via a Gaussian copula model. *Computational Statistics & Data Analysis*, 106:77–89.

190. Li, W. and Fearnhead, P. (2018a). Convergence of regression-adjusted approximate Bayesian computation. *Biometrika*, 105(2):301–318.

191. Li, W. and Fearnhead, P. (2018b). On the asymptotic efficiency of approximate Bayesian computation estimators. *Biometrika*, 105(2):285–299.

192. Li, Y., Cui, L., and Lin, C. (2017b). Modeling and analysis for multi-state systems with discrete-time Markov regime-switching. *Reliability Engineering & System Safety*, 166:41–49.

193. Lipková, J., Arampatzis, G., Chatelain, P., Menze, B., and Koumoutsakos, P. (2019). S-leaping: An adaptive, accelerated stochastic simulation algorithm, bridging tau-leaping and r-leaping. *Bulletin of mathematical biology*, 81(8):3074–3096.

194. Liu, J. S., Chen, R., and Wong, W. H. (1998). Rejection control and sequential importance sampling. *Journal of the American Statistical Association*, 93(443):1022–1031.

195. Liu, J. S. and Liu, J. S. (2001). *Monte Carlo strategies in scientific computing*, volume 10. Springer.

196. Llewellyn, J. (1984). The biology of *Isancistrum subulatae* n. sp., a monogenean parasitic on the squid, *Alloteuthis subulata*, at Plymouth. *Journal of the Marine Biological Association of the United Kingdom*, 64(2):285–302.

197. Lloyd, A. (2007). Introduction to epidemiological modeling: basic models and their properties. *Networks*, pages 1–166.

198. Louie, K., Vlassoff, A., and Mackay, A. (2005). Nematode parasites of sheep: extension of a simple model to include host variability. *Parasitology*, 130(4):437–446.

199. Louie, K., Vlassoff, A., and Mackay, A. (2007). Gastrointestinal nematode parasites of sheep: a dynamic model for their effect on liveweight gain. *International journal for Parasitology*, 37(2):233–241.

200. Lu, D. B., Rudge, J. W., Wang, T. P., Donnelly, C. A., Fang, G. R., and Webster, J. P. (2010). Transmission of *Schistosoma japonicum* in Marshland and hilly regions of China: Parasite population genetic and sibship structure. *PLoS Neglected Tropical Diseases*, 4(8).

201. Lutscher, F. (2019). *Integrodifference equations in spatial ecology.* Springer.

202. Luttbeg, B. and Schmitz, O. J. (2000). Predator and prey models with flexible individual behavior and imperfect information. *The American Naturalist*, 155(5):669–683.

203. Magurran, A. E. and Seghers, B. H. (1990). Population differences in predator recognition and attack cone avoidance in the guppy Poecilia reticulata. *Animal Behaviour*, 40(3):443–452.

204. Makowski, D., Ben-Shachar, M. S., and Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software*, 4(40):1541.

205. Malmberg, G. et al. (1970). The excretory systems and the marginal hooks as a

basis for the systematics of *Gyrodactylus* (Trematoda, Monogenea). *Arkiv for Zoologi*, 23(1/2):1–235.

206. Mandal, F. B. (2011). Does virulence offer benefit to the parasite ? *WebmedCentral Parasitology*, 2(10):1–9.

207. Manzini, G., Ettrich, T. J., Kremer, M., Kornmann, M., Henne-Bruns, D., Eikema, D. A., Schlattmann, P., and de Wreede, L. C. (2018). Advantages of a multi-state approach in surgical research: how intermediate events and risk factor profile affect the prognosis of a patient with locally advanced rectal cancer. *BMC Medical Research Methodology*, 18(1):23.

208. Marchuk, G., Asachenkov, A., Belykh, L., and Zuev, S. (1986). Mathematical modelling of infectious diseases. In *Immunology and Epidemiology*, pages 64–81. Springer.

209. Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180.

210. Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328.

211. Marshall, G. and Jones, R. H. (1995). Multi-state models and diabetic retinopathy. *Statistics in Medicine*, 14(18):1975–1983.

212. May, R. M. and Anderson, R. M. (1978). Regulation and stability of host-parasite population interactions: II. Destabilizing processes. *The Journal of Animal Ecology*, pages 249–267.

213. McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

214. McKane, A. J. and Newman, T. J. (2004). Stochastic models in population biology and their deterministic analogs. *Physical Review E*, 70(4):041902.

215. McKinley, T., Cook, A. R., and Deardon, R. (2009). Inference in epidemic models without likelihoods. *The International Journal of Biostatistics*, 5(1).

216. Meeds, E., Leenders, R., and Welling, M. (2015). Hamiltonian ABC. *arXiv preprint arXiv:1503.01916*.

217. Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C., and Andersen, P. K. (2009). Multi-state models for the analysis of time-to-event data. *Statistical methods in medical research*, 18(2):195–222.

218. Mettle, F. O., Osei Affi, P., and Twumasi, C. (2020). Modelling the transmission dynamics of tuberculosis in the Ashanti region of Ghana. *Interdisciplinary perspectives on infectious diseases*, 2020.

219. Metz, J. A. and Diekmann, O. (2014). *The dynamics of physiologically structured populations*, volume 68. Springer.

220. Midi, H. and Zahari, M. (2008). A simulation study on ridge regression estimators in the presence of outliers and multicollinearity. *Jurnal Teknologi*, pages 59â–74.

221. Mitrofani, I. A. and Koutras, V. P. (2021). A Branching Process Model for the Novel Coronavirus (Covid-19) Spread in Greece. *International Journal of Modeling and Optimization*, 11(3).

222. Morris, A., Börger, L., and Crooks, E. (2019). Individual variability in dispersal and invasion speed. *Mathematics*, 7(9):795.

223. Murphy, K. P. (2007). Conjugate Bayesian analysis of the Gaussian distribution. *def*, 1($2\sigma2$):16.

224. Nguyen, H. D., Arbel, J., Lü, H., and Forbes, F. (2020). Approximate Bayesian computation via the energy statistic. *IEEE Access*, 8:131683–131698.

225. Niss, M. (2005). History of the Lenz-Ising model 1920–1950: from ferromagnetic to cooperative phenomena. *Archive for history of exact sciences*, 59(3):267–318.

226. Norman, G. R., Sloan, J. A., and Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Medical care*, pages 582–592.

227. Numminen, E., Cheng, L., Gyllenberg, M., and Corander, J. (2013). Estimating the transmission dynamics of Streptococcus pneumoniae from strain prevalence data. *Biometrics*, 69(3):748–757.

228. Nunes, M. A. and Balding, D. J. (2010). On optimal selection of summary statistics for approximate Bayesian computation. *Statistical applications in genetics and molecular biology*, 9(1).

229. Ogawa, K. (1986). A monogenean parasite Gyrodactylus masu sp. n. (Monogenea: Gyrodactylidae) of salmonid fish in Japan. *Nippon Suisan Gakkaishi*, 52(6):947–950.

230. Ovaskainen, O. and Meerson, B. (2010). Stochastic models of population extinction. *Trends in ecology & evolution*, 25(11):643–652.

231. pada Das, K. et al. (2011). A mathematical study of a predator-prey dynamics with disease in predator. *ISRN Applied mathematics*, 2011.

232. Paisley, L., Karlsen, E., Jarp, J., and Mo, T. (1999). A Monte Carlo simulation model for assessing the risk of introduction of Gyrodactylus salaris to the Tana river, Norway. *Diseases of Aquatic Organisms*, 37(2):145–152.

233. Pandey, A., Atkins, K. E., Medlock, J., Wenzel, N., Townsend, J. P., Childs, J. E., Nyenswah, T. G., Ndeffo-Mbah, M. L., and Galvani, A. P. (2014). Strategies for containing Ebola in west Africa. *Science*, 346(6212):991–995.

234. Peeler, E., Gardiner, R., and Thrush, M. (2004). Qualitative risk assessment of routes of transmission of the exotic fish parasite *Gyrodactylus salaris* between river catchments in England and Wales. *Preventive Veterinary Medicine*, 64(2-4):175–189.

235. Peeler, E. and Thrush, M. (2004). Qualitative analysis of the risk of introducing

Gyrodactylus salaris into the United Kingdom. *Diseases of Aquatic Organisms*, 62(1-2):103–113.

236. Peeler, E., Thrush, M., Paisley, L., and Rodgers, C. (2006). An assessment of the risk of spreading the fish parasite *Gyrodactylus salaris* to uninfected territories in the European Union with the movement of live Atlantic salmon (Salmo salar) from coastal waters. *Aquaculture*, 258(1-4):187–197.

237. Pellis, L., Ball, F., Bansal, S., Eames, K., House, T., Isham, V., and Trapman, P. (2015). Eight challenges for network epidemic models. *Epidemics*, 10:58–62.

238. Perkins, S. E., Cagnacci, F., Stradiotto, A., Arnoldi, D., and Hudson, P. J. (2009). Comparison of social networks derived from ecological data: implications for inferring infectious disease dynamics. *Journal of Animal Ecology*, 78(5):1015–1022.

239. Peters, G. W., Fan, Y., and Sisson, S. A. (2012). On sequential Monte Carlo, partial rejection control and approximate Bayesian computation. *Statistics and Computing*, 22(6):1209–1222.

240. Phillips, J. M. and Venkatasubramanian, S. (2011). A gentle introduction to the kernel distance. *arXiv preprint arXiv:1103.1625.*

241. Pooley, C. M., Doeschl-Wilson, A. B., and Marion, G. (2022). Estimation of age-stratified contact rates during the COVID-19 pandemic using a novel inference algorithm. *medRxiv.*

242. Poulin, R. (2010). Network analysis shining light on parasite ecology and diversity. *Trends in parasitology*, 26(10):492–498.

243. Prangle, D. (2015). Summary statistics in approximate Bayesian computation. *arXiv preprint arXiv:1512.05633.*

244. Prangle, D. (2017). Adapting the ABC distance function. *Bayesian Analysis*, 12(1):289–309.

245. Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798.

246. Quine, M. and Szczotka, W. (1994). Generalisations of the Bienayme-Galton-Watson branching process via its representation as an embedded random walk. *The Annals of Applied Probability*, pages 1206–1222.

247. R Core Team (2019). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

248. Ramírez, R., Harris, P. D., and Bakke, T. A. (2012). An agent-based modelling approach to estimate error in gyrodactylid population growth. *International journal for parasitology*, 42(9):809–817.

249. Rathinam, M., Petzold, L. R., Cao, Y., and Gillespie, D. T. (2003). Stiffness in stochastic chemically reacting systems: The implicit tau-leaping method. *Journal of Chemical Physics*, 119(24):12784–12794.

250. Ratmann, O., Andrieu, C., Wiuf, C., and Richardson, S. (2009). Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proceedings of the National Academy of Sciences*, 106(26):10576–10581.

251. Ratmann, O., Jørgensen, O., Hinkley, T., Stumpf, M., Richardson, S., and Wiuf, C. (2007). Using likelihood-free inference to compare evolutionary dynamics of the protein networks of H. pylori and P. falciparum. *PLoS Computational Biology*, 3(11):e230.

252. Raynal, L., Marin, J.-M., Pudlo, P., Ribatet, M., Robert, C. P., and Estoup, A. (2019). ABC random forests for Bayesian parameter inference. *Bioinformatics*, 35(10):1720–1728.

253. Raynal, L. and Onnela, J.-P. (2021). Selection of Summary Statistics for Network Model Choice with Approximate Bayesian Computation. *arXiv preprint arXiv:2101.07766.*

254. Reimer, J. R., Bonsall, M. B., and Maini, P. K. (2017). The critical domain size of stochastic population models. *Journal of mathematical biology*, 74(3):755–782.

255. Renshaw, E. (1993). *Modelling biological populations in space and time*. Number 11. Cambridge University Press.

256. Richards, E. L., van Oosterhout, C., and Cable, J. (2010). Sex-specific differences in shoaling affect parasite transmission in guppies. *PLoS One*, 5(10):e13285.

257. Richards, G. and Chubb, J. (1996). Host response to initial and challenge infections, following treatment, of *Gyrodactylus bullatarudis* and *G. turnbulli* (Monogenea) on the guppy (*Poecilia reticulata*). *Parasitology Research*, 82(3):242–247.

258. Ridgway, J. (2017). Probably approximate Bayesian computation: nonasymptotic convergence of ABC under misspecification. *arXiv preprint arXiv:1707.05987*.

259. Robert, C. (2012). Contribution to the discussion of Fearnhead and Prangle (2012). *Journal of the Royal Statistical Society: Series B*, 74:447–448.

260. Robert, C. P., Cornuet, J.-M., Marin, J.-M., and Pillai, N. S. (2011). Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences*, 108(37):15112–15117.

261. Roberts, M., Andreasen, V., Lloyd, A., and Pellis, L. (2015). Nine challenges for deterministic epidemic models. *Epidemics*, 10:49–53.

262. Roberts, M. and Heesterbeek, J. (2003). *Mathematical models in epidemiology*, volume 215. EOLSS.

263. Rogers, J. L., Howard, K. I., and Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological bulletin*, 113(3):553.

264. Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, pages 1151–1172.

265. Rubio-Godoy, M., Muñoz-Córdova, G., Garduño-Lugo, M., Salazar-Ulloa, M., and Mercado-Vidal, G. (2012). Microhabitat use, not temperature, regulates intensity of *Gyrodactylus cichlidarum* long-term infection on farmed tilapia—Are parasites evading competition or immunity? *Veterinary parasitology*, 183(3-4):305–316.

266. Satsuma, J., Willox, R., Ramani, A., Grammaticos, B., and Cârstea, A. S. (2004). Extending the SIR epidemic model. *Physica A: Statistical Mechanics and its Applications*, 336(3-4):369–375.

267. Saulnier, E., Gascuel, O., and Alizon, S. (2017). Inferring epidemiological parameters from phylogenies using regression-ABC: A comparative study. *PLoS computational biology*, 13(3):e1005416.

268. Schelkle, B. (2012). *Gyrodactylid biology, transmission and control.* PhD thesis, Cardiff University.

269. Schwaferts, P. and Augustin, T. (2020). Bayesian decisions using regions of practical equivalence (ROPE): Foundations.

270. Scott, M. E. (1982). Reproductive potential of *Gyrodactylus bullatarudis* (Monogenea) on guppies (*Poecilia reticulata*). *Parasitology*, 85(2):217–236.

271. Scott, M. E. and Anderson, R. (1984). The population dynamics of *Gyrodactylus bullatarudis* (Monogenea) within laboratory populations of the fish host *Poecilia reticulata*. *Parasitology*, 89(1):159–194.

272. Sedki, M. and Pudlo, P. (2012). Contribution to the discussion of Fearnhead and Prangle (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B*, 74:466–467.

273. Shah, N. H. and Mittal, M. (2021). Introduction to Compartmental Models in Epidemiology. *Mathematical Analysis for Transmission of COVID-19*, page 1.

274. Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55.

275. Shaw, D. and Dobson, A. (1995). Patterns of macroparasite abundance and aggregation in wildlife populations: a quantitative review. *Parasitology*, 111(S1):S111–S133.

276. Shinn, A. A. P., Pratoomyot, J., Bron, J. E., Paladini, G. G., Brooker, E., and Brooker, A. J. (2015). Economic impacts of aquatic parasites on global finfish production. *Global aquaculture advocate*, (Setembro/Outubro):82–84.

277. Siler, W. (1979). A competing-risk model for animal mortality. *Ecology*, 60(4):750–757.

278. Silk, D., Filippi, S., and Stumpf, M. P. (2013). Optimizing threshold-schedules for sequential approximate Bayesian computation: applications to molecular systems. *Statistical applications in genetics and molecular biology*, 12(5):603–618.

279. Simola, U., Cisewski-Kehe, J., Gutmann, M. U., and Corander, J. (2021). Adaptive approximate Bayesian computation tolerance selection. *Bayesian analysis*, 16(2):397–423.

280. Simola, U., Pelssers, B., Barge, D., Conrad, J., and Corander, J. (2019). Machine learning accelerated likelihood-free event reconstruction in dark matter direct detection. *Journal of Instrumentation*, 14(03):P03004.

281. Sisson, S. and Fan, Y. (2018). ABC samplers. *Handbook of Approximate Bayesian Computation*, pages 87–123.

282. Sisson, S. A., Fan, Y., and Beaumont, M. (2018). *Handbook of approximate Bayesian computation*. CRC Press.

283. Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765.

284. Sisson, S. A., Fan, Y., and Tanaka, M. M. (2009). Correction for Sisson et al., Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 106(39):16889–16889.

285. Smith, D. L., Battle, K. E., Hay, S. I., Barker, C. M., Scott, T. W., and McKenzie, F. E. (2012). Ross, Macdonald, and a theory for the dynamics and control of mosquito-transmitted pathogens. *PLoS pathogens*, 8(4):e1002588.

286. Stephenson, J. F., Van Oosterhout, C., Mohammed, R. S., and Cable, J. (2015). Parasites of Trinidadian guppies: evidence for sex-and age-specific trait-mediated indirect effects of predators. *Ecology*, 96(2):489–498.

287. Stephenson, J. F., Young, K. A., Fox, J., Jokela, J., Cable, J., and Perkins, S. E. (2017). Host heterogeneity affects both parasite transmission to and fitness on subsequent hosts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1719):1–10.

288. Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., and Dessimoz, C. (2013). Approximate bayesian computation. *PLoS computational biology*, 9(1):e1002803.

289. Swinton, J., Woolhouse, M., Begon, M., Dobson, A., Ferroglio, E., Grenfell, B., Guberti, V., Hails, R., Heesterbeek, J., Lavazza, A., et al. (2002). Microparasite transmission and persistence.

290. Tallmon, D. A., Luikart, G., and Beaumont, M. A. (2004). Comparative evaluation of a new effective population size estimator based on approximate Bayesian computation. *Genetics*, 167(2):977–988.

291. Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518.

292. Taylor, N., Sommerville, C., and Wootten, R. (2005). A review of *Argulus* sp. occurring in UK freshwaters. *Bristol, UK, Environment Agency.*

293. The GIMP Development Team (2019). *GIMP.* https://www.gimp.org.

294. Thrall, P. H., Antonovics, J., and Hall, D. W. (1993). Host and pathogen coexistence in sexually transmitted and vector-borne diseases characterized by frequency-dependent disease transmission. *The American Naturalist*, 142(3):543–552.

295. Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202.

296. Tran, M. N. and Kohn, R. (2015). Exact ABC using importance sampling. *arXiv preprint arXiv:1509.08076.*

297. Treadway, T. and Twumasi, C. (2021). An Experimental Study of Lesions Observed in Bog Body Funerary Performances. *Experimental Archaeology Journal*, page 16. https://exarc.net/ark:/88735/10595.

298. Twumasi, C., Asiedu, L., and Nortey, E. N. (2019a). Markov chain modeling of HIV, tuberculosis, and Hepatitis B Transmission in Ghana. *Interdisciplinary perspectives on infectious diseases*, 2019.

299. Twumasi, C., Asiedu, L., and Nortey, E. N. (2019b). Statistical modeling of HIV, tuberculosis, and hepatitis B transmission in Ghana. *Canadian Journal of Infectious Diseases and Medical Microbiology*, 2019.

300. Twumasi, C. and Twumasi, J. (2022). Machine learning algorithms for forecasting and backcasting blood demand data with missing values and outliers: A study of Tema General Hospital of Ghana. *International Journal of Forecasting*, 38(3):1258–1277.

301. Uchmański, J. and Grimm, V. (1996). Individual-based modelling in ecology: what makes the difference? *Trends in Ecology & Evolution*, 11(10):437–441.

302. Vajargah, B. F. and Moradi, M. (2011). Period dependent branching process and its applications in epidemiology. *Infection, Genetics and Evolution*, 11(6):1225–1228.

303. van Oosterhout, C., Harris, P., and Cable, J. (2003). Marked variation in parasite resistance between two wild populations of the Trinidadian guppy, *Poecilia reticulata* (Pisces: Poeciliidae). *Biological Journal of the Linnean Society*, 79(4):645–651.

304. van Oosterhout, C., Joyce, D. A., and Cummings, S. M. (2006a). Evolution of MHC class IIB in the genome of wild and ornamental guppies, *Poecilia reticulata*. *Heredity*, 97(2):111–118.

305. van Oosterhout, C., Joyce, D. A., Cummings, S. M., Blais, J., Barson, N. J., Ramnarine, I. W., Mohammed, R. S., Persad, N., and Cable, J. (2006b). Balancing selection, random genetic drift, and genetic variation at the major histocompatibility complex in two wild populations of guppies (*Poecilia reticulata*). *Evolution*, 60(12):2562–2574.

306. van Oosterhout, C., Potter, R., Wright, H., and Cable, J. (2008). Gyro-scope: An individual-based computer model to forecast gyrodactylid infections on fish hosts. *International Journal for Parasitology*, 38(5):541–548.

307. van Oosterhout, C., Smith, A. M., Hänfling, B., Ramnarine, I. W., Mohammed, R. S., and Cable, J. (2007). The guppy as a conservation model: implications of parasitism and inbreeding for reintroduction success. *Conservation Biology*, 21(6):1573–1583.

308. van Wieringen, W. N. (2015). Lecture notes on ridge regression. *arXiv preprint arXiv:1509.09169*.

309. Vanhove, M. P., Boeger, W. A., Muterezi Bukinga, F., Volckaert, F. A., Huyse, T., and Pariselle, A. (2012). A new species of Gyrodactylus (Monogenea, Gyrodactylidae), an ectoparasite from the endemic Iranocichla hormuzensis (Teleostei, Cichlidae), the only Iranian cichlid. *European Journal of Taxonomy*, (30).

310. Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological methods*, 17(2):228.

311. Wallentin, G. and Neuwirth, C. (2017). Dynamic hybrid modelling: Switching between AB and SD designs of a predator-prey model. *Ecological Modelling*, 345:165–175.

312. Watson, H. W. and Galton, F. (1875). On the probability of the extinction of families. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 4:138–144.

313. Wegmann, D., Leuenberger, C., and Excoffier, L. (2009). Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, 182(4):1207–1218.

314. Weiss, H. H. (2013). The SIR model and the foundations of public health. *Materials matematics*, pages 0001–17.

315. Wermuth, N. (1972). *An empirical comparison of regression methods: a thesis.* PhD thesis, Harvard University.

316. Westlake, W. (1981). Bioequivalence testing–a need to rethink. *Biometrics*, 37(3):589–594.

317. Wilcox, R. R. and Serang, S. (2017). Hypothesis testing, p values, confidence intervals, measures of effect size, and Bayesian methods in light of modern robust techniques. *Educational and Psychological Measurement*, 77(4):673–689.

318. Wileman, D., Sangster, G., Breen, M., Ulmestrand, M., Soldal, A., and Harris, R. (1999). Roundfish and Nephrops survival after escape from commercial fishing gear. *EC Contract No: FAIR-CT95-0753. Final Report.*

319. Wilkinson, R. D. and Tavaré, S. (2009). Estimating primate divergence times by using conditioned birth-and-death processes. *Theoretical population biology*, 75(4):278–285.

320. Wilson, E. B. and Worcester, J. (1945). The law of mass action in epidemiology. *Proceedings of the National Academy of Sciences of the United States of America*, 31(1):24.

321. Wilson, K., Bjørnstad, O., Dobson, A., Merler, S., Poglayen, G., Randolph, S., Read, A., and Skorping, A. (2002). Heterogeneities in macroparasite infections: patterns and processes. *The ecology of wildlife diseases*, 44:6–44.

322. Wilson, K., Grenfell, B. T., and Shaw, D. J. (2006). Analysis of Aggregated Parasite Distributions: A Comparison of Methods. *Functional Ecology*, 10(5):592.

323. Wilson, W. G., Harrison, S. P., Hastings, A., and McCann, K. (1999). Exploring stable pattern formation in models of tussock moth populations. *Journal of animal ecology*, 68(1):94–107.

324. Wolfram, S. and Mallinckrodt, A. J. (1995). Cellular automata and complexity. *Computers in Physics*, 9(1):55–55.

325. Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104.

326. Yan, P. and Chowell, G. (2019). *Quantitative methods for investigating infectious disease outbreaks*, volume 70. Springer.

327. Zaric, G. S. (2002). Random vs. nonrandom mixing in network epidemic models. *Health care management science*, 5(2):147–155.

328. Zipkin, E. F., Jennelle, C. S., and Cooch, E. G. (2010). A primer on the application of Markov chains to the study of wildlife disease dynamics. *Methods in Ecology and Evolution*, 1(2):192–198.

Appendix A: Detailed visualization of fish heatmaps over eight body regions of fish over time

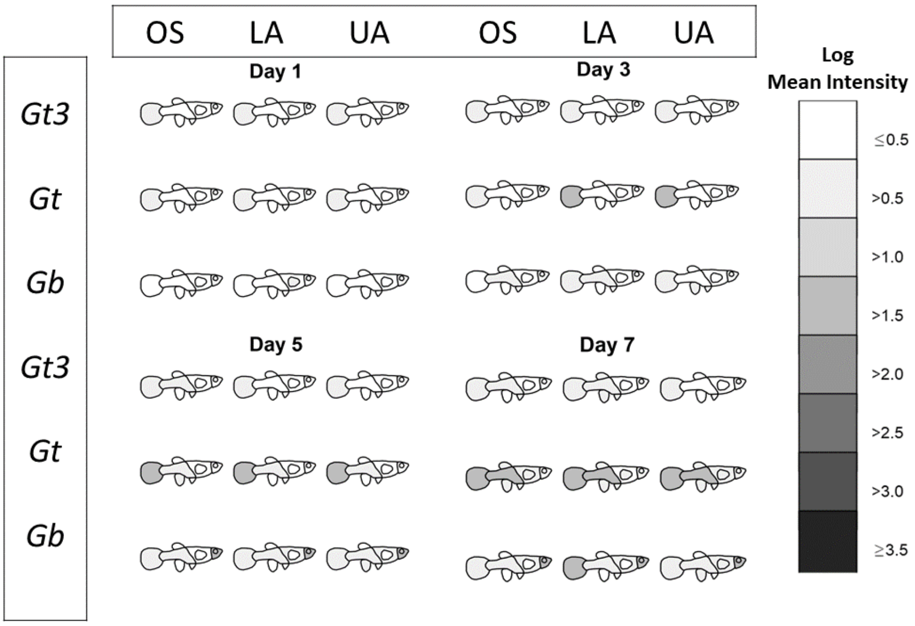

Figure A.1: Detailed visualization of fish heatmaps over eight body regions of fish across parasite strains and fish stocks from day 1 to 7.

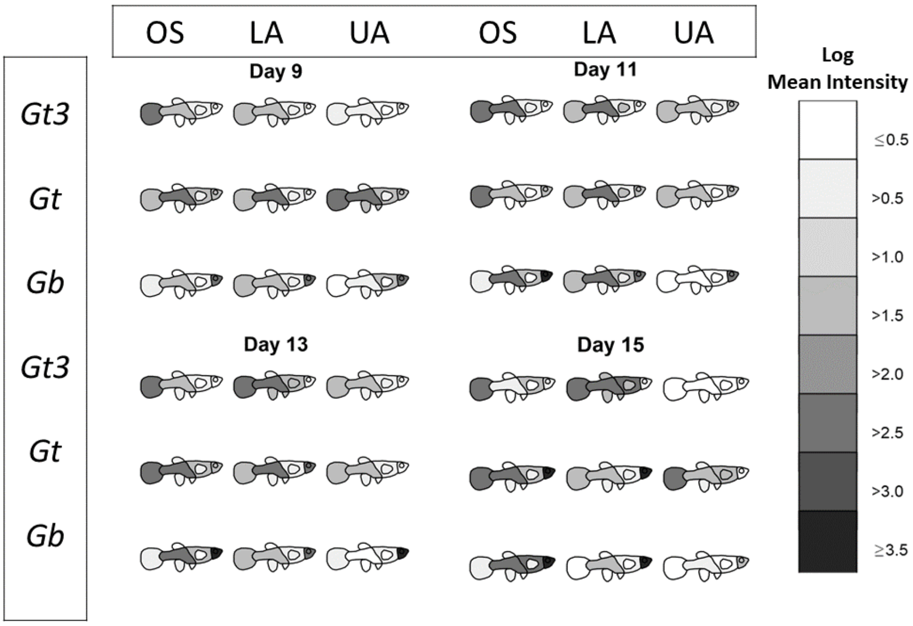

Figure A.2: Detailed visualization of fish heatmaps over eight body regions of fish across parasite strains and fish stocks from day 9 to 15.

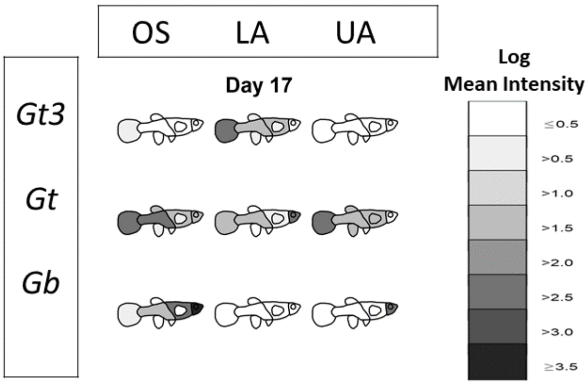

Figure A.3: Detailed visualization of fish heatmaps over eight body regions of fish across parasite strains and fish stocks on day 17.

Appendix B: Grouped barcharts showing variations in mean intensities at four main body regions of fish over time

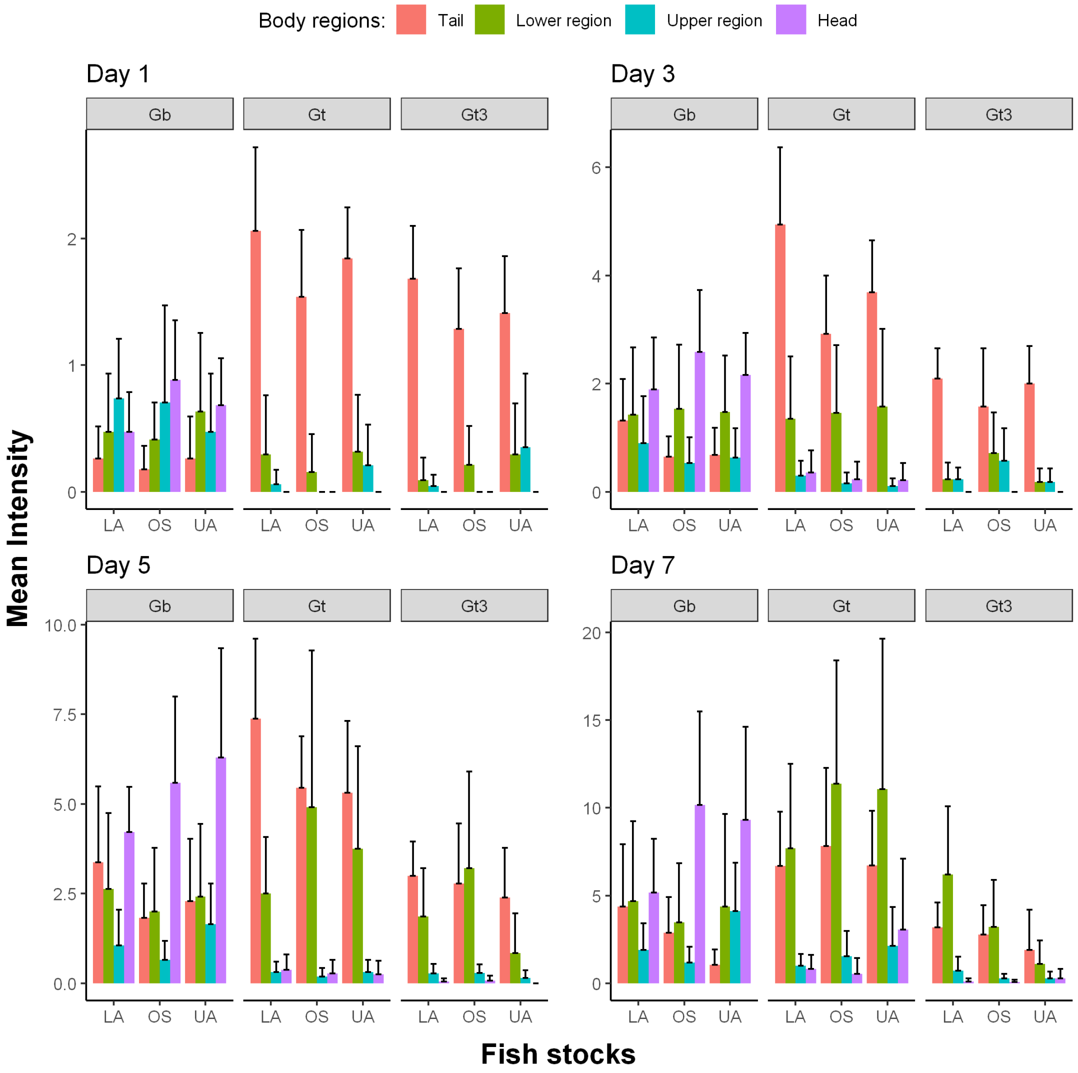

Figure A.4: Grouped barcharts showing variations in mean intensities at four main body regions of fish across parasite strains and fish stocks over surviving fish from day 1 to 7.

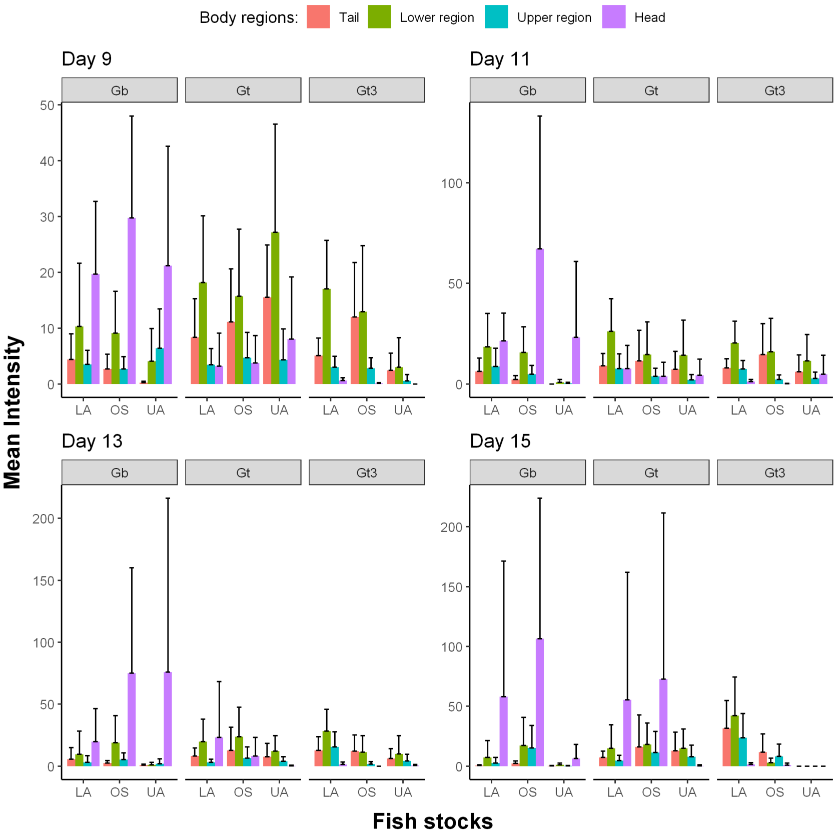

Figure A.5: Grouped barcharts showing variations in mean intensities at four main body regions of fish across parasite strains and fish stocks over surviving fish from day 9 to 15.

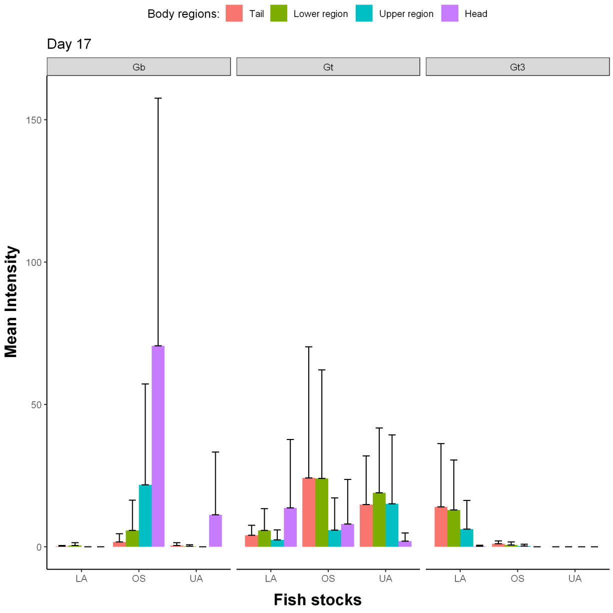

Figure A.6: Grouped barcharts showing variations in mean intensities at four main body regions of fish across parasite strains and fish stocks over surviving fish on day 17.

Appendix C: R Codes for Exact SSA of the B-D-C process

C.1: Function for updating events of the B-D-C process via exact SSA

```
SSA_update_event=function(X,fish_status,rate,total_rate){
#Let b,d & c be the birth, death and catastrophe parameters
#X be the number of parasites
  #fish_status <- 1 # fish starts out alive

  if (total_rate == 0) {
    return(list(X = X, t_incr = Inf)) # zero population
  }

  #Determine event occurence from single draw
  u<-runif(1,0, total_rate)
  if (u<abs(rate[1])){
    #birth of parasites
    X<-X+1
  } else if(u<abs(rate[1]+rate[2])){
    #death of parasites
    X<-X-1
  }  else {
    #catastrophe or death of fish
    X<-0
    fish_status<-2
  }
  t_incr <- rexp(1, total_rate) # time increment
  #Returns parasite numbers,time step and survival status
 return(list(X = X, t_incr=t_incr,fish_status=fish_status))

}
```

C.2: Function for exact stochastic simulation (SSA)

```
#Function for exact simulation of the B-D-C process
Exact_BDC<-function(X0,b,d,c,ti=0,tmax=30){
#Let ti be the initial time (set at 0)
#tfinal be the final simulation time
   rate<-numeric(3) #event rates
#stop simulation if total population exceeds this limit
  pop_max <- 10000
  #Time variable
  #ti<- 0; tmax<-30;
  save_ti <- 1:tmax #Discrete times to store simulation
  save_TF <- rep(FALSE, length(save_ti))
  #parasite pop over time
  pop <-matrix(NA,1,length(save_ti))
  # host host status at each time point
  alive <- rep(2, length(save_ti))
  alive_ti <- 1 #fish starts out alive
  X<-X0
  pop_ti <- sum(X)
  while(sum(save_TF) < length(save_ti)){
     #Calculate rate of events
     #probability of birth
     rate[1]<- b*X
     #probability of death
     rate[2]<- d*X
     #probability of catastrophe
     rate[3]<- c*X
     #tota rate
      total_rate<- rate[1]+rate[2]+rate[3]
     if(sum(pop_ti) > pop_max){
       cat("Popmax_exceeded","\n")
       break
                   }
     if(alive_ti == 2)  break

     output <-SSA_update_event(X,fish_status=alive_ti,
             rate=rate,total_rate=total_rate)
     #Update time to next event
     ti <- ti + output$t_incr
     #break if there is negative population
     if (X < 0) break
         # Events to occur
     save_new <- which((ti >= save_ti) & !save_TF)
     for (i in save_new){
       pop[,i]<- pop_ti
       alive[i] <- alive_ti
                     }
     save_TF <- (ti >= save_ti)
     X<- output$X
     pop_ti <- sum(X)
     alive_ti <- output$fish_status
            }
  #Returns the parasite numbers & survival status over time
  return(list(pop=pop,alive = alive))
}
```

Appendix D: Julia codes for computing the log-likelihood function

D.1: Computing constants of the B-D-C transition function

```
module BDCfit #begin module
using PolynomialRoots
export logL
function BDCconsts(lambda, mu, rho, t)
#lambda, mu, rho are B-D-C parameters respectively
# Computing constants of BDC process at time t
    rts = sort(real(roots([mu,-(lambda+mu+rho),lambda])))
    v0 = rts[1]
    v1 = rts[2]
    sigma = exp(-lambda*(v1 - v0)*t)
    k1 = v0*v1*(1 - sigma)/(v1 - sigma*v0)
    k2 = (v1*sigma - v0)/(v1 - sigma*v0)
    k3 = (1 - sigma)/(v1 - sigma*v0)
    return [k1, k2, k3]
end
function gamma_n_j(nmax)
# calculates gamma^n_j for n = 1, ..., nmax and j = 1, ...,n
    # used by ProbBDC
    gnj = zeros(BigInt, nmax, nmax)
    gnj[1,1] = 1
    if nmax > 1
        for n = 2:nmax
            gnj[n,1] = n*gnj[n-1,1]
        end
        for j = 2:nmax
            for n = j:nmax
                gnj[n,j] = gnj[n-1,j-1] + (n+j-1)*gnj[n-1,j]
            end
        end
    end
    return gnj
end
function delta_m_j(mmax, k1, k2, k3)
# calculates delta^m_j for n = 1, ..., mmax and j = 1, ...,n
# used by ProbBDC; k1, k2, k3 will be output from BDCconsts
    k = (k2 + k1*k3)/k1/k3
    dmj = zeros(BigFloat, mmax, mmax)
    dmj[1,1] = k
    if mmax == 1
        return dmj
    else
        for m = 2:mmax
            dmj[m,1] = k*m
            for j = 2:m
                dmj[m,j] = k*(m - j + 1)*dmj[m,j-1]
            end
        end
        return dmj
    end
end
```

D.2: Computing the B-D-C transition function

```
function ProbBDC(lambda, mu, rho, t, mmax, nmax)
    # P(X_t=n | X_0=m) for -1 <= m <= mmax and
    #-1 <= n <= nmax
    # where -1 indicates extinction by catastrophe
    cc = BDCconsts(lambda, mu, rho, t)
    k1 = cc[1]
    k2 = cc[2]
    k3 = cc[3]
    k4 = (k1 + k2)/(1 - k3)
    P = zeros(Float64, mmax+2, nmax+2)
    P[1,1] = 1
    P[2,2] = 1
    k1_powers = zeros(BigFloat, mmax)
    k3_powers = zeros(BigFloat, nmax)
    k4_powers = zeros(BigFloat, mmax)
    facts = zeros(BigFloat, nmax)
    k1_powers[1] = k1
    k4_powers[1] = k4
    P[3,1] = Float64(1 - k4)
    P[3,2] = Float64(k1)
    for m = 2:mmax
        k1_powers[m] = k1*k1_powers[m-1]
        k4_powers[m] = k4*k4_powers[m-1]
        P[m+2,1] = Float64(1 - k4_powers[m])
        P[m+2,2] = Float64(k1_powers[m])
    end
    k3_powers[1] = k3
    facts[1] = 1
    for n = 2:nmax
        k3_powers[n] = k3*k3_powers[n-1]
        facts[n] = n*facts[n-1]
    end
    gnj = gamma_n_j(nmax)
    dmj = delta_m_j(mmax, k1, k2, k3)
    for m = 1:mmax
        for n = 1:nmax
            x = BigFloat(0)
            for j = 1:(min(m,n))
                x = x + gnj[n,j]*dmj[m,j]
            end
 P[m+2,n+2] = Float64(x*k1_powers[m]*k3_powers[n]/facts[n])
        end
    end
    return P
end
```

D.3: Computing the B-D-C log-likelihood function

```
function logL(lambda, mu, rho, x)
 # calculate the log likelihood for params:
 # lambda, mu, rho and data x
 # each row of x are population at times:
 # t= 1, 3, 5, 7, 9, 11, 13, 15, 17
 # assume population at time 0 is 2;
 # state -1 indicates catastrophe
    mmax1 = 2
    nmax1 = Int64(max(maximum(x[:,1]), 2))
    P1 = ProbBDC(lambda, mu, rho, 1, mmax1, nmax1)
    mmax2 = Int64(max(maximum(x[:,1:8]), 2))
    nmax2 = Int64(max(maximum(x), 2))
    P2 = ProbBDC(lambda, mu, rho, 2, mmax2, nmax2)
    el = 0
    for i = 1:size(x, 1) # logL for observation i
        # time 0 to time 1 transition
        el = el + log(P1[4, Int64(x[i,1]+2)])
        for j = 1:8
            # time 2j-1 to time 2j+1 transition
       el = el + log(P2[Int64(x[i,j]+2), Int64(x[i,j+1]+2)])
        end
    end
    return el
end

end #module
```

E.1: Function for updating B-D-C Hybrid τ-leaping simulation

(see Appendix E)

```
#Function to update tau-leaping

tauleap_update<-function(X,tau,fish_status,rate,total_rate){
      #Inputs:
      #X=parasite number, tau=leap size, rate=event rates
      #fish_status=survival status, total_rate=total rate
      if(runif(1) < rate[3]*tau){ # catastrophe

          X <- 0
          fish_status<-2
      }else{ # births and deaths
          X <- X + rpois(1,  rate[1]*tau)
           - rpois(1,rate[2]*tau)
      }
    #Returns the parasite numbers & survival status
    return(list(X = X,fish_status=fish_status))
}
```

Appendix E: R Codes for B-D-C Hybrid τ-leaping algorithms

E.2: Function for τ-leaping based on Gillespie 2001

```
HTL2001<-function(X0,b,d,c,error,ti=0,tmax=30){
  #ti<-0 #initial time, X0=initial population size
  #tmax<-30  #final time
  rate<-numeric(3) #store event rates
  save_ti <- 1:tmax  #Times to simulate
  # host fish status at each time point
  alive <- rep(2, length(save_ti))
  alive_ti <- 1 #fish starts out alive
  save_TF <- rep(FALSE, length(save_ti))
  # parasite pop at observed time point
  pop <-matrix(NA,1,length(save_ti))
  X<-X0;pop_ti <- sum(X)
  while(ti<tmax){
    #Computing event rates (birth,death & catastrophe)
    rate[1]<- b*X;rate[2]<- d*X;rate[3]<- c*X
    #representing a0(x) or total rate
    total_rate<- rate[1]+rate[2]+rate[3]
    #Computing tau on Gillespie (2001)
    tau<-(error*(b+d))/(abs(b-d)*max(b,d))
    #Switching condition
    leap_condition<- 2/total_rate  #leap condition
    #Running Tau-leaping
    if(tau>leap_condition){#Execute tau-leaping
      ti <- ti + tau #update time
      output<-tauleap_update(X,tau=tau,fish_status=alive_ti,
      rate=rate,total_rate=total_rate)
      X<-output$X
    } #end of tau-leaping
     #Running exact SSA algorithm if tau<=leap_condition
    else {#Execute exact SSA
      output<SSA_update_event(X,fish_status=alive_ti,
      rate=rate,total_rate=total_rate)
      X<- output$X;ti <- ti +output$t_incr# update time
    if (X < 0) break #break if there is negative population
    if (alive_ti == 2) break
    # saving output
    save_new <- which((ti >= save_ti) & !save_TF)
    for (i in save_new){
       pop[,i]<- pop_ti; alive[i] <- alive_ti
                  }
       save_TF <- (ti >= save_ti)
    X<- X;pop_ti<- sum(X);alive_ti <- output$fish_status
            }
  #Returns the parasite numbers & survival status over time
  return(list(pop=pop,alive=alive))
}
```

E.3: Function for τ-leaping based on Gillespie and Petzold (2003)

```
HTL2003<-function(X0,b,d,c,error,ti=0,tmax=30){
  #ti<-0 #initial time, X0=initial population size
  #tmax<-30  #final time
  rate<-numeric(3) #store event rates
  Leap_sizes<- NULL #store leap size
  save_ti <- 1:tmax  #Times to simulate
  # host fish status at each time point
  alive <- rep(2, length(save_ti))
  alive_ti <- 1 #fish starts out alive
  save_TF <- rep(FALSE, length(save_ti))
  # parasite pop at observed time point
  pop <-matrix(NA,1,length(save_ti))
  X<-X0;pop_ti <- sum(X)
  while(ti<tmax){
    #Computing event rates (birth,death & catastrophe)
    rate[1]<- b*X;rate[2]<- d*X;rate[3]<- c*X
    #representing a0(x) or total rate
    total_rate<- rate[1]+rate[2]+rate[3]
    #Computing tau on Gillespie & Petzold 2003
    Leap_sizes[[1]]<- (error*(b+d))/(abs(b-d)*max(b,d))
    Leap_sizes[[2]]<- X*(error*(b+d))^2/((b+d)*max(b^2,d^2))
    tau<- min(Leap_sizes[[1]],Leap_sizes[[2]])#leap size
    #Switching condition
    leap_condition<- (1/(10*total_rate)) #leap condition
    #Running Tau-leaping
    if(tau>leap_condition){#Execute tau-leaping
      ti <- ti + tau #update time
      output<-tauleap_update(X,tau=tau,fish_status=alive_ti,
      rate=rate,total_rate=total_rate)
      X<-output$X
    } #end of tau-leaping
     #Running exact SSA algorithm if tau<=leap_condition
    else {#Execute exact SSA
      output<SSA_update_event(X,fish_status=alive_ti,
      rate=rate,total_rate=total_rate)
      X<- output$X; ti <- ti +output$t_incr# update time
    if (X < 0) break #break if there is negative population
    if (alive_ti == 2) break
    # saving output
    save_new <- which((ti >= save_ti) & !save_TF)
    for (i in save_new){
       pop[,i]<- pop_ti;alive[i] <- alive_ti
                  }
       save_TF <- (ti >= save_ti)
    X<- X;pop_ti<- sum(X);alive_ti <- output$fish_status
  }

  #Returns the parasite numbers & survival status over time
  return(list(pop=pop,alive=alive))
}
```

Appendix F: R Codes for the modified weighted-iterative ABC & ABC Post-Processing Regression Analysis

F.1: Functions for population projection & weighted distances

```
## 1. Function for population projection
#until day 17 after host mortality

#ga= gamma which is tuning parameter (set at 0.9)
project <- function(pop_single, alive_single, ga) {

  # project parasite numbers beyond fish mortality
  n <- length(alive_single)
  k <- sum(alive_single == 1)
  if (k == n) return(pop_single)
  if (k == 0) return(matrix(0, nrow=4, ncol=n))
  if (k == 1) return(matrix(pop_single[,1], nrow=4, ncol=n))
  z <- log(colSums(pop_single[,1:k],na.rm=T))
  al <- sum( (z[k] - z[1:(k-1)]) * ((k-1):1)
  * ga^((k-1):1),na.rm=T) /
  sum( ((k-1):1)^2 * ga^((k-1):1),na.rm=T)

  pop_single[,(k+1):n] <- pop_single[,k] %*%
   t( exp( (1:(n-k))*al ) )
  return(pop_single)
}

#converting function to byte-code compilation
project_compiler=cmpfun(project)


## 2. Function for computing weighted distance
#between simulated and observed summary statistics

w_distance <- function(S1, S2, weight)  {
  n<- dim(S1)[1]
  #squared difference between matrix S1 & S2
  Squared_diff_mat<- (S1-S2)^2
  #Multiplying vector to weights
  Weighted_sq_diff<- lapply(1:dim(S1)[1],
             function(k) weight*Squared_diff_mat[k, ])
  #total weighted distances (WSS)
  WSS<- do.call("sum",Weighted_sq_diff)
  #return a scaled weighted sum of squares distance
  return(sqrt(WSS/n))
}

#converting function to byte-code compilation
distance_compiler=cmpfun(w_distance)
```

F.2(i): External functions for Galton-Watson & GMM estimators for B-D-C parameter estimation

```
# 1. Function for computing BDC constants and PGF
BDCconsts <- function(lambda, mu, rho,t) {
# Constants used in calculating distribution of BDC process at time t

  roots <- sort(Re(polyroot(c(mu, -(lambda+mu+rho), lambda))))
  v0 <- roots[1]
  v1 <- roots[2]
  sigma <- exp(-lambda*(v1 - v0)*t)
  k1 <- v0*v1*(1 - sigma)/(v1 - sigma*v0)
  k2 <- (v1*sigma - v0)/(v1 - sigma*v0)
  k3 <- (1 - sigma)/(v1 - sigma*v0)
  return(list(k1=k1, k2=k2, k3=k3, sigma=sigma, v0=v0, v1=v1))
}



# 2. Function for the probability generating function G(z,t)
PGF_z<- function(lambda,mu,rho,t,z,m){
   #v0<-((lambda+mu+rho)-sqrt( ((lambda+mu+rho)^2)-4*mu*lambda))/(2*lambda)
   #v1<-((lambda+mu+rho)+sqrt( ((lambda+mu+rho)^2)-4*mu*lambda))/(2*lambda)
    constants=BDCconsts(lambda,mu,rho,t)
    v0<- constants$v0
    v1<- constants$v1
    sigma<- constants$sigma
    num<-(v0*v1*(1-sigma))+(z*(v1*sigma-v0))
    den<- v1-(sigma*v0)-(z*(1-sigma))
    return( (num/den)^m)
}

#3. Analytical probability of death due to catastrophe
#Estimating C(t)=P(catastrophe resulting in 0 population|host death)
Prob_catastrophe<- function(lambda,mu,rho,t,z=1,m=2){
             constant<- 1-PGF_z_compiler(lambda=lambda,
             mu=mu,rho=rho,t=t,z=z,m=m)
             #return the probability of catastrophic extinction
             return(constant)
    }



    #4. Function of the Exact mean/1st moment of the BDC process
First_moment<-function(b,d,c,t,m){
  #b,d,c are the birth,death and catastrophe rates; m=X0=2 and t=time
  roots <- sort(Re(polyroot(c(d, -(b+d+c), b))))
  v0 <- roots[1]
  v1 <- roots[2]
  sigma<-exp(-b*(v1-v0)*t)
  k1<-(v0*v1*(1-sigma))/(v1-(sigma*v0))
  k2<-((v1*sigma)-v0)/(v1-(sigma*v0))
  k3<-(1-sigma)/(v1-(sigma*v0))
  expectation=m*(((k1+k2)/(1-k3))^(m-1))*(k2+(k1*k3))*(1-k3)^-2
  return(expectation)#returns 1st moment
}
```

F.2(ii): External functions for Galton-Watson & GMM estimators for B-D-C parameter estimation

```
# 1. Function of the 2nd moment of the BDC process
Second_moment<-function(b,d,c,t,m){
        roots <- sort(Re(polyroot(c(d, -(b+d+c), b))))
        v0 <- roots[1]
        v1 <- roots[2]
        sigma<-exp(-b*(v1-v0)*t)
        k1<- (v0*v1*(1-sigma))/(v1-(sigma*v0))
        k2<-((v1*sigma)-v0)/(v1-(sigma*v0))
        k3<-(1-sigma)/(v1-(sigma*v0))
       expectation<- m*(((k1+k2)/(1-k3))^(m-1))*(k2+(k1*k3))*(1-k3)^-2

       Second_derivative_pgf<-((2*m*k3*(k2+k1*k3))*
       ((k1+k2)/(1-k3))^(m-1)*(1-k3)^-3 +
       m*(m-1)*(k2+k1*k3)^2*
       ((k1+k2)/(1-k3))^(m-2)*(1-k3)^-4)
       Variance<-(Second_derivative_pgf+ expectation)-(expectation)^2

      Second_moment_results<- Variance + expectation^2
      return(Second_moment_results)#returns 2nd moment
}


# 2. Function of the 3rd moment of the BDC process

Third_moment<-function(b,d,c,t,m){
       roots <- sort(Re(polyroot(c(d, -(b+d+c), b))))
       v0 <- roots[1]
       v1 <- roots[2]
       sigma<-exp(-b*(v1-v0)*t)
       k1<-(v0*v1*(1-sigma))/(v1-(sigma*v0))
       k2<-((v1*sigma)-v0)/(v1-(sigma*v0))
       k3<-(1-sigma)/(v1-(sigma*v0))
       expectation<- m*(((k1+k2)/(1-k3))^(m-1))*(k2+(k1*k3))*(1-k3)^-2

       Second_derivative_pgf<-((2*m*k3*(k2+k1*k3))
       *((k1+k2)/(1-k3))^(m-1)*(1-k3)^-3 +
       m*(m-1)*(k2+k1*k3)^2*
       ((k1+k2)/(1-k3))^(m-2)*(1-k3)^-4)

       Third_derivative_pgf<- 6*m*(k2+k1*k3)*(k3^2)*
       (((k1+k2)/(1-k3))^(m-1))*(1-k3)^(-4)+
                   6*m*(m-1)*((k2+k1*k3)^2)*k3*
                   (((k1+k2)/(1-k3))^(m-2))*(1-k3)^(-5)+
                   m*(m-1)*(m-2)*((k2+k1*k3)^3)*
                   (((k1+k2)/(1-k3))^(m-3))*(1-k3)^(-6)

       Variance<-(Second_derivative_pgf+ expectation)-(expectation)^2

       Second_moment_results<- Variance + expectation^2

       Third_moment_results<- Third_derivative_pgf
       +(3*Second_moment_results)-(2*expectation)

       return(Third_moment_results)#returns 3rd moment
}
```

F.2(iii): External functions for Galton-Watson & GMM estimators for B-D-C parameter estimation

```
# 1. Set the catastrophe state -1 to 0
zero.catastrophe <- function (x) {
    x[x<0] <- 0
    return(x)
}


# 2. Set the ratio Z(i)/Z(i-1) to 1 if NA
#(due to case of 0/0 in Z(i)/Z(i-1))
one.ratio <- function (x) {
    x[is.na(x)|x==Inf|x==-Inf] <- 1
    return(x)
}

# 3. functions for sample moments
sample_mean_1st<- function(x) sum(x)/length(x)
sample_mean_2nd<- function(x) sum(x^2)/length(x)
sample_mean_3rd<- function(x) sum(x^3)/length(x)

### Computing the 2-step GMM estimates ####

time<-seq(1,17,by=2)

# 4. Objective function for 1st step of GMM
g_objectivefunc_firstStep <- function(x,prob_sample,
        fixed=c(FALSE,FALSE,FALSE)) {
        Prob_catastrophe_analytical<- rep(NA,length=length(time))
        params<-fixed
        function(p){
        params[!fixed]<-p
        #The three parameters to be optimized
        b1<-params[1]
        d1<-params[2]
        c1<-params[3]


        #Computing theoritical prob of catastrophe
        for(i in seq_along(time)){
           Prob_catastrophe_analytical[i]<-Prob_catastrophe(
            lambda=b1,mu=d1,rho=c1,t=time[i])
                                }

        m1 <- First_moment(b=b1,d=d1,
        c=c1,t=seq(1,17,by=2),m=2)-
        apply(zero.catastrophe(x),1,sample_mean_1st)
        m2 <- Second_moment(b=b1, d=d1,
        c=c1,t=seq(1,17,by=2),m=2)-
        apply(zero.catastrophe(x),1,sample_mean_2nd)
        m3 <- Third_moment(b=b1, d=d1,
        c=c1,t=seq(1,17,by=2),m=2)-
        apply(zero.catastrophe(x),1,sample_mean_3rd)

       Catastrophe_Prob<- Prob_catastrophe_analytical- prob_sample

       gbar_theta<-c(mean(m1),mean(m2),mean(m3),mean(Catastrophe_Prob))

       Objective_func<- t(gbar_theta)%*%gbar_theta

                   }

    }
```

F.2(iv): External functions for Galton-Watson & GMM estimators for B-D-C parameter estimation

```
### Computing the 2-step GMM estimates(continued) ####

#First step of GMM
GMM_firstStep<-function(prob_sample,x){
      objec_func<- g_objectivefunc_firstStep(x=x,prob_sample=prob_sample)
      initial<-c(2, 1, 0.001)# initial values to optimize over
      estimates<-constrOptim(initial, objec_func, NULL,
                    ui=rbind(c(1,0,0),   # lambda>0
                            c(0,1,0),     # mu >0
                            c(0,0,1) # rho > 0
                    ),
      ci=c(0,0,0),method='Nelder-Mead')$par

      return(estimates)
  }


# Second step of GMM
#Second-step of the GMM optimization
#Function to calculating the weight matrix
Weight<-function(x,prob_sample,estimate1){
 est_step1<- c(estimate1)
 Prob_catastrophe_analytical1<- rep(NA,length=length(time))
 #Computing theoretical prob of catastrophe
 for(i in seq_along(time)){
     Prob_catastrophe_analytical1[i]<- Prob_catastrophe(
     lambda=est_step1[1],mu=est_step1[2],
     rho=est_step1[3],t=time[i])
                              }

  m1 <- First_moment(b=est_step1[1],
  d=est_step1[2],c=est_step1[3],t=seq(1,17,by=2),m=2)
  -apply(zero.catastrophe(x),1,sample_mean_1st)
  m2 <- Second_moment(b=est_step1[1],
  d=est_step1[2],c=est_step1[3],t=seq(1,17,by=2),m=2)
  -apply(zero.catastrophe(x),1,sample_mean_2nd)
  m3 <- Third_moment(b=est_step1[1],d=est_step1[2]
  ,c=est_step1[3],t=seq(1,17,by=2),m=2)
  -apply(zero.catastrophe(x),1,sample_mean_3rd)
  Catastrophe_Prob<- Prob_catastrophe_analytical1- prob_sample

  g<-cbind(m1,m2,m3,Catastrophe_Prob)

  covariance_matrix<- cov(g)
  #Setting off-diagonals to 0 to obtain an
  #invertible weighting (diagonal) matrix
  #by assuming that the moment conditions are uncorrelated
  covariance_matrix[lower.tri(covariance_matrix)] <- 0
  covariance_matrix[upper.tri(covariance_matrix)] <- 0

  #Finding inverse for the covariance diagonal matrix
     #finding reciprocal of entries
     weightmatrix<- 1/covariance_matrix
     weightmatrix[lower.tri(weightmatrix)] <- 0
     weightmatrix[upper.tri(weightmatrix)] <- 0
     weightmatrix

}
```

F.2(v): External functions for Galton-Watson & GMM estimators for B-D-C parameter estimation

```
### Computing the 2-step GMM estimates(continued) ####
#Second optimization step
g_objectivefunc_2ndStep <- function(x,prob_sample,
weighting_matrix,fixed=c(FALSE,FALSE,FALSE)) {
  Prob_catastrophe_analytical<-rep(NA,length=length(time))
  params<-fixed
  function(p){
        params[!fixed]<-p
        #The three parameters to be optimized
        b1<-params[1]
        d1<-params[2]
        c1<-params[3]


#Computing theoritical prob of catastrophe
 for(i in seq_along(time)){
     Prob_catastrophe_analytical[i]<- Prob_catastrophe(lambda=b1,mu=d1,
                                         rho=c1,t=time[i])
                            }

  m1 <-First_moment(b=b1, d=d1,c=c1,t=seq(1,17,by=2),m=2)
  -apply(zero.catastrophe(x),1,sample_mean_1st)
  m2 <-Second_moment(b=b1, d=d1,c=c1,t=seq(1,17,by=2),m=2)
  -apply(zero.catastrophe(x),1,sample_mean_2nd)
  m3 <-Third_moment(b=b1, d=d1,c=c1,t=seq(1,17,by=2),m=2)
  -apply(zero.catastrophe(x),1,sample_mean_3rd)

  Catastrophe_Prob<-Prob_catastrophe_analytical- prob_sample

  gbar_theta<-c(mean(m1),mean(m2),mean(m3),
  mean(Catastrophe_Prob))

  Objective_func<- t(gbar_theta)%*%
  weighting_matrix%*%gbar_theta

                 }

    }

#second step of GMM
GMM_2ndStep<-function(prob_sample,x,weighting_matrix){
    objec_func<- g_objectivefunc_2ndStep(x=x,
    prob_sample=prob_sample,weighting_matrix=
    weighting_matrix)
    # initial values to optimize over
    initial<-c(2, 1, 0.001)
    estimates=constrOptim(initial, objec_func, NULL,
                    ui=rbind(c(1,0,0),   # lambda>0
                            c(0,1,0),     # mu >0
                            c(0,0,1) # rho > 0
            ),
    ci=c(0,0,0),method='Nelder-Mead')$par
   return(estimates)
  }
```

F.2(vi): External functions for Galton-Watson & GMM estimators for B-D-C parameter estimation

```
#Restructuring data format for GW-GMM BDC estimation
RestructureData_BDC<- function(pop,alive,group){
        #Inputs:pop=parasite population per
                #region over time
                #alive= survival status over time
                # group=parasite-fish groups

       # to store parasite numbers over
       #time as a dataframe for each parasite-fish
        ParasiteData_combined<- NULL
       # to store  survival status as a
       #dataframe for each parasite-fish
        SurvStatus_combined<- NULL

       #Set NA in pop to state 0 denoting host
       # death for the B-D-C estimation

        for(pf in seq_along(group)){

            ParasiteData_combined[[pf]]<- matrix(NA,
            nrow=9, ncol=numF[[pf]])
            #Array for time steps fish was alive
            #for each combination
            SurvStatus_combined[[pf]]<-  matrix(NA,
            nrow=9, ncol=numF[[pf]])
            for(i in 1:numF[[pf]]){
                #total parasites over time for each
                #fish belonging to each parasite-fish group
                # state -1 in the BDC denote host death
                ParasiteData_combined[[pf]][,i]<-
                na.zero(apply(pop[[pf]][i,,],2,sum))
                SurvStatus_combined[[pf]][,i]<-
                alive[[pf]][i, ]
                }
        }
       return(list(PopTime_group=ParasiteData_combined,
       SurvTime_group=SurvStatus_combined))
}
```

F.2(vii): External functions for Galton-Watson & GMM estimators for B-D-C parameter estimation

```
#Function for finding Maximum likelihood estimates
#for the catastrophe rate given the GW estimate of
#birth and death rates for the B-D-C model
MLE_catastrophe<-function(b_est,d_est,dead_fish_time){
          log_like<-0
      #LogLikelihood function to maximize
      Catastrophe_Loglik<-function(param){
         rho<-param[1]

      #log likelihood function for catastraphe rate
        for(i in dead_fish_time){
        #sum across all dead fish for each group
            if(i>=3){#if the time to death>=3
               log_like<-log_like+
               na.zero(log(Prob_catastrophe(lambda=b_est,
               mu=d_est,rho=rho,t=i))-
               Prob_catastrophe(lambda=b_est,
               mu=d_est,rho=rho,t=(i-2)))
            }else{#if the time to death=1
               log_like<-log_like+
               na.zero(log(Prob_catastrophe(lambda=b_est,
               mu=d_est,rho=rho,t=i)-
               Prob_catastrophe(lambda=b_est,mu=d_est,
               rho=rho,t=0)))
                            }

                    }
        log_like
  }

    Catastrophe_Loglik_compiler=cmpfun(Catastrophe_Loglik)

  ## Inequality constraints:  rho>0

   estimates<-maxLik(logLik=Catastrophe_Loglik_compiler,
   start=c(rho= 1e-5))

  #returning estimates of catastrophe rate
  return(as.vector(estimates$estimate))
}

#External scripts
source("MLE_catastrophe-script.r")
source("GMM-1st2nd-Steps-script.r")
```

F.3: Function for computing the B-D-C model parameters as extra ABC summary statistics using Galton-Watson & GMM estimations

```
GW_GMM_BDCestimator<-function(X0,pop,alive,group){
  #X0= initial parasites
 Parasite_data<- NULL; survival_data<- NULL
 #re-structuring the format of the data into the
 # 9 parasite-fish groups
 data<- RestructureData_BDC(pop=pop,alive=alive,group=group)

 time<-seq(1,17,by=2)
 # Parasite_data[[pf]][,fish_index]
 Prob_catastrophe_analytical=Prob_catastrophe_sample=
  matrix(0,nrow=length(time),ncol=length(group))
             ## Initialize GMM   ###
 #Computing catastrophic probability analytically
 # & based on the sample data

 #computing sample probability of catastrophe

 time_index<- seq_along(time)
 for (pf in seq_along(group)){
    Parasite_data[[pf]]<- data$PopTime_group[[pf]]
    survival_data[[pf]]<- data$SurvTime_group[[pf]]
    for(i in time_index){
       if(any(survival_data[[pf]][i,]==2)==TRUE){
         #print(paste("time=",time[i]))
          fish_dead_sim<-length(which(
          survival_data[[pf]][i, ]==2))
          #print( fish_dead_sim)
          Prob_catastrophe_sample[i,pf]<-
          fish_dead_sim/dim(survival_data[[pf]])[2]
            }
         }

     }
   #Let Zi_t be the population for fish i at time t
   # Let alive_status be the survival status of each fish
    Z=NULL; alive_status=NULL
    for(pf in seq_along(group)) {
        Z[[pf]]<-list()
        alive_status[[pf]]<-list()
                            }

##GW_GMM_BDCestimator function continues at the next page##
+++
#Continuation of GW_GMM_BDCestimator function
for(pf in seq_along(group)){
        for(k in 1:numF[[pf]]){
           Z[[pf]][[k]]<- Parasite_data[[pf]][,k]
           alive_status[[pf]][[k]]<-survival_data[[pf]][,k]
                       }
               }
    #Computing the mean and variance for the
    #Galton-Watson process based on fish survival
       # And for each k replicate
     mean_GW=NULL; var_GW=NULL; mean_sum_num=NULL;
     mean_sum_den=NULL;var_sum=NULL
            #Computing the mean of GW process
     for(pf in seq_along(group))
      #initial summation for the GW mean
      mean_sum_num[[pf]]=mean_sum_den[[pf]]=0
         for(pf in seq_along(group)){
           for(k in 1:numF[[pf]]){
               if(all(survival_data[[pf]][,k]==1)==TRUE){

                 mean_sum_num[[pf]]<-mean_sum_num[[pf]]+
                 sum(Z[[pf]][[k]][1:9])# sum from t1-t17
                 mean_sum_den[[pf]]<-mean_sum_den[[pf]]+
                 sum(Z[[pf]][[k]][1:8])+X0 #sum from t0-t15
                                                  }
                                            }
               mean_GW[[pf]]<- one.ratio(mean_sum_num[[pf]]/
               mean_sum_den[[pf]])#if 0/0=1
                                     }
       #computing the variance of GW process
       #initial summation for GW variance
     for(pf in seq_along(group)) var_sum[[pf]]<-0
         for(pf in seq_along(group)){
            for(k in 1:numF[[pf]]){
               if(all(survival_data[[pf]][,k]==1)==TRUE){
                  var_sum[[pf]]<-var_sum[[pf]]+
                  sum(Z[[pf]][[k]][1:9]*
                   (one.ratio(Z[[pf]][[k]][1:9]/
                   c(X0,Z[[pf]][[k]][1:8])) -
                   mean_GW[[pf]])^2)
                                               }
                                      }
                 var_GW[[pf]]<- var_sum[[pf]]/
                 (numF[[pf]]*length(time))
                 }
    ###    GMM estimation ###
    birth_rate=NULL;death_rate=NULL; c_estimates<-
    NULL;delta_t=2;
    BDC_estimates=NULL
    GMM_resultsStep1=NULL; GMM_resultsStep2=NULL;
     weighting_matrix_cov=NULL; method=NULL

    #Estimating the catastrophe rate
    #using MLE when m>1 for GW estimation
    #time at death for each fish i and
    #replicate/simulation run k
    t_death<-NULL;
    for(pf in seq_along(group)){ t_death[[pf]]<-rep(NA,length=numF[[pf]])   }
     for(pf in seq_along(group)){
           for(k in 1:numF[[pf]]){
               #time to death
              t_death[[pf]][k]<-time[which(
              survival_data[[pf]][,k]==2)[1]]
                           }
                    }
 ##GW_GMM_BDCestimator function continues at the next page##
+++
#Continuation of GW_GMM_BDCestimator function
 for(pf in seq_along(group)){### begining of GW and GMM
       if(mean_GW[[pf]]>1){####Consider GW if mean_GW>1
            method[[pf]]<-"GW␣estimation"
          birth_rate[[pf]]<-((log(mean_GW[[pf]])
          /(2*delta_t))*(one.ratio(var_GW[[pf]]
                 /(mean_GW[[pf]]*(mean_GW[[pf]]-1))) +1))
          death_rate[[pf]]<- ((log(mean_GW[[pf]])
          /(2*delta_t))*(one.ratio(var_GW[[pf]]/
          (mean_GW[[pf]]*(mean_GW[[pf]]-1))) -1))

         #Computing MLE of catastrophe rate
         #based on estimated birth and death rates
          if(all(is.na(t_death[[pf]]))==FALSE){
          #if at least some fish are dead
         #estimates of the catastrophe rate
             c_estimates[[pf]]<-MLE_catastrophe_compiler(
             b_est=birth_rate[[pf]],d_est<-death_rate[[pf]],
             dead_fish_time=na.omit(t_death[[pf]][k]))
                 }else if(all(is.na(t_death[[pf]]))==TRUE){
                 #if no fish is dead
             c_estimates[[pf]]<-0
                     }

         BDC_estimates[[pf]]<-c(birth_rate[[pf]],
         death_rate[[pf]],c_estimates[[pf]])
    }else if(mean_GW[[pf]]<=1){ #Consider GMM
            method[[pf]]<-"GMM␣estimation"
           #First stage of GMM
           GMM_resultsStep1[[pf]]<- GMM_firstStep(
           prob_sample=Prob_catastrophe_sample[,pf],
           x=as.data.frame(Parasite_data[[pf]]))

           weighting_matrix_cov[[pf]]<-Weight(x=
           as.data.frame(Parasite_data[[pf]]),
           prob_sample=Prob_catastrophe_sample[,pf],
           estimate1=GMM_resultsStep1[[pf]])

          #Second stage of GMM
          GMM_resultsStep2[[pf]]<- GMM_2ndStep(
          prob_sample=Prob_catastrophe_sample[,pf],
          x=as.data.frame(Parasite_data[[pf]]),
          weighting_matrix=weighting_matrix_cov[[pf]])

          BDC_estimates[[pf]]<-  GMM_resultsStep2[[pf]]
                                                       } ####GMM estimation ends
                            } #### end of GW and GMM
  #Returning the B-D-C parameters and method used
 BDC_estimates_df<-do.call("rbind", BDC_estimates)
 return(list(BDC_estimates=BDC_estimates_df,
 method_used=unique(unlist(method))))
}
```

F.4(i): Functions for initial prior & sampling proposals

```
#Prior distribution of model parameters (on log scale)

prior<- function() {
  lb1<- runif(2, -4, 1)# birth of parasites (Gt3)
  # birth rate for young parasites based on lb (Gt3)
  logb11 <- max(lb1)
  logb12<- min(lb1)# birth rate for older parasites based on lb1 (Gt3)

  lb2<- runif(2, -4, 1)# birth of parasites (Gt)
  logb21 <- max(lb2)  # birth rate for young parasites based on lb2 (Gt)
  logb22<- min(lb2)# birth rate for older parasites based on lb2 (Gt)

  lb3<- runif(2, -4, 1)# birth of parasites (Gb)
  logb31 <- max(lb3)# birth rate for young parasites based on lb3 (Gb)
  logb32<- min(lb3)# birth rate for older parasites based on lb3 (Gb)

  ld1 <- runif(2, -5, 2) #death rates (Gt3)
  logd11 <- min(ld1)# death rate without an immune response (Gt3)
  logd12 <- max(ld1)# death rate with immune response (Gt3)

  ld2 <- runif(2, -5, 2)# death rates (Gt)
  logd21 <- min(ld2)#death rate without an immune response (Gt)
  logd22 <- max(ld2)# death rate with immune response (Gt)

  ld3 <- runif(2, -5, 2) # death rates (Gb)
  logd31 <- min(ld3)# death rate without an immune response (Gb)
  logd32 <- max(ld3)# death rate with immune response (Gb)

  logm<- runif(1, -4, 1) #movement rate

  logr <- runif(1, -10, 1)#immune response rate (base rate)

  # immune response (adjustment for LA fish)
  logr1 <- runif(1, -10, 1)
  logr2 <- runif(1, -10, 1)# immune response rate (adj for OS fish)
  logr3 <- runif(1, -10, 1)# immune response rate (adj for male fish)

  logs <- runif(1, -8, -2)#fish mortality rate (base rate)

  logs1 <- runif(1, -8, -2)#fish mortality (adj for male fish)

  loge1 <- runif(1, -8, log(2)) #movement rate adj (Gt3)

  loge2 <- runif(1, -8, log(2))#movement rate adj (Gt)

  loge3 <- runif(1, -8, log(2))#movement rate adj (Gb)

  log_kappa <- runif(1, 4.5, 6.5)#effective carrying capacity

  #Returns the prior samples on log scale
  return(c(logb11, logb12,logb21, logb22,logb31,
   logb32,logd11, logd12,logd21, logd22,logd31, logd32,
   logm, logr,logr1,logr2,logr3,logs,logs1,loge1,
   loge2,loge3,log_kappa))
}


#view next page for the perturbation kernel function
```

```
#MVN kernel given optimal bandwidth matrix H
#For peturbation
MultivNorm_rkernel<- function(Num,bandwidth_matrix){
   dim_k<- dim(bandwidth_matrix)[2]
   mean_vector<- rep(0,dim_k)
   #return random noise from MVN kernel
   return(tmvtnorm::rtmvnorm(n=1, mean=mean_vector,
   sigma=bandwidth_matrix,
   lower=rep(-.1,dim_k),upper=rep(.1,dim_k),
   algorithm=c("gibbs")))
}


## Function for importance proposal sampling
post <- function(samp=tha_post,importance_weight=weight,
                 optimal_bw_matrix=Sigma_optimal_t){

  ##new proposal based on accepted priors (samp)##
  #number of previous accepted samples
  n <- dim(samp)[1]
  sample.particle<-sample(n, 1,prob=importance_weight)
  # Perturbing sampled particle based on MVN kernel
  KDE_sampler<- samp[sample.particle, ]
                +MultivNorm_rkernel(Num=1,
                 bandwidth_matrix=optimal_bw_matrix)

  new_proposal<- KDE_sampler; x<- new_proposal

  # birth rate of young>old
  x[1:2] <- sort(x[1:2], decreasing=TRUE)
  x[3:4] <- sort(x[3:4], decreasing=TRUE)
  x[5:6] <- sort(x[5:6], decreasing=TRUE)
  #death rates (without and with immune response)
  x[7:8] <- sort(x[7:8],decreasing=FALSE)
  x[9:10] <- sort(x[9:10],decreasing=FALSE)
  x[11:12] <- sort(x[11:12],decreasing=FALSE)
  return(x)
}
```

F.4(ii): Functions for computing initial summary statistics weights & setting other initial conditions for the modified ABC

```
## Computing initial weights ###
A0 <- matrix(0, 4, 2)
A0[1, 1] <- 2   #Intial parasites at the tail
B0 <- rep(1, 4) #initial immune response at 4 body regions

#Transition matrix
J<- matrix(c(0,    1,    0,     0,
             1/2,   0,    1/2,   0,
             0,    1/2,    0,   1/2,
             0,     0,     1,    0), 4, 4, byrow=TRUE)


# initial summary statistics weights estimate
dimS<-17 #length of summary statistics for ABC
n0 <- 100 #number of simulations for initial weights
#saving summary statistic for each group sim realisation
#for computing intial weights for ABC fitting
SummaryStats_sim <- NULL;SummaryStats_sim_combined<-NULL

for (i in 1:n0) {
    theta<- prior()
    output<- SimGroup_tauleap(theta1=theta,
    fish_sex=fishSex,fish_type=Fish_stock,
    strain=Strain,fish_size=fishSize,error=0.01)

    #B-D-C parameter estimates for the
    #parasite-fish groups based on simulated data
    #for each simulation realisation
    BDC_estimates_sim<-GW_GMM_BDCestimator(X0=2,
    pop=output$pop_sim,output$alive_sim,
    group=parasite_fish)$BDC_estimates

    #Computing the summary stats for each sim realisation
    SummaryStats_sim[[i]] <- Summary_stats(
    pop=output$pop_sim,alive=output$alive_sim,
    BDC_estimates=BDC_estimates_sim)
    #combining for all summary stats of
    #parasite-fish groups for each simulation realisation
    SummaryStats_sim_combined[[i]]<-do.call("rbind",
    SummaryStats_sim[[i]])
    }
#dimension is rows=(n0*total_fish) by cols=17
S0<- do.call("rbind",SummaryStats_sim_combined)
#initial weight (inverse of summary statistics)
w <- 1/apply(S0, 2, var, na.rm = TRUE)
print(w)# printing initial  weights
```

F.4(iii): Functions for returning priors, summaries and distances

```
#Function for returning priors, summaries and distances
ABC <- function(fork, pftn , n, w ) {
  # pftn is prior function or sampling proposals
  # n is number of samples or proposals
  # w are summary statistics weights
  dimS<-17 #dimension or number of ABC summary statistics
  number_of_parameters<- 23 #number of model parameters
  # matrix of prior distributions
  theta  <- matrix(nrow = n, ncol = number_of_parameters)
  #storing the summary stats across all simulations
  S_i <- NULL
  #S is a matrix(nrow = n*total_fish, ncol = dimS)
  d <- rep(NA, n)# weighted distance
  SummaryStats_sim <- NULL
  w<- w/sum(w) #normalising summary statistics weights

  for (i in 1:n) {
    theta[i,] <- pftn()
    output<- SimGroup_tauleap(theta1=theta[i,],fish_sex=fishSex,
    fish_type=Fish_stock,strain=Strain,fish_size=fishSize,
    error=0.01)
    #B-D-C parameter estimates for the parasite-fish groups
    # based on simulated data & simulation realisations
    BDC_estimates_sim<- GW_GMM_BDCestimator(X0=2,
    pop=output$pop_sim,output$alive_sim,
    group=parasite_fish)$BDC_estimates
    #Computing the all summary stats for each group
    SummaryStats_sim[[i]]<-Summary_stats(pop=output$pop_sim,
    alive=output$alive_sim,BDC_estimates=BDC_estimates_sim)
    #combining for all summary stats of parasite-fish
    #groups for each simulation realisation
    SummaryStats_sim_combined<-do.call("rbind",
    SummaryStats_sim[[i]])
    #Combining the observed summaries for the groups
    SummaryStats_obs_combined<- do.call("rbind", summaries_obs)

    #Storing weighted distances between summaries
    #of observed and simulated data
    S_i[[i]] <- SummaryStats_sim_combined
    d[i] <- w_distance(S1=S_i[[i]],
     S2=SummaryStats_obs_combined, weight=w)
  }
  # summary stats matrix(nrow = n*total_fish,
  #ncol = dimS)
  S<-do.call("rbind",S_i)
  #returns priors (theta), simulated summaries (S)
  #& distances (d)
  return(list(theta=theta, S=S, d=d))
}
```

F.4(iv): The modified weighted-iterative ABC (with SMC & SIS)

```
#Function to obtain the final posterior distribution iteratively
#using the ABC() function
Weighted_iterative_ABC<- function(N=500,dimS=17,
fish_total=Total_fish,numCores=numCores,
ABC_time_steps=10){
  # N= total number of samples
  n_cores <- numCores;n<- N/n_cores #Run on  n cores
  #number of parameters to be estimated
  number_of_parameters<- 23
  #Storing importance weight for sequential sampling
  import_weights<- NULL
  #Storing weights corresponding to accepted samples
  w_accepted<- NULL
  #saving number of particles for each iteration
  dim_tha_post<-NULL
  #saving summaries of all fish for each simulation
  S_i <- NULL
  #ABC_time_steps= time for the algorithm to terminate
  eps<-NULL # storage for index of accepted particles
  #proportion of sample to retain during SIS
  if(N<1000){
    epsilon<- c(0.5,0.43,0.4,0.35,0.3,
    0.2,0.1,0.08,0.06,0.02)
  }else if(N>=1000){
    epsilon<-c(0.5,0.3,0.2,0.1,0.08,
    0.07,0.06,0.03,0.02,0.01)
  }
  d_i<-NULL;d<-NULL#storing weighted distances
  #For storing parameter values at time t
  theta_i<- NULL;theta<-NULL
  # for density plots (256 used here is
  #the number of equally spaced points
  #at which the density is to be estimated)
  #range of prior distribution (on log scale)
  x <- seq(from = -10, to = 7, length.out = 256)
  fx <- array(dim=c(ABC_time_steps+1,
  number_of_parameters, 256))
   time0<- proc.time()
  for (t in 1:ABC_time_steps) {
    cat("ABC_time_steps", t, "\n")
    if (t == 1) {
      pftn <- prior
      ABC_out <- mclapply(1:n_cores, ABC,
      pftn=pftn, n=n, w=w, mc.cores=n_cores)
      for (i in 1:n_cores) {
        theta_i[[i]] <- ABC_out[[i]]$theta
        S_i[[i]] <- ABC_out[[i]]$S
        d_i[[i]] <- ABC_out[[i]]$d
      }
    }else{#if t>1
      #Calculate optimal MVN kernel bandwidth matrix
      #parameter values for a new proposal sample
      #N0= number of accepted particles
      #N1= total number of proposal samples
+++
#Calculate optimal MVN kernel bandwidth matrix H
# denoted by Sigma_optimal_t
#eps[[t]]= index of accepted samples
#w_accepted[[t]]= weights of accepted particles
#tha_post= accepted proposals at time t
 Sigma_optimal_t<- matrix(0,nrow=number_of_parameters,
 ncol=number_of_parameters)
      N1<-dim(theta[[t-1]])[1]
      N0<- dim(tha_post)[1]
      for(i in 1:N1) {
        for(k in 1:N0){
          Sigma_optimal_t<- Sigma_optimal_t+
          (import_weights[[t-1]][i]*w_accepted[[t-1]][k]
          *(matrix(tha_post[k,]-theta[[t-1]][i, ])
          %*%t(matrix(tha_post[k,]-theta[[t-1]][i, ]))))
        }
      }

      #Sampling from MVN Perturbation kernel
      weight<-w_accepted[[t-1]]
      pftn <- function() post(tha_post,weight,
      Sigma_optimal_t)
      ABC_out <- mclapply(1:n_cores, ABC,
      pftn=pftn, n=n, w=w, mc.cores=n_cores)
      for (i in 1:n_cores) {
        theta_i[[i]] <- ABC_out[[i]]$theta
        S_i[[i]] <- ABC_out[[i]]$S
        d_i[[i]] <- ABC_out[[i]]$d
      }
      #Combining theta at time t
      theta[[t]]<- as.matrix(na.zero(
      do.call("rbind",theta_i)))#N by 23 matrix
      #Re-weighting for importance sampling
      import_weights[[t]]<-rep(NA,
      length=dim(theta[[t]])[1])

    #Evaluating the perturbation kernel
    #for each particle at time t
      dMVN_func<- function(i)
      mvtnorm::dmvnorm(x=theta[[t]][i, ],
      mean = theta[[t-1]][i, ],sigma =Sigma_optimal_t)
      K_normal_kernel<- mclapply(1:dim(theta[[t]])
      [1],dMVN_func,mc.cores=n_cores)
      #### KDE of proposal distn####
      #Estimating the optimal bandwidth
      density_proposals<- matrix(NA,
       nrow=length(import_weights[[t]])
      ,ncol=number_of_parameters)
      N1<-length(unlist(K_normal_kernel))

+++
for(i in seq_along(import_weights[[t]])){
        density_proposals[i, ]<-
        ks::kde(x = theta[[t]][i, ],eval.points =
        theta[[t]][i, ])$estimate
        #KDE value for each proposal sample
        par.weight.numerator<-mean(density_proposals[i, ])
        par.weight.denominator<- sum(import_weights[[t-1]]
        [1:N1]*unlist(K_normal_kernel))
        import_weights[[t]][i]<- par.weight.numerator/
        par.weight.denominator
      }
      #normalizing weights
      import_weights[[t]]<- import_weights[[t]]/
      sum(import_weights[[t]])
    }
#Combining results from the ncores
    # N by 23 matrix
     theta[[t]]<- as.matrix(
     na.zero(do.call("rbind",theta_i)))#N by 23 matrix
    d[[t]]<- na.zeros(do.call("c",d_i))#length of N
    #number of draw for posterior samples
    small_draws<- epsilon[t]*N
    #adding the computed distance as extra column of theta
    theta_dist<- cbind(theta[[t]],d[[t]])
    #smallest distance index
    eps[[t]]<- order(theta_dist[,24])[1:small_draws]

    # choose posterior samples
    tha_post<-theta_dist[eps[[t]],][,-24]
    dim_tha_post[[t]]<- dim(tha_post)[1]
    #initialize importance weight for sequential sampling
    if(t==1)  import_weights[[1]]<- rep(1/N,length=N)

    #Weights corresponding to accepted proposal samples
    w_accepted[[t]]<- import_weights[[t]][eps[[t]]]
    w_accepted[[t]]<- w_accepted[[t]]/
    sum(w_accepted[[t]])#normalising accepted weights

    # update summary statistics weights
    #max least distance
    eps_dist_max <- sort(d[[t]])[small_draws]
    #combining the summaries[(N*fish_total) by 17 matrix]
    S<-na.omit(do.call("rbind",S_i))
    w1inv<-apply(S[rep(d[[t]],fish_total)<=eps_dist_max,]
    ,2, var, na.rm = TRUE)
    w <- na.zero(2/(1/w + w1inv))


+++
 # densities
    if (t == 1) {
      for (k in 1:number_of_parameters){
        fx[1,k,]<- density(theta[[1]][ ,k], from=-10, to=7, n=256)$y
        #saving the densities for each iteration
        write.csv(fx[1, ,],file =
        paste0("density_post_", 1, ".csv"))
      }
    }
    for (k in 1:number_of_parameters){
      fx[t+1,k,]<- density(tha_post[, k],from=-10, to=7, n=256)$y
      #saving the densities for each iteration
      write.csv(fx[t+1, ,], file =
      paste0("density_post_", t+1, ".csv"))
    }
###saving importance weights
    write.csv(import_weights[[t]],
    file = paste0("importance_weights_",t, ".csv"))
    #accepted particles at each iteration
    write.csv(tha_post, file =
    paste0("theta_post_", t, ".csv"))
    #saving weighted distance
    write.csv(d[[t]],file =
    paste0("weighted_distance_", t, ".csv"))
  }
  timef<- proc.time()-time0
  CPUtime<-sum(as.vector(timef)[-3])
  write.csv(CPUtime,file=paste0("CPUtime_", N, ".csv"))
  #Returns estimated densities & final posterior
  return(list(fx=fx,final_posteior=tha_post))
} #end of the weighted-iterative ABC algorithm
```

```
#External functions in the posterior adjustment func.

#Gaussian kernel with bandwidth delta
guass_kernel<- function(dist,delta){
    #bandwidth=delta for regression adjustment
    #is optimally determined using the kedd package
    kern<-(sqrt(2*pi*delta))*exp(-(dist^2)/(2* delta^2))
    return(kern)
}

#To deal with any possible unknown irregularity
na.inf.zero<- function(x){
    x[is.na(x)|is.finite(x)==FALSE]<- 0
    return(x)
}
```

F.4(v): Function for proposed ABC post-processing analysis

```
#Function for the modified local-linear regression
# based on weighted ridge regression
require("kedd")
Post_Ridge_reg_adj<- function(post_distn,summary_obs){
  #k=biasing parameter or  penalty parameter
  # post_dtn is the posterior sample
  # w are summary statistics weights
  #storing the summary stats across simulations
  S_i <- NULL
  no_of_parameters<- 23
  #storing adjusted posterior means
  posterior_mean_adj<- rep(NA,no_of_parameters)
  #Combining the summary stats for
  # the observed data for the parasite-fish groups
  SummaryStats_obs_combined<- do.call("rbind",
  summaries_obs)
  m<- dim(post_distn)[1]# m=number of posterior samples
  d<- rep(0, m)#weighted distances given observed data
  p<- dim(summary_obs)[2] #dimension of summary statistics
  Unadj_dist<- post_distn
  SummaryStats_sim <- NULL
  X_Design_matrix<- matrix(NA, ncol=p,nrow=m) #design matrix
  # Weights based on Gaussian kernel
  #for local-linear regression adjustment
  W<- matrix(0, ncol=m,nrow=m)
  #saving weighted column means of design matrix
  X_bar=numeric(length=p)
  for (i in 1:m) {
    theta<- as.vector(unlist(post_distn[i,]))
    output_sim<- SimGroup_tauleap(theta1=theta,
    fish_sex=fishSex,fish_type=Fish_stock,
    strain=Strain,fish_size=fishSize,error=0.01)
    #B-D-C parameter estimates for the
    #parasite-fish groups based on simulated data
    #for each simulation realisation
    BDC_estimates_sim<- GW_GMM_BDCestimator(X0=2,
    pop=output_sim$pop_sim,
       output_sim$alive_sim,
       group=parasite_fish)$BDC_estimates

    #Computing the summary stats for each
    #group simulation realisation
    SummaryStats_sim[[i]] <- Summary_stats(
    pop=output_sim$pop_sim,alive=output_sim$alive_sim,
    BDC_estimates=BDC_estimates_sim)

    #combining for all summary stats of
    #parasite-fish groups for each simulation realisation
    SummaryStats_sim_combined<-do.call("rbind",
    SummaryStats_sim[[i]])
    mean_diff<- apply(SummaryStats_sim_combined-
    summary_obs,2,mean,na.rm = TRUE)
+++
    #storing each row of design matrix X
    X_Design_matrix[i, ]<- mean_diff

    # Computing weights based on
    #Storing weighted distances between
    #summaries of observed and simulated data)
    S_i[[i]] <- SummaryStats_sim_combined
    #Updating summary statistics weights
    w <-apply(S_i[[i]], 2, var, na.rm = TRUE)
    w<- w/sum(w) #normalising summary weights
    d[i] <- w_distance(S1=S_i[[i]],
    S2=summary_obs, weight=w)
  }

  distances<-na.inf.zero(d)
  #Adaptively choosing the bandwidth
  #of the Gaussian kernel based on the distances
  bandwidth<- kedd::h.amise(x=distances,
  deriv.order =0,kernel = c("gaussian"))$h
  diag(W)<- guass_kernel(dist= distances,delta=bandwidth)
  theta_post<- as.matrix(post_distn)
  # (normalising) main diagonal of Weighting matrix
  weights<- diag(W)/sum(diag(W))

  #Transforming X and Y (posterior distribution
  #and summary statistics)
  for(j in seq_along(posterior_mean_adj)){
    #For each jth model parameter, j=1,2,...23
    X<- X_Design_matrix
    Y<- theta_post[ ,j]

    #Step 1 (Mean centring X and Y)
    for (k in 1:p) X_bar[k]<- sum(weights*X[,k])

    for (k in 1:p) {
      X[, k]<- X[, k]-X_bar[k]
    }
    #finding the weighted mean of Y and mean centring
    Y_bar <- sum(weights*Y)
    Y<- Y- Y_bar

+++
    #Step 2: scaling (centred X and Y) by weights

    for(k in 1:p) X[, k]<- sqrt(weights)*X[, k]
    Y<- sqrt(weights)*Y

    #Choose optimal value of k (the penalty paramters)
    # Using cross validation glmnet
    # Setting the range of lambda values
    options(warn = -1)
    lambda_seq <- 10^seq(2, -2, by = -.1)
    ridge_cv <- cv.glmnet(X, Y, alpha = 0,
     lambda =lambda_seq)
    # Best lambda value
    best_lambda <- ridge_cv$lambda.min
    k<-best_lambda
    # calculate beta estimates corresponding
    #to summary statistics X (standardised coefficients)
    beta_ridge_std <- solve(t(X) %*%W%*%X+
     k*diag(p)) %*% t(X)%*%W%*%Y


    # calculating beta estimates of predictors
    beta_ridge <- solve(t(X) %*%W%*%X+ k*diag(p))
    %*% t(X)%*%W%*%Y


    #calculate intercept estimates (adjusted posterior mean)
    posterior_mean_adj[j]<- exp(Y_bar - X_bar%*%beta_ridge)

    #Adjusting the posterior distribution
    Unadj_dist[,j]<- post_distn[, j]-X_Design_matrix
    %*%beta_ridge
  }

  posterior_mean_uadj<- exp(apply(post_distn,2,mean))
  Posterior_mean_output<- data.frame(Adj_posterior_mean=
  posterior_mean_adj,Uadj_posterior_mean=
  posterior_mean_uadj)

  #returns the design data matrix, adjusted & unadjusted
  # means, and the adjusted posterior distribution
  return(list(X_Design_matrix=X,
  Posterior_mean_output=Posterior_mean_output,
  Adjusted_posterior_dist=Unadj_dist))
}
```

Appendix G: R Codes for the multidimensional simulation model

G.1: Description of state variables and simulation parameters

```
## 1. State variables ##

# A[j,k] gives the number of parasites at location j, age k, where
  #    j = 1 for Tail population
  #    j = 2 for Lower region population
  #    j = 3 for Upper region population
  #    j = 4 for head population
  #    k = 1 for young parasites (yet to give birth)
  #    k = 2 for old parasites (have given birth)

  # B[j] = immune response at location j (1 for no response; 2 for a response)

  # X = state of fish (1 for alive; 2 for dead)


## 2. Base simulation parameters ##

  # b1[k,el] = birth rate for parasites age k, when immune state is el (for Gt3)
  # b2[k,el] = birth rate for parasites age k, when immune state is el (for Gt)
  # b3[k,el] = birth rate for parasites age k, when immune state is el (for Gb)
  # d1[k,el] = death rate for parasites age k, when immune state is el (for Gt3)
  # d2[k,el] = death rate for parasites age k, when immune state is el (for Gt)
  # d3[k,el] = death rate for parasites age k, when immune state is el (for Gb)
  # m[k,el] = movement rate for parasites age k, when immune state is el
 # e = the adjustment to the movement rate for forward/backward movement
  # r = rate a single parasite increases immune state (base rate)
  # kappa = effective carrying capacity per unit area of each body region
  # s = rate a single parasite causes fish mortality


## 3. Additional simulation parameters ##

  # r1 = immune response rate adjustment for LA fish (ref: UA fish)
  # r2 = immune response rate adjustment for OS fish (ref: UA fish)
  # r3 = immune response rate adjustment for male fish (ref: female fish)
  # e1, e2, e3 = movement rate adjustment depending on
 parasite type (Gt3,Gt, Gb respectively)
  #s = must depend on total parasite numbers, fish sex and fish size
  #s1 = host mortality rate with adjustment for male fish (ref: female)


## 4. Experiment descriptors ##

  # fish_type (1 for UA, 2 for LA & 3 for OS)
  # Parasite type (Gt3, Gt & Gb)
  # fish_sex (1 for female fish & 2 for male fish)
  # f= area of each body part (depends on size and gender)
  # a= fish size

#Function to convert NA's to 0 where necessary
 na.zero<-function(x){
    x[is.na(x)]<-0
    return(x)
}

#Loading packages (R packages to install)
library(transport)#for Wasserstein distance computation
library(parallel)# for parallizing R codes
RNGkind("L'Ecuyer-CMRG")#Dealing with distinct seed numbers
library(compiler)# byte code compilation
library("maxLik")#for MLE/optimization
library("R.utils")
```

G.2: Function for computing event rates

```
# Function for computing rates based on fish sex, fish type and parasite strain
compute_rates_func<- function(A, B, b1,b2,b3, d1,d2,d3, m,
                      r,r1,r2,r3,s,s1,e1,e2,e3,
                      kappa,f,a,fish_sex,fish_type,strain){

#Matrix of immune rates (additive effect of covariates)
  r_matrix<- matrix(c(r,r+r1,r+r2,r+r3,r+r1+r3,r+r2+r3),nrow=2,ncol=3,byrow=T)
 #r_selected= selected rate based on adjustments (adj) for fish sex &fish type
   #selecting the immune response rate depending on fish sex and fish type
   if(fish_sex=="F" & fish_type=="UA"){r_selected<-  r_matrix[1,1]}#base rate
   if(fish_sex=="F" & fish_type=="LA"){r_selected<-r_matrix[1,2]}#adj for LA fish
   if(fish_sex=="F" & fish_type=="OS"){r_selected<-r_matrix[1,3]}#adj for OS fish
   if(fish_sex=="M" & fish_type=="UA"){r_selected<-r_matrix[2,1]}#adj for male fish
   if(fish_sex=="M" & fish_type=="LA"){r_selected<-r_matrix[2,2]}#adj for M & LA
   if(fish_sex=="M" & fish_type=="OS"){r_selected<-r_matrix[2,3]}#adj for M & OS

    # selecting which host mortality rate & body areas given fish sex
   if(fish_sex=="F"){
       s_selected<- s #base host mortality rate
       #f=body_area
       f<-as.vector(f[,1])# body areas for female fish
   }
   if(fish_sex=="M"){
       s_selected<- s+s1 #host mortality rate with adjustment for male fish
       #f=body_area
       f<-as.vector(f[,2])## body areas for male fish
   }
    #selecting microhabitat preference rate depending on parasite strain
    if(strain=="Gt3"){e_selected<- e1}
    if(strain=="Gt"){e_selected<- e2}
    if(strain=="Gb"){e_selected<- e3}
     #selecting birth and deaths rates depending on parasite strain
    if(strain=="Gt3"){b_selected<-b1; d_selected<-d1}
    if(strain=="Gt"){b_selected<- b2; d_selected<- d2}
    if(strain=="Gb"){b_selected<- b3; d_selected<- d3}

    # birth rates; death rates; movement rates; immune response
  QB <- matrix(0, 4, 2) # QB[k,j] = birth rate for parasites location j age k
  QD <- matrix(0, 4, 2) # QD[k,j] = death rate for parasites location j age k
  QM_forward <- matrix(0, 4, 2) # QM[k,j] = movement rate for j age k
  QM_backward <- matrix(0, 4, 2)
  QI <- rep(0, 4) # QI[j] = rate at which location j increases immune response
  for (j in 1:4) {
    QI[j] <- sum(A[j, ]) * r_selected
    for (k in 1:2) {
      QB[j, k] <- A[j, k] * (1-(A[j, k]/(f[j]*a*kappa)))*b_selected[k, B[j]]
      QD[j, k] <- A[j, k] * (1-(A[j, k]/(f[j]*a*kappa)))*d_selected[k, B[j]]
      QM_forward[j, k] <- A[j, k] *m[k, B[j]]*e_selected
      QM_backward[j, k] <- A[j, k] *m[k, B[j]]*(1-e_selected)
                  }
          }
  # total rates
  laB <- sum(QB) #total birth rate
  laD <- sum(QD) #total death rate
  laM_forward <- sum(QM_forward) # total rate for forward movement
  laM_backward <- sum(QM_backward) # total rate for backward movement
  laI <- sum(QI)# total rate of immuune response
  laX <- sum(A) * s_selected # host fish mortality rate
  #overall total
  la <- na.zero(abs(laB + laD + laM_forward+laM_backward+ laI + laX))

#Returns rates in relation to birth, death, movement,
# immune response, host mortality and total rate (la)
    return(list(laB=laB,laD=laD,laM_forward=laM_forward,
                laM_backward=laM_backward,laI=laI,laX=laX,
                la=la,QB=QB,QD=QD,QM_forward=QM_forward,
                 QM_backward=QM_backward,QI=QI))
 }
```

G.3: Function for extracting parasite numbers & experimental descriptors of the empirical data

```
Experiment_descriptors<-function(empirical_data){
###Fish-parasite combinations/groups###
parasite_fish<-c("Gt3-OS","Gt3-LA","Gt3-UA","Gt-OS","Gt-LA","Gt-UA","Gb-OS","Gb-LA","Gb-UA")
levels(empirical_data$Sex_fish)<-c("F","M")

empirical_data$LowerRegion<-empirical_data$LB+empirical_data$Pelvic+
empirical_data$Anal+empirical_data$Dorsal
empirical_data$UpperRegion<-empirical_data$UB +Combined_data$Pectoral

### Data across the four recategorized body regions###
Data_fourRegions<-empirical_data[,c(1,15,16,8,9,10,12,13,11,14)]
#head(Data_fourRegions,n=4)

###Data across parasite strains###
Gt3_data<-Data_fourRegions[Data_fourRegions$Parasite_strain=="Gt3",]
Gt_data<-Data_fourRegions[Data_fourRegions$Parasite_strain=="Gt",]
Gb_data<-Data_fourRegions[Data_fourRegions$Parasite_strain=="Gb",]

#To store extracted information
Parasite_fish_data=NULL;fishID=NULL;
numF=NULL;pop_obs=NULL;alive_obs=NULL;
fishSize=NULL;fishSex=NULL;Size=NULL;Sex=NULL; Parasite_strain=NULL;Strain=NULL;
Fish_type=NULL;Fish_stock=NULL

##Data of each parasite strain across fish stocks ##
Parasite_fish_data[[parasite_fish[1]]]<-split(Gt3_data,Gt3_data$Fish_strain)$"OS"
Parasite_fish_data[[parasite_fish[2]]]<-split(Gt3_data,Gt3_data$Fish_strain)$"LA"
Parasite_fish_data[[parasite_fish[3]]]<-split(Gt3_data,Gt3_data$Fish_strain)$"UA"
Parasite_fish_data[[parasite_fish[4]]]<-split(Gt_data,Gt_data$Fish_strain)$"OS"
Parasite_fish_data[[parasite_fish[5]]]<-split(Gt_data,Gt_data$Fish_strain)$"LA"
Parasite_fish_data[[parasite_fish[6]]]<-split(Gt_data,Gt_data$Fish_strain)$"UA"
Parasite_fish_data[[parasite_fish[7]]]<-split(Gb_data,Gb_data$Fish_strain)$"OS"
Parasite_fish_data[[parasite_fish[8]]]<-split(Gb_data,Gb_data$Fish_strain)$"LA"
Parasite_fish_data[[parasite_fish[9]]]<-split(Gb_data,Gb_data$Fish_strain)$"UA"

for (pf in 1:length(parasite_fish)){
   #Assigning unique ID for  data
   fishID[[pf]]<- unique(Parasite_fish_data[[parasite_fish[pf]]]$Fish_ID)
   #Total number of fish used for data
   numF[[pf]] <- length(fishID[[pf]])
   #Observed data or matrix across 4 regions
   pop_obs[[pf]] <- array(dim = c(numF[[pf]], 4, 9))
   #Array for time steps fish was alive for each combination
   alive_obs[[pf]] <- array(dim = c(numF[[pf]], 9))
   #NB: Fish size & sex  over time
   #Array of fish size across the 9 time steps  for each combination
    Size[[pf]]<- array(dim = c(numF[[pf]], 9))
    Sex[[pf]]<- array(dim = c(numF[[pf]], 9))
    Parasite_strain[[pf]]<-  array(dim = c(numF[[pf]], 9))
    Fish_type[[pf]]<-  array(dim = c(numF[[pf]], 9))

    for(i in 1:numF[[pf]]){
         pop_obs[[pf]][i,,] <-
         t(Parasite_fish_data[[parasite_fish[pf]]]
         [Parasite_fish_data[[parasite_fish[pf]]]
         $Fish_ID==fishID[[pf]][i], 1:4])
         alive_obs[[pf]][i, ] <-ifelse(is.na(pop_obs[[pf]][i,1,]), 2, 1)
          Size[[pf]][i, ]<-
          Parasite_fish_data[[parasite_fish[pf]]]
          [Parasite_fish_data[[parasite_fish[pf]]]
          $Fish_ID==fishID[[pf]][i], 9]
         #1=Female fish & 2=Male fish
          Sex[[pf]][i, ]<-
          paste(Parasite_fish_data[[parasite_fish[pf]]]
          [Parasite_fish_data[[parasite_fish[pf]]]
          $Fish_ID==fishID[[pf]][i], 10])
          Parasite_strain[[pf]][i, ]<-
          paste(Parasite_fish_data[[parasite_fish[pf]]]
          [Parasite_fish_data[[parasite_fish[pf]]]
          $Fish_ID==fishID[[pf]][i], 7])
          Fish_type[[pf]][i ,]<-
          paste(Parasite_fish_data[[parasite_fish[pf]]]
          [Parasite_fish_data[[parasite_fish[pf]]]
          $Fish_ID==fishID[[pf]][i], 8])
                        }

### Experiment descriptors ####
fishSize[[pf]]<-apply(Size[[pf]],1,unique)
fishSex[[pf]]<- apply(Sex[[pf]],1,unique)
Strain[[pf]]<- apply(Parasite_strain[[pf]],1,unique)
Fish_stock[[pf]]<- apply(Fish_type[[pf]],1,unique)
                    }

    #return data on experiment descriptors (fish size, sex,
    #fish type & strain) for each parasite-fish group
    return(list(fishSize=fishSize,fishSex=fishSex,
    Strain=Strain,Fish_stock=Fish_stock,numF=
    numF,fishID=fishID,pop_obs=pop_obs,
    alive_obs=alive_obs,fishID=fishID))
     }
```

G.4: Function for updating exact SSA

```
# Function for updating exact SSA
#For for updating simulation events across the 4 body regions
(Tail, Lower region, Upper region, Head)

SSA_update_event <- function(A,B,J,X,laB,laD,laM_forward,
                     laM_backward,laI,laX,la,QB,QD,
                     QM_forward,QM_backward,QI) {

  #Inputs:
  # A[j,k] gives the number of parasites at location j, age k, where
  # B[j] = immune response at location j (1 for no response; 2 for a response)
  #J is transition matrix
  #X is survival status (1= alive, 2=dead)
  #And all rates in relation to birth,
  #death, movement, immune response, host mortality
  # and total rate (la)

 if (la == 0) {
    return(list(A = A, B = B, t_incr_SSA = Inf, X = X)) # zero population
          }

  U <- runif(1, 0, la)  #uniform random number/generator

  if (U < laB) {# birth
    i <- sample(8, 1, prob = abs(QB))
    j <- ((i-1) %% 4) + 1 # location
    k <- ((i-1) %/% 4) + 1 # age
    if (k == 1) {
      A[j, 2] <- A[j, 2] + 1
    } else {
      A[j, 1] <- A[j, 1] + 1
    }
  } else if (U < sum(c(laB,laD))) {# death
    i <- sample(8, 1, prob = abs(QD))
    j <- ((i-1) %% 4) + 1 # location
    k <- ((i-1) %/% 4) + 1 # age
    A[j, k] <- A[j, k] - 1
  } else if(U < sum(c(laB,laD,laM_forward))){#forward movement

    i <- sample(8, 1, prob = abs(QM_forward))
    j <- ((i-1) %% 4) + 1 # location
    k <- ((i-1) %/% 4) + 1 # age
    j_new <- sample(4, 1, prob =abs(J[j,]))#new location
    A[j, k] <- A[j, k] - 1
    A[j_new, k] <- A[j_new, k] + 1
  } else if (U < sum(c(laX,laI,laB,laD,laM_forward)) ){#backward movement
    i <- sample(8, 1, prob = abs(QM_backward))
    j <- ((i-1) %% 4) + 1 #location
    k <- ((i-1) %/% 4) + 1 # age
    j_new <- sample(4, 1, prob =abs(J[j,]))#new location
    A[j, k] <- A[j, k] - 1
    A[j_new, k] <- A[j_new, k] + 1
  }else if(U < sum(c(laB,laD,laM_forward,laM_backward,laI)) ){#immune response

    i <- sample(4, 1, prob = abs(QI))
    B[i] <- 2
  } else {# fish death
    X <- 2
  }
   t_incr_SSA <- rexp(1, la) #time increment for exact SSA

   #Output: returns A[j,k] the number of parasites
   # at location j, age k
   # where B[j] = immune response at location j
   #(1 for no response; 2 for a response)
   # t_incr_SSA= time increment for exact SSA
   # X survival status
  return(list(A = A, B = B,t_incr_SSA=t_incr_SSA, X = X))
}
```

G.5: Function for updating hybrid τ-leaping

```
#Function for updating tau-leaping
taulealping_update_event <-function(A,B,J,X,laB,laD,
                      laM_forward,laM_backward,laI,laX,la
                     ,QB,QD,QM_forward,QM_backward,QI,tau){

  #Inputs:
  # A[j,k] gives the number of parasites at location j, age k, where
  # B[j] = immune response at location j (1 for no response; 2 for a response)
  #J is transition matrix
  #X is survival status (1= alive, 2=dead)
  # tau is the leap size
  #And all rates in relation to birth,
  #death, movement, immune response, host mortality
  # and total rate (la)

             U <- runif(1, 0, la)
             if(U<laX)  X <-2 #Fish mortality
             else if(U<sum(c(laX,laI))){#Immune response

                j <- sample(4, 1, prob = abs(QI))
                B[j] <- 2
           } else if (U<sum(c(laX,laI,laB,laD,laM_forward))){
             #brith, death or forward movement
                i <- sample(8, 1, prob = abs(QB+ QD+QM_forward))
                j <- ((i-1) %% 4) + 1 # current location
                k <- ((i-1) %/% 4) + 1 # age
                j_new <- sample(4, 1, prob =abs(J[j,]))# new location
                A[j,k]<- A[j,k] + rpois(1,abs(laB*tau))-
                rpois(1,abs(laD*tau))
                -rpois(1,abs(laM_forward*tau))
                A[j_new,k]<-  A[j_new,k]
                +rpois(1,abs(laB*tau))
                -rpois(1,abs(laD*tau))
                +rpois(1,abs(laM_forward*tau))
          } else if (U< sum(c(laX,laI,laB,laD,laM_forward,laM_backward))){
          #birth, death or backward movement
                i <- sample(8, 1, prob = abs(QB+ QD+QM_backward))
                j <- ((i-1) %% 4) + 1 # current location
                k <- ((i-1) %/% 4) + 1 # age
                j_new <- sample(4, 1, prob =abs(J[j,]))# new location
                A[j,k]<- A[j,k]+rpois(1,abs(laB*tau))-
                rpois(1,abs(laD*tau))-
                rpois(1,abs(laM_backward*tau))
                A[j_new,k]<-A[j_new,k]
                +rpois(1,abs(laB*tau))-rpois(1,abs(laD*tau))
                +rpois(1,abs(laM_backward*tau))
             }

    #Output: returns A[j,k] gives the number of parasites at location j, age k,
    #where B[j] = immune response at location j
    #(1 for no response; 2 for a response)
    # X=survival status
         return(list(A = A, B = B, X = X))

    }
```

G.6: Function for simulating infection dynamics for a single fish

```
#tau-leaping simulation for a single fish
sim_tauleap_singlefish<- function(A0, B0,J, b1,b2,b3,
                        d1,d2,d3, m, r,r1,
                       r2,r3,s,s1,e1,e2,e3,kappa,
                 f,a,fish_sex,fish_type,strain,error){
 #Inputs: inital conditions, parameter values, fish sex,
 #fish type, parasite strain and error bound
 #f=body area (dependent on fish sex and size)
 #parasite_fish=c("Gt3-OS","Gt3-LA","Gt3-UA",
 #"Gt-OS","Gt-LA","Gt-UA","Gb-OS","Gb-LA","Gb-UA")
 #strain-parasite type to be simulated
 parasite_fish<-paste(strain,"-",fish_type)
 pop=NULL; alive =NULL; exploded=NULL;Leap_sizes=NULL;
 A<-A0; B<- B0
 #observed discrete times
 save_ti <- c(1, 3, 5, 7, 9, 11, 13, 15, 17)
 save_TF <- rep(FALSE, length(save_ti))
 ti<- 0 # initial time
 #parasite pop at each location (rows) & timepoint (cols)
  pop[[parasite_fish]] <- matrix(NA, 4, length(save_ti))
 # host fish status at each time point
  alive[[parasite_fish]] <- rep(2, length(save_ti))
  pop_ti <- rowSums(A)
  # host survival status (alive=1; dead=2)
  alive_ti <- 1
  exploded[[parasite_fish]] <- FALSE
  # stop the simulation if total population>pop_max
  pop_max <- 10000
  X <- 1 # fish starts out alive
 while(sum(save_TF) < length(save_ti)){
  #### Computing the rates ######
        computed_rates<-compute_rates(A=A, B=B,b1=b1,b2=b2,
      b3=b3, d1=d1,d2=d2,d3=d3,m=m,r=r,r1=r1,r2=r2,r3=r3
      ,s=s,s1=s1,e1=e1, e2=e2,e3=e3,kappa=kappa,f=f,a=a,
      fish_sex=fish_sex,fish_type=fish_type,strain=strain)
  laB<-computed_rates$laB
  laD<-computed_rates$laD
  laM_forward<-computed_rates$laM_forward
  laM_backward<-computed_rates$backward
  laI<-computed_rates$laI
  laX<-computed_rates$laX
  la<-computed_rates$la
  QB<-computed_rates$QB
  QD<-computed_rates$QD
  QM_forward<-computed_rates$QM_forward
  QM_backward<-computed_rates$QM_backward
  QI<-computed_rates$QI

## sim_tauleap_singlefish function continues at next page##
++++
#sim_tauleap_singlefish function continuation

# Determining the  switching condition between the exact SSA
#and the Tau-leaping algorithm
        #selecting birth and deaths rates
        #depending on parasite strain for leap size
       if(strain=="Gt3"){b_selected<-b1; d_selected<-d1}
       if(strain=="Gt"){b_selected<- b2; d_selected<-d2}
       if(strain=="Gb"){b_selected<- b3; d_selected<-d3}
     #finding average birth & death rates (eqn 6.1)
   b_avg<- mean(b_selected[,1]);d_avg<-mean(d_selected[, 1])
    Leap_sizes[[1]]<-(error*(b_avg+d_avg))/
    (abs(b_avg-d_avg)*max(b_avg,d_avg))
        Leap_sizes[[2]]<- sum(A)*(error*(b_avg+d_avg))^2
        /((b_avg+d_avg)* max(b_avg^2,d_avg^2))
    # the leap size: time increment for tau-leaping
        tau<- na.zero(min(Leap_sizes[[1]],Leap_sizes[[2]]))
        leap_condition<- na.zero((1/(10*la)))
        if (sum(pop_ti) > pop_max) {
            exploded[[parasite_fish]] <- TRUE
            break }
        if (alive_ti == 2)  break
      #Running tau-leaping if tau >leap_condition
        if(tau >leap_condition){ #Execute tau-leaping
            out<-taulealping_update_event(A=A,
            B=B,J=J,X=X,laB=laB,laD=laD,laM_forward=laM_forward,
              laM_backward=laM_backward,laI=laI,laX=laX,
              la=la,QB=QB,QD=QD,QM_forward=QM_forward,
              QM_backward=QM_backward,QI=QI,tau=tau)
              X<- out$X;A<- out$A;B<- out$B;
              time_increment=tau
                               } #end of tau-leaping
        else if(tau <=leap_condition){
        #Execute exact SSA if tau <=leap_condition
              out <- SSA_update_event(A=A,B=B,J=J,X=X,
              laB=laB,laD=laD,laM_forward=laM_forward,
              laM_backward=laM_backward,laI=laI,laX=laX,
              la=la,QB=QB,QD=QD,QM_forward=QM_forward,
              QM_backward=QM_backward,QI=QI)
              #time increment for SSA
              time_increment<- out$t_incr_SSA
              X<- out$X; A<- out$A;B<- out$B
                                   } #end of exact SSA
  ti <- ti +time_increment #updating time ti
  save_new <- which((ti >= save_ti) & !save_TF)
         for (i in save_new) {
              pop[[parasite_fish]][,i] <- pop_ti
              alive[[parasite_fish]][i] <- alive_ti}
              save_TF <- (ti >= save_ti)
              pop_ti <- rowSums(A)
              alive_ti <- X
     #break if parasite number<0 at any body region
         if(any(pop_ti<0) ==TRUE) break
            }
  #Output: returns pop (parasite pop at each location and time)
  #alive: survival status of fish
  # exploded: explosion status
  #(whether parasite numbers>pop_max=10000)
  # parasite_fish: the host-parasite group being simulated
  return(list(pop = pop[[parasite_fish]],
   alive = alive[[parasite_fish]],
  exploded =exploded[[parasite_fish]],
  parasite_fish=parasite_fish))
  }
```

G.7: Exporting external scripts and extracting relevant information from the empirical data for group simulation

```
#tau-leaping simulation for a group of fish
###exporting external scripts###
#Script of function for computing event rates
source("Computing-rates-script.r")
# Script of function for updating exact SSA
source("Update-exactSSA-script.r")
# Script of function for updating tau-leaping
source("Update-tauleaping-script.r")
# Script of function experimental descriptors
# (fish type, strain, fish size, fish sex &
#areas of the 4 body regions)
source("Descriptors-Data-script.r")
# Script of function for simulating parasites
#only a single fish over time and across body regions
source("Simulation-single-fish-script.r")

#Importing empirical data
Combined_data <- read.csv(file="Parasite_Data.csv")

#Importing data for area of the 8
#body parts across 18 fish (measured in mm^2)
Bodyparts_area<-read.csv(file="Area_Fish_bodyParts.csv")
#Experimental descriptors
Descriptors<-Experiment_descriptors(empirical_data=
             Combined_data)
fishSize  <- Descriptors$fishSize #fish size
fishSex   <- Descriptors$fishSex #fish sex
Strain    <- Descriptors$Strain # parasite strain
Fish_stock<-Descriptors$Fish_stock #fish stock
# total fish for each parasite-fish group
numF     <-  Descriptors$numF
# fish IDs for each parasite-fish group
fishID   <-  Descriptors$fishID
#observed parasite numbers for each parasite-fish group
pop_obs  <-  Descriptors$pop_obs
# observed surviva status for each parasite-fish group
alive_obs<- Descriptors$alive_obs
#body areas for female (column 1) & male (column 2) fish
Area_normalized<-Body_area(Area_data=Bodyparts_area)


#Initial simulation inputs for A (parasite numbers)
#and  B (immune status)
A0 <- matrix(0, 4, 2)
A0[1, 1] <- 2  #Intial parasites at the tail
#initial immune response at 4 body regions
#(1=no response, 2=response)
B0 <- rep(1, 4)
#Transition matrix(between body regions)
J<- matrix(c(0,     1,      0,      0,
            1/2,    0,     1/2,     0,
             0,    1/2,     0,     1/2,
             0,     0,      1,      0), 4, 4, byrow=TRUE)
```

G.8: Function for simulating infection dynamics for a group of fish corresponding to the empirical data

```
#To simulate group of fish for each parasite-fish combination as observed
#in the empirical data

SimGroup_tauleap<-function(theta1,fish_sex,
                   fish_type,strain,fish_size,error){
    #Inputs: theta1= parameter values from prior distribution
    #fish_sex= sex of fish
    #fish_type= type of fish
    #strain= parasite strain
    #fish_size= fish size
    #error= error bound of tau-leaping
    pop_sim<-NULL; alive_sim<- NULL;
    exploded_sim<-NULL;results<-NULL;group<-NULL

  for(pf in 1:9){
     pop_sim[[pf]]<- array(dim = c(numF[[pf]], 4, 9))
      #Array for time steps fish was alive for each combination
     alive_sim[[pf]]<-array(dim = c(numF[[pf]], 9))
      #Array for time steps parasites>pop_max for each combination
     exploded_sim[[pf]]<-array(dim = c(numF[[pf]], 9))

      for(i in 1:numF[[pf]]){
          results[[pf]]<-sim_tauleap_singlefish(A0=A0,
           B0=B0,J=J,b1=matrix(exp(theta1[1:2]), 2,2),
           b2=matrix(exp(theta1[3:4]), 2, 2),
           b3=matrix(exp(theta1[5:6]), 2,2),
           d1=matrix(exp(theta1[7:8]), 2, 2, byrow=TRUE),
           d2=matrix(exp(theta1[9:10]), 2,2,byrow=TRUE),
           d3=matrix(exp(theta1[11:12]),2, 2,byrow=TRUE),
           m=matrix(exp(theta1[13]), 2,2),
           r=exp(theta1[14]),r1=exp(theta1[15]),
           r2=exp(theta1[16]),r3=exp(theta1[17]),
           s=exp(theta1[18]),s1=exp(theta1[19]),
           e1=exp(theta1[20]),e2=exp(theta1[21]),
           e3=exp(theta1[22]),kappa=exp(theta1[23]),
           f=Area_normalized,a=fish_size[[pf]][i],
           fish_sex=fish_sex[[pf]][i],
           fish_type=fish_type[[pf]][i],
           strain=strain[[pf]][i],error=error)
           pop_sim[[pf]][i, ,]<- results[[pf]]$pop
           alive_sim[[pf]][i, ]<-  results[[pf]]$alive
           exploded_sim[[pf]][i, ]<- results[[pf]]$exploded
          group[[pf]]<-results[[pf]]$parasite_fish
                               }
                     }
    #Output: returns
    #(pop_sim=parasite pop per region and time)
    #alive_sim: survival status of fish
    # exploded_sim: explosion status
    #(whether parasite numbers>pop_max=10000)
    # group: the host-parasite groups being simulated
  return(list(pop_sim=pop_sim,alive_sim=alive_sim,
  exploded_sim=exploded_sim, group= unlist(group)))
    }
```

G.9: Function for performing ROPE+HDI Bayesian hypothesis testing

```
# Function to perform Region of Practical Equivalence (ROPE) and Highest
#Density Interval (HDI)
require(bayestestR)
ROPE_Cred_Int<- function(theta_distn_diff,
parameter_labels, ci_percent=0.89){

    if(is.list(theta_distn_diff)==FALSE){
    #standard deviation of differenced posterior samples
        sigma_d<- sd(theta_distn_diff)
        output<-bayestestR::equivalence_test(

    #ROPE range recommended by Norman et al (2003)
        theta_distn_diff, range =c(-.5*sigma_d,.5*sigma_d),
        ci = ci_percent,ci_method = "HDI")
    final_output<- cbind(parameter_labels,output)
    names(final_output)[1]<- "Parameter"
    return(final_output)
    }


    #theta_distn_diff=a list of posterior samples of
    # differences of parameters of interest
    output<-list() #save ROPE+HDI results
    for(i in seq_along(parameter_labels)){
    #standard deviation of differenced posterior samples
     sigma_d<- sd(theta_distn_diff[[i]])

      #ROPE range is recommend by Norman et al (2003)
    output[[i]]<- bayestestR::equivalence_test(
    theta_distn_diff[[i]],range =c(-.5*sigma_d,.5*sigma_d),
     ci = ci_percent,ci_method = "HDI")
        }


     #Function returns ROPE interval,
     #ROPE Percentage or coverage probability,
     #ROPE equivalence decision and the corresponding HDI

    final_output<- do.call("rbind",output)
    final_output<- cbind(parameter_labels,final_output)
    names(final_output)[1]<- "Parameters"
    return(final_output)
}
```